# A Comprehensive Survey on Knowledge Distillation

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Deep Neural Networks (DNNs) have achieved notable performance in the fields of computer vision and natural language processing with various applications in both academia and industry. However, with recent advancements in DNNs and transformer models with a tremendous number of parameters, deploying these large models on edge devices causes serious issues such as high runtime and memory consumption. This is especially concerning with the recent large-scale foundation models, Vision-Language Models (VLMs), and Large Language Models (LLMs). Knowledge Distillation (KD) is one of the prominent techniques proposed to address the aforementioned problems using a teacher-student architecture. More specifically, a lightweight student model is trained using additional knowledge from a cumbersome teacher model. In this work, a comprehensive survey of knowledge distillation methods is proposed. This includes reviewing KD from different aspects: distillation sources, distillation schemes, distillation algorithms, distillation by modalities, applications of distillation, and comparison among existing methods. In contrast to most existing surveys, which are either outdated or simply update former surveys, this work proposes a comprehensive survey with a new point of view and representation structure that categorizes and investigates the most recent methods in knowledge distillation. This survey considers various critically important subcategories, including KD for diffusion models, 3D inputs, foundational models, transformers, and LLMs. Furthermore, existing challenges in KD and possible future research directions are discussed.

## 1 Introduction

With the emergence of DNNs, the fields of Computer Vision (CV) and Natural Language Processing (NLP) have been revolutionized, and most tasks in these domains are now solved using DNNs. While a simple ResNet (He et al., 2016) model or BERT (Kenton & Toutanova, 2019) can be easily trained on most of existing GPUs, the advent of large models like LLMs and foundational vision models has made training and inference a significant challenge in terms of runtime and memory usage, especially for deployment on edge devices like mobile phones due to their widespread use. Despite their high performance, these large models often have complex architectures and are heavily over-parameterized (Kaplan et al., 2020; Fischer et al., 2024). For example, the weight matrices in LLMs have been shown to be low-rank matrices (Hu et al., 2021).

To address the issue of large models, various solutions have been proposed, such as efficient network architectures, compression methods, pruning, quantization, low-rank factorization, and knowledge distillation. Efficient network blocks, including MobileNet (Howard, 2017), ShuffleNet (Zhang et al., 2018), and BiSeNet (Yu et al., 2018) have been introduced in recent years. Pruning methods, a type of compression technique, aim to remove unnecessary layers and parameters from models with minimal impact on performance. Low-rank factorization methods reduce parameters through matrix decomposition. Unlike other methods, KD does not modify the network's layers or parameters. In KD, a lightweight network (student) is trained under the supervision of a more complex model (teacher) with deeper architecture and greater number of parameters. This concept, first proposed by Buciluă et al. (2006) and later popularized by Hinton (2015) as KD, introduced the idea of training the student network by mimicking the teacher's output distribution as soft labels.

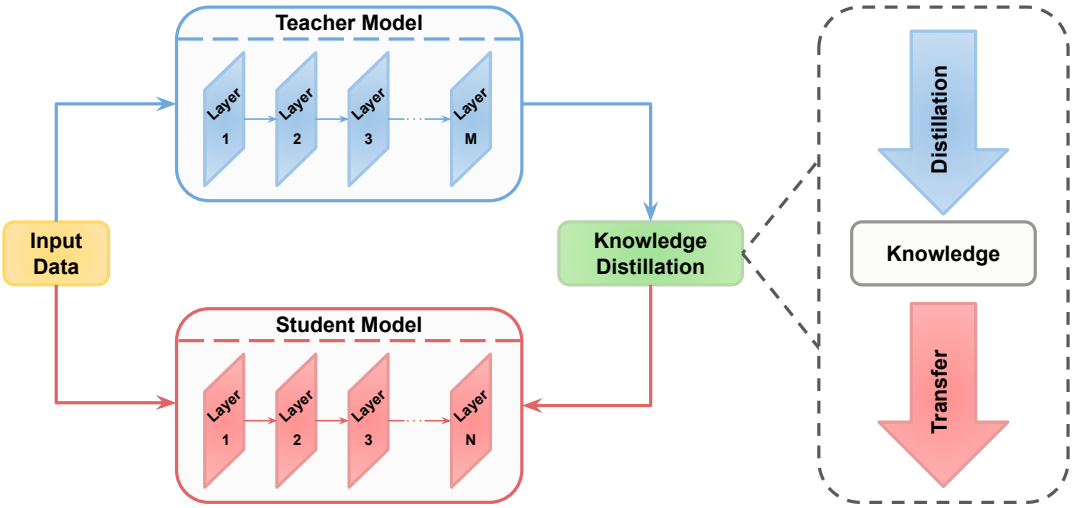


Figure 1: A general teacher-student framework for knowledge distillation.

Aside from the large number of parameters, large models also require a significant amount of labeled data for training. Another purpose of KD is knowledge transfer, which can involve transferring knowledge from a source task to a target task that lacks sufficient labeled data. Additionally, in cases where data privacy is a concern, data-free knowledge distillation becomes helpful. This approach addresses the issue by generating synthetic data, eliminating the need to store sensitive data. The key challenges in KD are determining what knowledge to transfer, selecting the appropriate algorithm, and designing the student and teacher architectures. A diagram of a general teacher-student KD method is shown in Figure 1.

With the significant growth in the number of published papers in the field of KD and its wide applications across various tasks and domains, several survey papers have been presented to review KD from different perspectives. The prominent work Gou et al. (2021) provides a comprehensive survey on KD, reviewing it from the aspects of knowledge types, algorithms, and schemes, while including comparisons between different methods. Another notable work, Wang & Yoon (2021), offers an overview of model compression based on the teacher-student architecture specifically for vision tasks. Alkhulaifi et al. (2021) presents a survey on KD and proposes a new metric for comparing distillation methods based on size and accuracy, while Yang et al. (2023c) categorizes existing approaches based on the type of knowledge source used for distillation. More recently, Hu et al. (2023a) delivers a comprehensive survey, exploring knowledge optimization objectives associated with various representations, and Moslemi et al. (2024) updates earlier surveys by providing an extensive review of KD methods. Additionally, they briefly investigate distillation in VLMs and discuss the challenges of distillation under limited data scenarios.

While each of these surveys provides a detailed summary of papers from different perspectives, they also have some drawbacks. First, although these surveys analyze existing approaches in each category, they fail to address recent advancements, particularly in feature-based distillation methods. Despite the significant growth of feature-based distillation, which now dominates state-of-the-art methods, many recent and impactful works are overlooked. Furthermore, adaptive distillation and contrastive distillation algorithms are rarely discussed. Second, with the recent emergence of foundation models and LLMs, these surveys have largely overlooked their potential for distilling knowledge into other models. One significant shortcomings of Gou et al. (2021), as the most prominent survey in the field, is its lack of explanation regarding the applications of distillation in newly introduced models such as foundation models and LLMs, which have gained significant attention in recent years. Third, none of the existing surveys have investigated KD for 3D inputs like point clouds. As 3D tasks garner increasing attention in top research venues, the lack of sufficient approaches and models in comparison to the image domain highlights the importance of KD as an effective method for training the models in this area. Table 1 shows a summarized comparison between existing surveys and this work.

Table 1: Comparison between existing KD surveys and the current survey. New topics not covered in previous works are highlighted in bold to indicate the progress and scope of this survey.

| Paper | Source | Algorithm | Modality | Application |
|---|---|---|---|---|
| Gou et al. (2021) | Logit, Feature, Similarity | Adversarial, Attention, Cross-modal, Data-free, Graph, Lifelong, Multi-teacher, NAS, Quantized | Image, Text, Speech, Video | Visual Recognition |
| Wang & Yoon (2021) | Logit, Feature, Similarity | Adversarial, Cross-modal, Data-free, Few-shot, Graph, Incremental, Multi-teacher, Reinforcement | Image, Video | Visual Recognition, Self-supervised Learning |
| Hu et al. (2023a) | Logit, Feature, Similarity, Mutual Information | Cross-modal, Federated, Graph, Multi-teacher | Image, Text, Speech, Video | Visual Recognition, Ranking, Generation, Regression |
| Moslemi et al. (2024) | Logit, Feature, Similarity | Adversarial, Attention, Cross-modal, Data-free, Graph, Lifelong, Multi-teacher, NAS, Quantized | Image, Text | Vision-Language Models, Medical Image |
| This paper | Logit, Feature, Similarity | **Adaptive**, Adversarial, Attention, **Contrastive**, Cross-modal, Graph, Multi-teacher | **3D/Multi-view**, Image, Text, Speech, Video | Visual Recognition, **Foundation Models**, **Transformers**, **LLMs**, **Diffusion Models**, Self-supervised Learning |

In this work, a comprehensive survey on knowledge distillation is presented. This includes reviewing existing KD methods from different perspectives: the source of distillation, distillation algorithms, distillation schemes, distillation by modalities, and applications of KD.

The sources reviewed include logit-based, feature-based, and similarity-based distillation methods. In contrast to existing surveys, a more comprehensive classification of these methods is presented, highlighting recent advancements, particularly in feature-based distillation.

Regarding algorithms, methods in attention-based, adversarial, multi-teacher, cross-modal, graph-based, adaptive, and contrastive distillation are reviewed. Notably, adaptive and contrastive distillation are two recent categories that, despite their importance, have yet to be presented in previous works.

In terms of schemes, offline, online, and self-distillation methods are covered. Compared to previous surveys, a new section is introduced that classifies existing KD methods by modalities such as video, speech, text, multi-view, and 3D data.

In applications, a comprehensive exploration is conducted on the applications of KD in important areas including self-supervised learning, foundational models, transformers, diffusion models, visual recognition tasks, and LLMs. Finally, a quantitative comparison of prominent methods is provided, along with a discussion of current challenges and future directions.

Figure 2 shows the organization of this paper. In summary, the main contributions of this work are:

- Presenting a comprehensive survey on knowledge distillation as an essential area of research. This includes reviewing existing methods based on distillation sources, distillation algorithms, distillation schemes, modalities, and applications of distillation.
- Classifying recent distillation methods by sources, with a particular focus on feature distillation methods due to their importance and widespread application.
- Introducing two new distillation algorithms: adaptive distillation and contrastive distillation. These important categories have gained significance in distillation, especially with the advent of foundation models like CLIP, where contrastive distillation plays an increasingly crucial role.

- Reviewing the distillation methods for multi-view and 3D data. Due to the significant role of KD in 3D tasks nowadays, exploring distillation in the 3D domain is crucial, yet it has been overlooked in previous comprehensive surveys.
- Exploring the applications of KD in self-supervised learning, foundation models, transformers, diffusion models, and LLMs. Distillation from foundation models is a hot topic, and distillation in LLMs is particularly critical due to their large number of parameters.
- Comparing prominent distillation methods and discussing the existing challenges and future directions of KD.

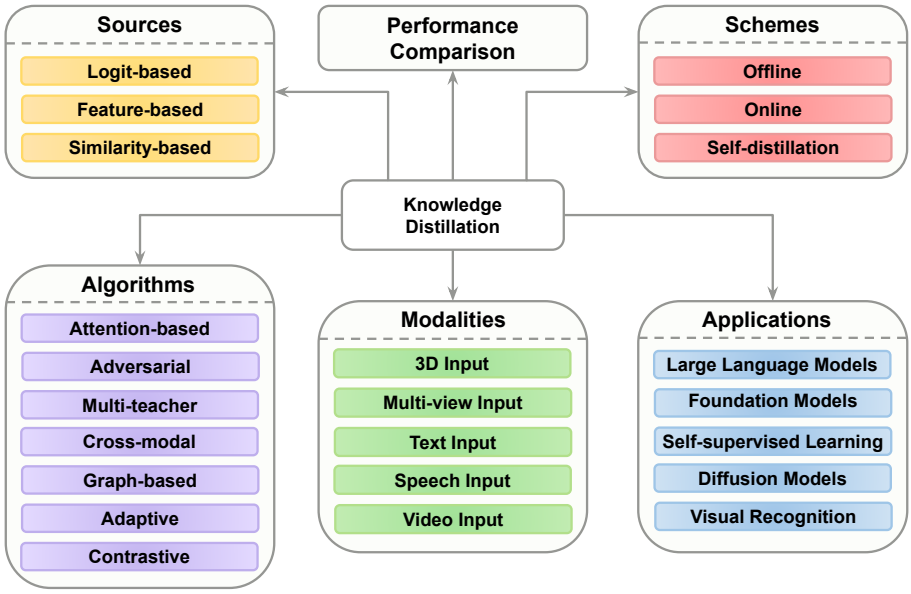


Figure 2: Diagram of the contents of this paper. Knowledge distillation is reviewed from different aspects, including sources, schemes, algorithms, modalities, and applications.

## 2 Sources

The primary component of a distillation method is the source of knowledge being distilled. Various sources of knowledge can be utilized for distillation. In the initial concept of distillation (Hinton, 2015), logits were employed as the teacher's final output. Alongside logit distillation (Ba & Caruana, 2014; Mirzadeh et al., 2020; Hao et al., 2023; Sun et al., 2024a), the activations (Heo et al., 2019b), neurons, and features (Romero et al., 2014; Zagoruyko & Komodakis, 2016) of intermediate layers play a crucial role in guiding the student network. Moreover, relationships and similarities among channels (Liu et al., 2021c), instances (Tung & Mori, 2019), and classes (Wang et al., 2020d) provide higher-order information that facilitates the transfer of knowledge from the teacher to the student. In this section, KD methods based on their sources of distillation are examined. Specifically, existing methods are classified into three categories: Logit-based, Feature-based, and Similarity-based as shown in Figure 3. Table 2 summarizes logit-based and similarity-based distillation methods and Table 3 presents detailed information on existing feature distillation methods. Subsequently, each type of distillation is explained in detail.

### 2.1 Logit-based Distillation

The most straightforward source for KD is logits. Logits are outputs of the last layer of the model, and the idea is to force the student model to mimic the final predictions of the teacher, which are supposed to

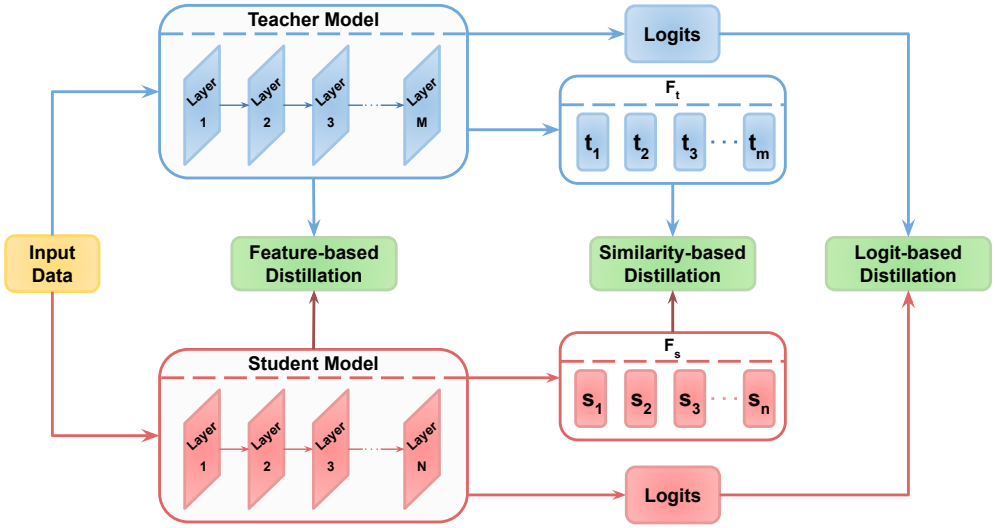


Figure 3: Illustration of the sources for distillation, including logits, features, and similarities. Logits are the model's last layer output, features are intermediate outputs of the model, and similarities are relationships between features, channels, samples, etc.

contain informative dark knowledge (Hinton, 2015) of the teacher. The prediction from the last layer can vary for each task. For example, logits in classification contain predictions for each class, in detection they contain predicted coordinates, and in pose estimation, they can represent heatmaps. In general, let $Z_T$ and $Z_S$ be logits of the last layers of the teacher and student, respectively. Then the logit-based distillation loss ($L_{ld}$) is defined as

$$L_{ld}(Z_T, Z_S) = \ell_{logit}(Z_T, Z_S) \tag{1}$$

where $\ell_{logit}$ is a metric for measuring the difference between the logits.

The first papers on KD were proposed for image classification, where soft-labels were introduced by applying softmax on logits (Hinton, 2015; Ba & Caruana, 2014). These output distributions, or soft-labels, provide information on the probability of the input belonging to each class. The distillation loss function for soft-labels is expressed as

$$p(Z_i, \tau) = \frac{exp(Z_i/\tau)}{\Sigma_i exp(Z_i/\tau)} \tag{2}$$

$$L_{ld}(p(Z_t, \tau), p(Z_s, \tau)) = KL(p(Z_t, \tau), p(Z_s, \tau)) \tag{3}$$

where $exp$ is softmax function, $\tau$ is the temperature factor that controls the softening of logits, and $KL(\cdot)$ is the Kullback–Leibler Divergence (KL) function. By optimizing the above objective, the student can mimic the output distribution of the teacher and make better predictions.

Although the initial concept of KD was simple and effective, many different logit-based methods have been proposed in recent years (Sau & Balasubramanian, 2016; Mirzadeh et al., 2020; Song et al., 2023b; Hao et al., 2023; Xu et al., 2023d; Lu et al., 2025). These methods mainly aim to normalize logits before distillation (Guo et al., 2020a; Chi et al., 2023; Miles & Mikolajczyk, 2024; Zheng & Yang, 2024; Sun et al., 2024a), soften

or smoothen the logits (Sau & Balasubramanian, 2016; Li et al., 2023a; Yuan et al., 2024b), or decouple the logit-distillation loss (Zhao et al., 2022a; Li et al., 2024a; Ding et al., 2024; Cui et al., 2024; Wei et al., 2024b; Liu et al., 2024d). Guo et al. (2020a) finds that the magnitude of confidence is not necessary for KD and proposes spherical KD to reduce the gap between teacher and student. Chi et al. (2023) shows that it is not feasible to soften all the samples with a constant temperature and proposes NormKD to customize the temperature for each sample according to the characteristics of the sample's logit distribution. Miles & Mikolajczyk (2024) and Zheng & Yang (2024) investigate the role of projectors and temperature in the KD process respectively. Sun et al. (2024a) suggests logit standardization by setting the temperature as the weighted standard deviation of the logit and performing a plug-and-play Z-score pre-process. On the other hand, among methods that try to soften the logits, Sau & Balasubramanian (2016) perturbs the logits for distillation with some noises, and Yuan et al. (2024b) softens the logits before KD and utilizes a learning simplifier with an attention module. However, most existing methods aim to decouple the distillation loss. For instance, Zhao et al. (2022a); Li et al. (2024a); Ding et al. (2024) separate the distillation of target-class and non-target class distributions. Wei et al. (2024b) reshapes the logits for multi-scale logit distillation, Liu et al. (2024d) uses different weights for distilling edges and bodies of objects, and Cui et al. (2024) decouples the KL to a weighted MSE and a cross-entropy loss that incorporates soft-labels.

In summary, logit distillation is a simple and straightforward method that can be likened to label smoothing (Kim & Kim, 2017) and regularization (Müller et al., 2019; Ding et al., 2019), aiming to enhance the student network's performance by preventing overfitting (Gou et al., 2021). However, due to the reliance of logits on predefined classes, its application is limited to supervised learning. Moreover, mimicking the last-layer predictions does not provide insights into the intermediate layers of the teacher network. It can be challenging for the student to learn effectively solely from the final outputs of the teacher, especially when teacher and student networks have different architectures. Therefore, in addition to logit distillation, there is a necessity to distill intermediate outputs and similarities from the teacher network to facilitate effective knowledge transfer.

### 2.2 Feature-based Distillation

DNNs are proven to extract different levels of features in their layers, and these multi-level representations from the intermediate layers of a network make them a valuable source for distillation, known as feature distillation. These intermediate representations provide step-by-step information that leads to the final prediction, which can be more valuable, especially in tasks like representation learning where the output of the model is a representation rather than logits.

The concept of feature distillation was first explored by Romero et al. (2014), where they proposed providing the student network with hints from the teacher. More specifically, they suggested minimizing the differences in features between the teacher and student to align them. Generally, let $F_s$ and $F_t$ be intermediate features of the teacher and student, respectively. A general feature distillation loss ($L_{fd}$) between teacher and student can then be defined as follows

$$L_{fd} = \ell_{feature}(\Phi_t(F_t), \Phi_s(F_s)) \tag{4}$$

where $\ell_{feature}$ is a similarity function for aligning the features, and $\Phi$ is a transformation function that matches the spatial size or number of channels between the features of the teacher and the student.

Inspired by the idea of FitNet, several feature-based distillation methods have been proposed. Initially, AT(Zagoruyko & Komodakis, 2016) introduced attention transfer by minimizing the $L_2$ norm of the difference between the feature maps of the teacher and student. Huang & Wang (2017) focused on matching the distributions of neuron selectivity patterns between the teacher and student by minimizing the Maximum Mean Discrepancy (MMD). Heo et al. (2019b) transferred the activation boundaries formed by hidden neurons, while Kim et al. (2018) transferred the factors of the features. Heo et al. (2019a) presented a comprehensive overhaul of feature distillation, proposing to distill the pre-activation feature maps, and Shu et al. (2021) introduced channel-wise distillation by applying softmax to channels of features and matching the distributions between teacher and student.

Table 2: Summary of logit-based and similarity-based distillation methods.

| Source | Sub-category | Description | Reference |
|---|---|---|---|
| Logit | Logit Normalization | Standardizes the logits before distillation | KD (Hinton, 2015), SphericalKD (Guo et al., 2020a), NormKD (Chi et al., 2023), Miles et al. (Miles & Mikolajczyk, 2024), TTM (Zheng & Yang, 2024), Sun et al. (Sun et al., 2024a) |
| | Logit Softening | Smooths the logits by adding noise | Sau et al. (Sau & Balasubramanian, 2016), Li et al. (Li et al., 2023a), SFKD (Yuan et al., 2024b) |
| | Decoupling | Splits the logit distillation loss into separate terms | DKL (Cui et al., 2024), BPKD (Liu et al., 2024d), SDD (Wei et al., 2024b), Ding et al. (Ding et al., 2024), NTCE (Li et al., 2024a) |
| | Revisiting Logit Distillation | Reduces the gap between teacher and student | TAKD (Mirzadeh et al., 2020), VanillaKD (Hao et al., 2023), EffDstl (Song et al., 2023b), Xu et al. (Xu et al., 2023d), BDCKD (Lu et al., 2025) |
| Similarity | Instance-level Similarity | Pairwise similarity between different data samples | MTKD (You et al., 2017), RKD (Park et al., 2019), SPKD (Tung & Mori, 2019), IRGKD (Liu et al., 2019f), Chen et al. (Chen et al., 2020b), CSKD (Chen et al., 2020f), Wang et al. (Wang et al., 2024g), Cheng et al. (Cheng et al., 2024a), GIRKD (Hu et al., 2024b) |
| | Feature/Channel-level Similarity | Pairwise similarity between features/channels | Yim et al. (Yim et al., 2017), ICKD (Liu et al., 2021c), CCD (Wang et al., 2023d) |
| | Class-level Similarity | Pairwise similarity between classes | CSKD (Chen et al., 2020f), DSD (Feng et al., 2021), DistKD (Huang et al., 2022a), IDD (Zhang et al., 2022d), AICSD (Mansourian et al., 2025), Dist+ (Huang et al., 2025a), IKD (Wang et al., 2025a) |
| | Others | Similarity between regions/ pixels/ neighbors/ views | SKD (Liu et al., 2019d), CIRKD (Yang et al., 2022c), PRRD (Wang et al., 2023b), BCKD (Wang et al., 2023k), NRKD (Xin et al., 2024), CRLD (Zhang et al., 2024g) |

Subsequently, some methods explored cross-layer feature distillation. SimKD (Chen et al., 2022a) employed a reused teacher classifier. Chen et al. (2021d) introduced Knowledge Review, where the feature maps of each layer of the teacher should match the student's feature maps in the corresponding layer and all features from the previous layers. Chen et al. (2021a) proposed a more general framework where each feature from the student should match all of the teacher's features, learning weights to determine the connection between the feature maps. This framework is a generalized version of FitNet, where the weights for corresponding features are set to one.

Recently, the focus of many papers has shifted towards the transformation function of the features, highlighting its role in feature matching beyond just matching feature sizes. Heo et al. (2019b) trains an autoencoder to transform the student's features for better alignment with the teacher's features. Yang et al. (2022f) randomly masks some pixels of the student's feature map and uses two convolution layers to transform the masked features, aligning them with the corresponding features in the teacher's model to encourage the student to produce feature maps similar to those of the teacher. Liu et al. (2023c) demonstrated that masking is not necessary and a simple Multi-Layer Perceptron (MLP) layer for transforming the student's features is sufficient.

Most recently, Huang et al. (2023b) combined the idea of diffusion (Ho et al., 2020) with KD, training a diffusion model on the teacher's feature maps and using it to denoise the student's feature maps. It considers the student's features as a noisy version of the teacher's features and employs diffusion as a transformation for

the student's features. In a concurrent work, Mansourian et al. (2024) used Convolutional Block Attention Mechanism (CBAM) (Woo et al., 2018) as a transformation, applying channel and spatial attention to the teacher's and student's features to highlight the most discriminative parts. It defines an MSE loss between the refined feature maps of the teacher and student.

Concurrent with the research path outlined, several other methods have been proposed. Liu et al. (2024g) rethinks raw feature alignment and decomposes the loss function of FitNets (Romero et al., 2014) into a magnitude difference term and an angular difference term. Yuan et al. (2024a) matches the features of the teacher with augmented versions of the student's features, while Zhang et al. (2024i) distills features in the frequency domain. Liu et al. (2023a) proposes N-to-one matching between the features of the student and teacher, respectively, and Passban et al. (2021) suggests that distilling corresponding features results in the disregard of some layers of the teacher. Instead, they propose to fuse all the features of the teacher for distillation.

In summary, feature-based distillation methods are more generalizable as the student can access information from the intermediate layers of the teacher, providing richer knowledge. While most state-of-the-art methods are feature-based and lead to improved performance, selecting the appropriate layers for distillation, aligning corresponding features from the teacher and student, and matching feature sizes pose significant challenges in feature distillation. This is especially true in cases where the teacher and student have different architectures, a common scenario with the recent growth of foundation models. In such cases, there can be a significant decrease in performance due to the capacity gap between the teacher and student networks. Furthermore, comparing different aspects of feature-based methods is not straightforward, as their performance can vary significantly depending on factors like the selection of layers, feature alignment, and feature transformation components. The sensitivity of these methods to such factors complicates their comparison and evaluation.

Table 3: Summary of feature distillation methods.

| **Method** | **Teacher Transformation** | **Student Transformation** | **Knowledge Type** | **Distillation Loss** |
|---|---|---|---|---|
| FitNet (Romero et al., 2014) | – | $1 \times 1$ Conv | Feature Representation | $\mathcal{L}_2(\cdot)$ |
| AT (Zagoruyko & Komodakis, 2016) | Channel Aggregation | Channel Aggregation | Attention Map | $\mathcal{L}_2(\cdot)$ |
| NST (Huang & Wang, 2017) | – | $1 \times 1$ Conv | Neuron Selectivity Pattern | $\mathcal{L}_{MMD}(\cdot)$ |
| Jacobian (Srinivas & Fleuret, 2018) | Gradient | Gradient | Jacobian of Feature | $\mathcal{L}_2(\cdot)$ |
| FT (Kim et al., 2018) | Auto-encoder | Auto-encoder | Paraphraser | $\mathcal{L}_1(\cdot)$ |
| AB (Heo et al., 2019b) | Binarization | $1 \times 1$ Conv | Activation Boundary | Marginal $\mathcal{L}_2$ |
| Heo et al. (Heo et al., 2019a) | Margin ReLU | $1 \times 1$ Conv | Pre-ReLU Feature | Partial $\mathcal{L}_2$ |
| He el al. (He et al., 2019) | Auto-encoder | $3 \times 3$ Conv | Adapted Feature | $\mathcal{L}_1(\cdot)/\mathcal{L}_2(\cdot)$ |
| FN (Xu et al., 2020b) | $L_2$ Normalization | $L_2$ Normalization | Feature Representation | $\mathcal{L}_{CE}(\cdot)$ |
| CWD (Shu et al., 2021) | Softmax$(\cdot)$ | Softmax$(\cdot)$ | Activation Map | $KL(\cdot)$ |
| MGD (Yang et al., 2022f) | – | Masking + Conv | Feature Representation | $\mathcal{L}_2(\cdot)$ |
| MLP (Liu et al., 2023c) | – | MLP | Feature Representation | $\mathcal{L}_2(\cdot)$ |
| DiffKD (Huang et al., 2023b) | – | Diffusion Process | Denoised Feature | $KL(\cdot)/\mathcal{L}_2(\cdot)$ |
| FAKD (Yuan et al., 2024a) | – | Feature Augmentation | Activation Boundary | $KL(\cdot)$ |
| LAD (Liu et al., 2024g) | – | $1 \times 1$ Conv | Angular Information of Feature | $\mathcal{L}_2(\cdot)$ |
| AttnFD (Mansourian et al., 2024) | Channel/Spatial Attention | Channel/Spatial Attention | Refined Feature | $\mathcal{L}_2(\cdot)$ |

### 2.3 Similarity-based Distillation

In addition to logit and feature distillation, similarity distillation is another form of distillation that equips the student with the structural knowledge and relationships derived from the data learned by the teacher. Rather than relying solely on exact predictions or feature values, similarity distillation transfers a higher order of knowledge by considering the pairwise similarities between features or instances.

One of the pioneering works was Yim et al. (2017), which introduced the Flow of Solution Procedure (FSP). This method involved computing the inner product between features from two layers in the teacher network and defining the squared L2 norm as the cost function for each pair of layers between the teacher and student networks. In a related work, You et al. (2017) suggested distilling the relative dissimilarity between intermediate representations of different examples. Subsequent studies proposed distilling the 8-neighborhood similarity of logits (Xie et al., 2018) and distilling similarity of features and logits among multiple teachers (Zhang & Peng, 2018).

In general, the similarity distillation loss ($L_{SD}$) is defined between a relation in the teacher and student as follows

$$L_{SD} = \ell_{similarity}(\phi(F_t^i, F_t^j), \phi(F_s^i, F_s^j)) \tag{5}$$

where, $F_t$ and $F_s$ represent the features/logits of the teacher and student, respectively, and $F^i$ and $F^j$ denote different features/logits from either two different layers or two different samples. The function $\phi$ calculates the similarity knowledge between the features/logits of the teacher and student, and $\ell_{similarity}$ is a correlation function that measures the similarities between the teacher and student.

Recently, similarity distillation has garnered increased attention, with numerous methods proposed, each focusing on different levels of similarity such as instance/sample-level similarity (Park et al., 2019; Tung & Mori, 2019; Liu et al., 2019f; Peng et al., 2019; Chen et al., 2020b; Cheng et al., 2024a; Hu et al., 2024b; Wang et al., 2024g; Xin et al., 2024), feature/channel-level similarity (Wang et al., 2020d; Liu et al., 2021c; Wang et al., 2023d), and class-level similarity (Chen et al., 2020f; Feng et al., 2021; Zhang et al., 2022d; Huang et al., 2022a; Mansourian et al., 2025; Huang et al., 2025a; Wang et al., 2025a). Remarkable works have emerged in instance-level methods. For example, RelationKD (Park et al., 2019) distills the distance and anglewise correlations between features of different samples. Tung & Mori (2019) adopts a similar approach by distilling the correlation of samples within a batch through the inner-product calculation of features of each sample. Liu et al. (2019f) leverages an instance relationship graph where features are represented as nodes and relations as edges for each sample. CIRKD (Yang et al., 2022c), a popular method, extends instance similarity from the batch-level by utilizing a memory bank to consider global relations between samples. Hu et al. (2024b) conducts similar work by distilling pairwise similarity relations based on stored features to reveal more complete instance relations.

In feature/channel-level similarity, IFVD (Wang et al., 2020d) and ICKD (Liu et al., 2021c) are two pioneering works. The former defines a prototype for each class and distills the distances of pixels from each prototype, while the latter creates a channel correlation matrix using the dot-product of a feature map with its transpose for channel-level similarity distillation. Wang et al. (2023d) compels the student to imitate the teacher by minimizing the distance between the channel correlation maps of the student and the teacher.

In class-level similarity, DistKD (Huang et al., 2022a; 2025a) is a popular method that significantly enhances the student's performance by distilling inter- and intra-class similarities. Feng et al. (2021) distills the pairwise similarity of classes by multiplying the logits with its reshaped version. Zhang et al. (2022d) suggests distilling position information and inter-class distances for semantic segmentation. AICSD (Mansourian et al., 2025) distills inter-class similarities by introducing intra-class distributions for each class and subsequently calculating pairwise similarity of these distributions using KL for semantic segmentation. IKD (Wang et al., 2025a) facilitates knowledge sharing within the same class to ensure consistent predictions.

In addition to the similarity levels mentioned, spatial similarity is also utilized for distilling pixel-to-pixel or pixel-to-region (Liu et al., 2019d; Yang et al., 2022c; Wang et al., 2023a;b;k; Zhang et al., 2024g) similarities.

Liu et al. (2019d) was among the first to propose pooling features into nine regions and distilling the similarity of these regions. CIRKD (Yang et al., 2022c) focused on pixel-to-pixel and pixel-to-region similarities across entire images. Wang et al. (2023b) transfers the multi-scale pixel-region relation, Wang et al. (2023k) distills correlations between adjacent blocks of the teacher, Wang et al. (2023a) employs patch-level similarity for distillation, and Zhang et al. (2024g) introduces distillation of the logits' similarity from local and global perspectives.

## 3 Schemes

After the knowledge source, the distillation scheme is one of the key choices in a teacher-student architecture. Based on the architecture of the teacher and its training mode, distillation methods can be categorized into three main schemes: offline distillation, online distillation, and self-distillation. In the offline scheme, which includes most of the distillation methods, the teacher is a larger pre-trained network. In the online scheme, the teacher and student are trained simultaneously, and in self-distillation, the teacher and student are the same network. Each of these schemes has its own applications: offline is preferred when a larger pre-trained teacher is available, online is most suitable when access to a pre-trained model is not available, and self-distillation plays an important role in self-supervised scenarios where labeled data is not available. Figure 4 shows the overall diagram of each scheme. In the following sections, the details of each scheme are explained.

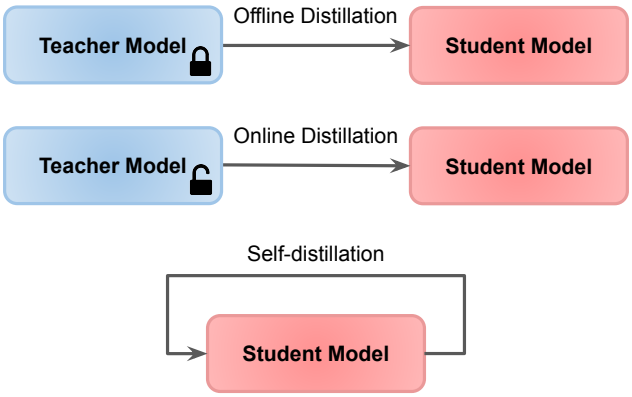


Figure 4: Illustration of distillation schemes. The teacher model in offline distillation is pre-trained and frozen, while in online distillation, the teacher and student are trained simultaneously. In self-distillation, the teacher and student are the same network.

### 3.1 Offline Distillation

The most straightforward approach to knowledge distillation is offline distillation, which consists of two stages. First, a large teacher model is pre-trained on a dataset. Then, during the training of the student model, the teacher's knowledge is transferred while its weights remain frozen.

Most existing distillation methods follow this offline approach, and the majority of the studies mentioned in the previous section adopt this technique. Offline distillation was first introduced in Hinton (2015), and numerous methods have since been proposed to extend and improve it. However, offline distillation has some limitations. One major challenge is the memory and computational overhead required to load a large teacher model, which can be a significant issue in resource-constrained environments. Additionally, since the teacher model remains fixed, the student model's performance is inherently constrained by the teacher's capabilities. Any biases present in the teacher can also be transferred to the student, potentially affecting its generalization (Guo et al., 2020a; Huang et al., 2022a).

To address these issues, incorporating adaptive distillation techniques alongside offline distillation can help improve student performance (Cho & Hariharan, 2019; Shu et al., 2021; Mansourian et al., 2025). Moreover,

in LLM distillation, where the teacher model is proprietary and only accessible via an API, offline distillation remains the only practical approach. In summary, despite its challenges, offline distillation continues to be the most widely used method in knowledge distillation due to its simplicity and effectiveness.

### 3.2 Online Distillation

Unlike offline distillation, the training of online distillation occurs in a single stage (Chen et al., 2020a). Wang et al. (2024a) showed that online distillation offers advantages by using a group of students as a virtual teacher, eliminating the need for a large pretrained model. Early approaches, such as Guo et al. (2020b) and Zhu et al. (2018), utilized an ensemble of multi-branch logits to create an on-the-fly teacher and trained the student accordingly. In addition, Wu & Gong (2021) extended this approach by using the mean of previous predictions as an auxiliary guide to enhance learning. Li et al. (2021b) introduced variation in branch depth, with each Hourglass network mimicking both the final heatmap and the fused heatmap of all branches. Yang et al. (2023b) proposed a framework for mutual contrastive learning, where contrastive embeddings from the same or different networks were used with the anchor during the training phase.

Recent works have focused on improving other aspects of online KD. Rao et al. (2023) explores methods to enhance KD by reducing the sharpness and variance gap in the logit space between the teacher and student. Zhang et al. (2023h) aims to achieve flatter loss minima to improve the generalization and stability of student models. Jacob et al. (2023) enhances multi-task networks by leveraging single-task networks as teachers and employing adaptive feature distillation with online task weighting. Similarly, Ypsilantis et al. (2024) applies this approach to multi-domain tasks using multiple teachers, where each teacher model specializes in a specific domain.

### 3.3 Self-distillation

Self-distillation is a unique form of online distillation where there is no larger pre-trained teacher; instead, the teacher and student are identical networks (Furlanello et al., 2018; Zhang et al., 2019b; Lan et al., 2019; Xu & Liu, 2019; Mobahi et al., 2020; Zhang et al., 2021c; Kim et al., 2021a). The concept was initially introduced in Furlanello et al. (2018), where a teacher is initially trained on labels, and subsequently, a new identical model is initialized from a different random seed and trained under the guidance of the former teacher. In a similar approach, Lan et al. (2019) first trains the network with labels and then utilizes labels and features of the model trained so far for subsequent training steps. Kim et al. (2021a) incorporates a linear combination of hard targets and predictions from the model in the previous epoch (teacher) to train the network in the current epoch (student). Zhang et al. (2019b) partitions the model into several sections and transfers knowledge from deeper sections to shallower ones, while Zhang et al. (2021c) employs multiple shallow classifiers at different layers and transfers knowledge from the deepest classifier to shallower ones. Mobahi et al. (2020) is the first to provide a theoretical analysis of self-distillation, and Zhang & Sabuncu (2020) establishes a theoretical link between self-distillation and label smoothing.

Recently, a variety of diverse self-distillation algorithms have been proposed (Yang et al., 2023g; Zhang et al., 2023b; Xiao et al., 2023b; Zheng et al., 2024b; Wu et al., 2024a; Chen et al., 2024a; Lin et al., 2024b; Li et al., 2024i; Liu et al., 2024e; Liang et al., 2024; Qin et al., 2025; Wang et al., 2025c), each offering a new perspective on the problem. Some methods leverage the deeper layers of the model as teachers to distill knowledge to shallower parts of the model (Liang et al., 2024; Qin et al., 2025). Other approaches utilize predictions from previous epochs as teachers to supervise the model in subsequent epochs (Liu et al., 2024e; Li et al., 2024i; Lin et al., 2024b). Chen et al. (2024a) and Zheng et al. (2024b) employ iterative pruning and discarding of parts of the model to create a student for self-distillation and MSKD(Wang et al., 2025c) proposes a self-distillation method for multi-label learning. Wu et al. (2024a) and Zhang et al. (2023b) integrate the concept of self-distillation with graphs, while Xiao et al. (2023b) combines data augmentation, deep contrastive learning, and self-distillation. In this method, different views of the same sample are input to the model for enhanced learning.

# 4 Algorithms

Aside from the source and scheme, different algorithms have been proposed to effectively transfer knowledge from the teacher to the student. In this section, the most important algorithms are reviewed. These include attention-based distillation, adversarial distillation, multi-teacher distillation, cross-modal distillation, graph-based distillation, adaptive distillation, and contrastive distillation.

## 4.1 Attention-based Distillation

Attention-based distillation focuses on transferring rich information from the teacher to the student using attention mechanisms, as attention can reflect neuron selectivity. Attention maps highlight crucial regions of the features in the teacher model and enable the student to automatically and effectively focus on mimicking the important information from the teacher. Instead of exact prediction matching, this approach allows the student to concentrate on the more critical regions of the image, similar to the teacher model.

The first work exploring attention distillation was Attention Transfer (AT) (Zagoruyko & Komodakis, 2016), which proposed the aggregation of feature channels in the student and teacher models. By minimizing the loss between attention maps of the teacher and student, the attention of the teacher can be effectively transferred to the student. Feng et al. (2021) transfers residual attention maps of features from two different layers, while An et al. (2022) employs a self-attention module to adaptively aggregate context information from the student and teacher feature maps. Yuan et al. (2024b) uses self-attention to simplify the learning process by softening the logits before distillation.

Several methods have focused on utilizing Class Activation Maps (CAM) (Zhou et al., 2016) or Gradient-Weighted Class Activation Maps (GradCAM) (Selvaraju et al., 2017) for attention distillation by transferring regions with more attention (Bavandpour & Kasaei, 2020; Cho & Kang, 2022; Guo et al., 2023d). Karine et al. (2024) applies channel self-attention and position self-attention on the features for distillation, Gou et al. (2023) employs different types of attention to transfer knowledge at various levels of deep representation learning for KD, and Guo et al. (2024b) computes attention using student features as queries and teacher features as key values, implementing sparse attention values through random deactivation. These sparse attention values are then used to reweight the feature distance of each teacher-student feature pair to prevent negative transfer.

Recently, powerful attention distillation methods have been proposed (Yang et al., 2023d; Huang et al., 2023b; Mansourian et al., 2024). DiffKD (Huang et al., 2023b) suggests training a diffusion model on the teacher's features and using it to denoise the student's features to highlight important areas for distillation. AttnFD (Mansourian et al., 2024) investigates the role of attention in distillation for semantic segmentation and employs CBAM (Woo et al., 2018) attention mechanism to refine features before distillation by considering both channel and spatial attention. This refinement focuses on transferring crucial information and emphasizes important regions, particularly enhancing the student network's performance in scenarios involving small objects.

## 4.2 Adversarial Distillation

A promising direction for leveraging teacher information during distillation is to embed the process within an adversarial framework inspired by the min-max game of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). Three distinct approaches emerge in this context. First, generative networks can be employed to synthesize the teacher's training data when the original data is unavailable. Second, a discriminator is integrated to provide refined guidance by distinguishing between teacher and student outputs during knowledge transfer. Third, adversarial training is applied to minimize discrepancies between teacher and student models when confronted with perturbed samples, thus pushing the student to learn from more challenging examples. Collectively, these methods aim to enhance the performance and robustness of the distilled student model, and the following sections will elaborate on these three approaches in detail.

#### 4.2.1 Adversarial Data-Free Knowledge Distillation

Adversarial distillation is applied in various ways, depending on the goal of the distillation process. One approach involves data-free distillation (Chen et al., 2019a; Ye et al., 2020; Fang et al., 2019; Choi et al., 2020; Liao et al., 2024; Patel et al., 2023; Wang et al., 2024o; Do et al., 2022; Zhao et al., 2022b; Zhou et al., 2024d; Liu et al., 2018; Micaelli & Storkey, 2019; Zhang et al., 2022b), where adversarial examples are generated in the absence of access to the teacher's original training data, often due to privacy concerns, transmission issues, or legal constraints (Fang et al., 2019). Adversarial samples are synthesized either to mimic the original dataset or to augment it, thereby enhancing the distillation process. Methods in this category (Chen et al., 2019a; Choi et al., 2020; Wang et al., 2024o; Ye et al., 2020; Zhao et al., 2022b; Micaelli & Storkey, 2019) often leverage the pre-trained teacher network and its internal statistics as a discriminator to guide the generator. Other techniques address distribution shifts in generated samples by proposing the use of an Exponential Moving Average (EMA) of the student as a more reliable learning source (Do et al., 2022) or framing the process as a meta-learning problem that balances knowledge acquisition and retention (Patel et al., 2023). Self-adversarial strategies (Zhou et al., 2024d) have also been explored to extend robustness in data-free settings. Lastly, Liao et al. (2024) investigated the challenges of data-free distillation with imbalanced datasets, where the teacher model is pre-trained on skewed class distributions, introducing class-aware strategies to mitigate this imbalance.

#### 4.2.2 Discriminator in the Loop

The use of a discriminator has proven effective in better aligning the prediction distributions between the teacher and student networks. Leveraging this intuition, one of the prevalent approaches in adversarial distillation involves incorporating a discriminator network to align the teacher and student logits (Xu et al., 2017; Liu et al., 2019a; Ghorbani et al., 2020; Croitoru et al., 2024) or intermediate feature representations (Liu et al., 2020a; Wang et al., 2020a; Chung et al., 2020; Wang et al., 2020d; Zhang et al., 2020b; Li et al., 2020b; Liu et al., 2019d; Shen et al., 2019d; Belagiannis et al., 2018; Wang et al., 2018c). Additionally, some methods employ a three-player framework for discrimination (Wang et al., 2019; 2018b), enhancing the effectiveness of the distillation process.

Adversarial distillation has also been extended beyond standard networks to the compression of GANs (Aguinaldo et al., 2019; Wang et al., 2020c; Li et al., 2020a; Zhang et al., 2022a), facilitating the distillation of heavy generator networks into more efficient ones. Notably, the discrimination module is not limited to the GAN framework and has been adapted for diffusion models to reduce the number of inference steps, significantly improving their inference efficiency (Sauer et al., 2024; Liu et al., 2024a; Kong et al., 2024; Lin et al., 2024c). Furthermore, advanced techniques have been introduced to utilize the score function instead of the final outputs or logits for distillation, leading to approaches such as adversarial score distillation (Wei et al., 2024a) and variational score distillation (Wan et al., 2024). These methods employ score matching and variational approximations to achieve improved robustness and efficiency in knowledge transfer.

#### 4.2.3 Adversarial Robust Knowledge Distillation

As previously discussed, KD transfers knowledge from a large, pre-trained teacher network to a smaller student network, yet it does not inherently guarantee robustness against adversarial attacks. Adversarial Robust Distillation (ARD) (Goldblum et al., 2020) addresses this limitation by combining adversarial training with distillation. In ARD, the objective is to minimize the discrepancy between the student's predictions and the teacher's predictions under adversarial perturbations, defined as follows

$$\min_{\theta} \mathbb{E}_{(X,y)\sim\mathcal{D}}\Big[\underbrace{\alpha\tau^2 D_{\mathrm{KL}}(S^{\tau}_{\theta}(X+\delta_{\theta}), T^{\tau}(X))}_{\text{Adversarially Robust Distillation loss}} + \underbrace{(1-\alpha)\ell(S^{t}_{\theta}(X), y)}_{\text{Classification loss}}\Big], \tag{6}$$

where $S$ and $T$ denote the student and teacher networks, $\tau$ is the temperature constant, and $\delta_\theta$ (computed as $\delta_\theta = \arg\max_{\|\delta\|_p<\epsilon} \ell(S^t_\theta(X), y)$) represents the adversarial perturbation.

Building upon ARD, several approaches have been developed to enhance the transfer of robust knowledge. One group refines teacher prediction reliability: RSLAD (Zi et al., 2021) replaces the classification loss in eq. (6) with a KL loss between teacher and student predictions to capitalize on robust soft labels from adversarially trained teachers, while introspective adversarial distillation (IAD)(Zhu et al., 2021a) and MTARD(Zhao et al., 2022e) further allow selective trust in teacher outputs and employ dual teachers for clean and adversarial scenarios, respectively. Another group enhances alignment and adaptability by synchronizing gradients and adaptively deriving perturbations through IGDM (Lee et al., 2023a) and AdaAD (Huang et al., 2023a); additional refinements are achieved in SmaraAD (Yin et al., 2024b) and IGKD-BML (Wang et al., 2021b) via Spearman Correlation, Class Activation Mapping, and attention-guided distillation to better emulate the teacher's decision-making process.

Advanced strategies tackle limitations in robustness transfer by introducing novel training schemes. DARWIN (Dong et al., 2024b) incorporates intermediate adversarial samples along with a triplet-based loss to balance natural, adversarial, and intermediate distributions, while PeerAiD (Jung et al., 2024) trains a peer model alongside the student for dynamic adaptation. Furthermore, DGAD (Park & Min, 2024) partitions the dataset into standard and adversarial distillation groups with consistency regularization to manage data imbalance, and both STARSHIP (Dong et al., 2024a) and TALD (Nguyen-Duc et al., 2023) further bolster robustness by transferring statistical attributes and sampling diverse adversarial examples via a Teacher Adversarial Local Distribution using Stein Variational Gradient Descent. These grouped techniques collectively advance the robustness and overall performance of student models in adversarial environments.

### 4.3 Multi-teacher Distillation

In contrast to typical distillation scenarios where the student is trained using a single teacher, multi-teacher algorithms aim to train the student network by amalgamating knowledge from multiple teachers. This approach has the potential to enhance the student's performance by leveraging an ensemble of teachers with diverse knowledge, thereby bolstering the student's generalization capabilities. This concept was initially proposed in Hinton (2015), which utilized the average of logits from multiple teachers as softened labels. Subsequent works have explored variations on this theme: You et al. (2017) employed a voting strategy, while Sau & Balasubramanian (2016) recommended perturbing the teacher's logits by adding noise, treating these perturbed outputs as distinct teachers that can act as a regularizer. TeKAP (Hossain et al., 2025) proposed an augmentation technique that generates multiple synthetic teacher knowledge by perturbing the knowledge of a single pretrained teacher. Fukuda et al. (2017) utilized a pool of teachers, randomly selecting a teacher for distillation each time. Papernot et al. (2016a) trained multiple teachers, with each focusing on a subset of the dataset, and employed a voting system to distill the knowledge of each expert teacher. Iordache et al. (2025) distilled knowledge from multiple teachers trained on distinct datasets. Furlanello et al. (2018) adopted a step-by-step strategy in which the student, at each stage, served as a teacher for the student in the subsequent step.

In recent years, a variety of multi-teacher methods have been introduced, each addressing distinct issues. Managing the knowledge aggregation adaptively poses a challenge due to the capacity gap between the student and each teacher (Li & Bilen, 2020; Wang et al., 2023e; Zhang et al., 2023a; Shang et al., 2023b; Lin et al., 2024d; Cheng et al., 2024b; Wang & Xu, 2024; Li et al., 2024e). Lin et al. (2024d) employs adaptive temperature and a diverse aggregation strategy to enhance distillation performance, while Cheng et al. (2024b) separates the vanilla KD loss and assigns adaptive weights to each teacher based on the entropy of their predictions. Wang & Xu (2024) leverages data relation knowledge to dynamically allocate weights to teachers, and Zhang et al. (2023a) employs a meta-network for weighting each teacher effectively.

Other methodologies have been proposed to effectively aggregate the knowledge of teachers for distillation (Luo et al., 2019; Chen et al., 2019b; Park & Kwak, 2020; Asif et al., 2020; Cao et al., 2023). Chen et al. (2019b) trains two distinct teachers, one for distilling intricate features and another for transferring general decision features. Park & Kwak (2020) and Asif et al. (2020) introduce additional branches to the student network for learning the features of teachers, while Cao et al. (2023) suggests a progressive approach for

distilling knowledge from multiple teachers. Another significant research avenue in multi-teacher distillation involves using ensemble pre-trained teachers, referred to as Knowledge Amalgamation (KA) (Shen et al., 2019a;b; Vongkulbhisal et al., 2019; Luo et al., 2020; Amirkhani et al., 2021; Xu et al., 2022b; Zhang et al., 2023c; Li et al., 2024b). Through KA, publicly available trained networks can be repurposed in the context of multi-teacher distillation. Some approaches involve employing multiple teachers, each trained on different datasets (Dong et al., 2024b) or tasks (Shen et al., 2019a; Luo et al., 2019). Shen et al. (2019b) initially extracts task-specific knowledge from heterogeneous teachers sharing the same sub-task and then combines this extracted knowledge to construct the student network. Li et al. (2024b) proposes a KA framework based on uncertainty suppression, and Vongkulbhisal et al. (2019) suggests training a unified classifier from pre-trained classifiers from each source along with an unlabeled set of generic data.

In summary, multi-teacher distillation furnishes the student network with a broader range of diverse and enriched information from multiple teachers, ultimately enhancing the student's generalization capabilities. Nonetheless, the effective aggregation of teachers' knowledge, determining the individual contributions of each teacher in distillation, and managing the complexity and computational costs are the primary challenges associated with multi-teacher distillation methods.

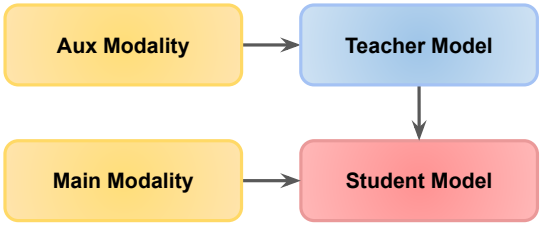


Figure 5: General overview of a cross-modal distillation method.

### 4.4 Cross-modal Distillation

In many scenarios, some modalities have rich annotations, while others lack sufficient labels. Cross-Modality Knowledge Distillation (CMKD) addresses this challenge by leveraging a well-annotated modality to improve the learning of a less-annotated one. This technique enables the model to benefit from the additional modality during training while remaining functional, even when that modality is unavailable at inference time. Gupta et al. (2016) introduced an innovative framework to transfer knowledge from a high-resource modality to a low-resource one, a method that has since been widely adopted in various studies. Figure 5 illustrates the overall concept of cross-modal distillation and Table 4 summarizes existing methods.

The early approach for CMKD was to use two parallel streams for two modalities during training, followed by a single stream during inference to produce feature maps similar to the teacher's modality.Garcia et al. (2018; 2019) applied this method to generate depth map from RGB. Later works focused on improving the imitation of the student's output to match the teacher's and reducing the gap between the modalities (Huo et al., 2024), with Thoker & Gall (2019) employing two student networks, each attempting to replicate the other's output. Wang et al. (2021d) adopted bidirectional distillation between events and frames.

The model proposed in Liu et al. (2021g) learned to generate 3D features from paired 2D inputs during the training phase and used them during inference. The structures in Hong et al. (2022b); Zhou et al. (2023b); Wang et al. (2023p); Klingner et al. (2023) consisted of separate encoders and decoders to extract features from LiDAR and RGB inputs, aiming to align the different features from the two modalities and distill spatial knowledge from LiDAR to RGB. Zhao et al. (2024b) used the same technique, replacing LiDAR with Radar, which provided lower spatial information, alongside RGB. Instead of using voxels for alignment, Zhang et al. (2024c) generated depth and semantic features from both networks and aimed to make them similar. Zhang et al. (2023i) employed a shared model to extract features from Thermal and RGB inputs in the student network, while using two heavier models for the teacher. The method proposed in Lee et al.

(2023b) decomposed the process into task-specific components and enabled the distillation of knowledge from optical flow effects on the right part of the network.

To distill knowledge between text and image modalities, Andonian et al. (2022) extracted features separately and trained them on aligned instances, using soft alignment to supervise the student network. Wang et al. (2024j) applied masking techniques and used text features to guide the network in predicting masked images. Sarkar & Etemad (2024) masked audio and video frames, attempting to reconstruct them to achieve a good representation for both modalities and align the two features effectively.

Table 4: Summary of Cross-modal Distillation Methods.

| **Method** | **Train Modality** | **Inference Modality** |
|---|---|---|
| Gupta et al.(Gupta et al., 2016) | RGB | Depth/Optical Flow |
| Garcia et al.(Garcia et al., 2019) | RGB + Depth | RGB |
| EvDistill(Wang et al., 2021d) | RGB + Event | Event |
| Hong et al.(Hong et al., 2022b)<br>DistillBEV(Wang et al., 2023p) | RGB + LiDAR | RGB |
| UniDistill(Zhou et al., 2023b) | RGB + LiDAR | LiDAR |
| Lee et al.(Lee et al., 2023b) | RGB + Optical Flow | RGB |
| Radocc(Zhang et al., 2024c) | Multi-view + LiDAR | Mult-view |
| CRKD(Zhao et al., 2024b) | Multi-view + LiDAR | Multi-view + Radar |
| Andonian et al. (Andonian et al., 2022)<br>CM-MaskSD (Wang et al., 2024j) | RGB + Text | RGB + Text |
| XKD(Sarkar & Etemad, 2024) | Video | RGB |

### 4.5 Graph-based Distillation

Graph-based knowledge distillation extends traditional KD methods by incorporating higher-order relational information between features or instances. Unlike approaches that primarily rely on pairwise similarity, graph-based methods utilize graph structures as a generic framework to capture intra-dependent relationships. In such methods, features or instances are represented as vertices, while the dependencies between them are modeled through edge weights. These graph structures encapsulate both local and global relationships, enabling the transfer of richer and more structured knowledge from teacher models to student models (Zhang & Peng, 2018; Chen et al., 2020b; Park et al., 2019; Tung & Mori, 2019; Peng et al., 2019; Liu et al., 2019f; Lee & Song, 2019; Passalis et al., 2020; Zhou et al., 2021a; Chen et al., 2021e; Ghorbani et al., 2021; Lee & Song, 2021). Furthermore, graph structures can be effectively employed to model the flow of information during distillation, particularly in scenarios involving multiple teacher models, by regulating the knowledge transfer process (Luo et al., 2018; Minami et al., 2020).

In recent years, the advent of Graph Neural Networks (GNNs) has garnered significant attention, particularly in the fields of data mining and knowledge graph modeling. GNN distillation is primarily utilized to achieve two main objectives (Tian et al., 2023a): enhancing the performance of the original teacher model (performance improvement) (Zhang et al., 2020a; 2021e; Chen et al., 2020e; Rezayi et al., 2021; Guo et al., 2023c; Huo et al., 2023; Zhu et al., 2024b; Yu et al., 2022a; Guo et al., 2022) or creating a lightweight version of the GNN (compression). The latter involves distilling a larger and more complex GNN into a compact and efficient GNN (Dong et al., 2023c; Yang et al., 2022b; He et al., 2022; Deng & Zhang, 2021) or even an MLP (Yang et al., 2021; Zhang et al., 2021d; Zheng et al., 2021; Tian et al., 2022a), making the distilled model suitable for real-time applications. To accomplish these goals, the information transferred between the teacher and the student typically includes logits (Guo et al., 2023c; Huo et al., 2023; Zhang et al., 2021e; 2020a; 2021d; Yang et al., 2021; Tian et al., 2022a; Dong et al., 2023c; He et al., 2022; Deng & Zhang, 2021), feature representations (Yu et al., 2022a; Huo et al., 2023; Rezayi et al., 2021; Zhang et al., 2020a; He et al., 2022), and the graph structure itself (Guo et al., 2022; Rezayi et al., 2021; Chen et al., 2020e; Zhang et al., 2021e; Zheng et al., 2021; Tian et al., 2022a; Yang et al., 2022b), which encapsulates the connectivity between the elements of the graph. These components collectively enable an effective transfer of structured

knowledge, ensuring that the distilled model retains essential characteristics while meeting its respective performance or efficiency objectives.

### 4.6 Adaptive Distillation

Adaptive distillation has recently garnered increased attention. By dynamically adjusting the parameters of the distillation process instead of maintaining a constant setup, the effectiveness of knowledge transfer can be enhanced. Table 5 summarizes adaptive distillation methods in different categories.

Adaptive distillation can manifest in various levels and forms, such as adaptive loss definition (Park & Heo, 2020; Wang et al., 2024g; Ma et al., 2024), adaptive loss weighting (Cho & Hariharan, 2019; Zhou et al., 2020; Mansourian et al., 2025), adaptive teacher pruning (Mirzadeh et al., 2020; Sarridis et al., 2025; Lee et al., 2024; Passalis et al., 2020; Guo et al., 2024a), and dynamically weighting different aspects of distillation based on sample importance (Huang et al., 2022b; Tian et al., 2022c; Sun et al., 2023b; 2024b; Kim et al., 2024c; Gou et al., 2024; Huang et al., 2025b).

One primary concern is that the teacher network may produce incorrect outputs that should not be directly transferred to the student. To address this issue, some approaches have introduced adaptive losses. Park & Heo (2020) suggests an adaptive Cross Entropy loss that replaces incorrect probability maps with ground truth labels. Wang et al. (2024g) proposes a calibrated mask to prevent the teacher model's erroneous representations from interfering with the student model's training, while Ma et al. (2024) integrates the positive prediction distribution of two teacher networks based on their correctness and cross-entropy magnitudes to provide a more accurate output distribution for guiding the student network.

Moreover, certain studies have delved into the effects of altering the distillation loss throughout the distillation process. Cho & Hariharan (2019) was among the first to explore strategies where reducing the influence of the teacher network on the student can enhance performance. Zhou et al. (2020) indicates that the impact of the distillation loss should diminish as training progresses, with the student gradually assuming control of the training process towards the end. Mansourian et al. (2025) introduces a weight decaying process to merge similarity distillation with vanilla distillation.

Given the capacity disparity between teacher and student models, some works aim to narrow this gap. A notable example is TAKD (Mirzadeh et al., 2020), which employs a teaching assistant network as an intermediary teacher to facilitate better knowledge comprehension by the student model. Similarly, Passalis et al. (2020) utilizes an auxiliary teacher model before distillation, while Lee et al. (2024) and Guo et al. (2024a) advocate for channel and feature pruning, respectively, before distillation. Sarridis et al. (2025) implements filter pruning to reduce channels and applies a curriculum learning strategy to distill layers from easy to challenging levels.

Lastly, certain methodologies adaptively conduct distillation based on the significance of a region or sample. Tian et al. (2022c) extracts detailed context-specific information from each training sample, Sun et al. (2024b) adaptively identifies confusing samples, and Kim et al. (2024c) adjusts the temperature parameter based on each sample's energy level. Huang et al. (2022b) incorporates adaptive weights for distilling each mask in feature distillation, while Gou et al. (2024) employs ground truth information to mask crucial regions for logit distillation. Sun et al. (2023b) calculates channel divergences between the teacher and student networks, converting them into distillation weights, and Kang et al. (2024) assigns greater weight to losses in layers where training discrepancies between the teacher and student models are more pronounced during distillation. CALD (Huang et al., 2025b) leverages a meta-network to generate class-adaptive weights for distillation.

### 4.7 Contrastive Distillation

Contrastive learning seeks to learn representations by contrasting similar (positive) and dissimilar (negative) samples in the feature space (Hadsell et al., 2006). In KD, most research employs feature-based contrastive learning approaches, where anchor, positive, and negative sets are defined using features extracted by teacher and student models as shown in Figure 6. The contrastive distillation loss can be formulated as

Table 5: Summary of adaptive distillation methods.

| Method | Reference |
|---|---|
| Adaptive Loss Definition | Park et al.(Park & Heo, 2020), Wang et al.(Wang et al., 2024g), CAG-DAKD (Ma et al., 2024), Dist+ (Huang et al., 2025a) |
| Adaptive Loss Scaling | Cho et al.(Cho & Hariharan, 2019), Zhou et al.(Zhou et al., 2020), AICSD(Mansourian et al., 2025) |
| Adaptive Teacher Pruning | TAKD(Mirzadeh et al., 2020), InDistill(Sarridis et al., 2025), Passalis et al.(Passalis et al., 2020), PCD(Lee et al., 2024), AFMPM(Guo et al., 2024a), ADG-KD(Li et al., 2025b) |
| Adaptive Sample Importance | MasKD(Huang et al., 2022b), APD(Tian et al., 2022c), HWD(Sun et al., 2023b), CS-KD (Sun et al., 2024b), Energy KD(Kim et al., 2024c), Gou et al.(Gou et al., 2024), CALD(Huang et al., 2025b) |

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(F_i^s, F_i^t)/\tau)}{\sum_{j \in \mathcal{N}} \exp(\text{sim}(F_i^s, F_j)/\tau)}, \tag{7}$$

where anchors ($F_i^s$) are typically features from the student's embeddings, positives ($F_i^t$) are corresponding features from the teacher for the same input, and negatives ($F_j$) are unrelated features from other data points. These methods employ contrastive loss to align the student's feature space with the teacher's, enabling efficient knowledge transfer. Other studies (Gao et al., 2021; Fan et al., 2023) innovate by introducing new contrastive loss functions that address challenges such as noisy negatives and capacity mismatches between teacher and student models, enhancing the robustness and effectiveness of the distillation process.

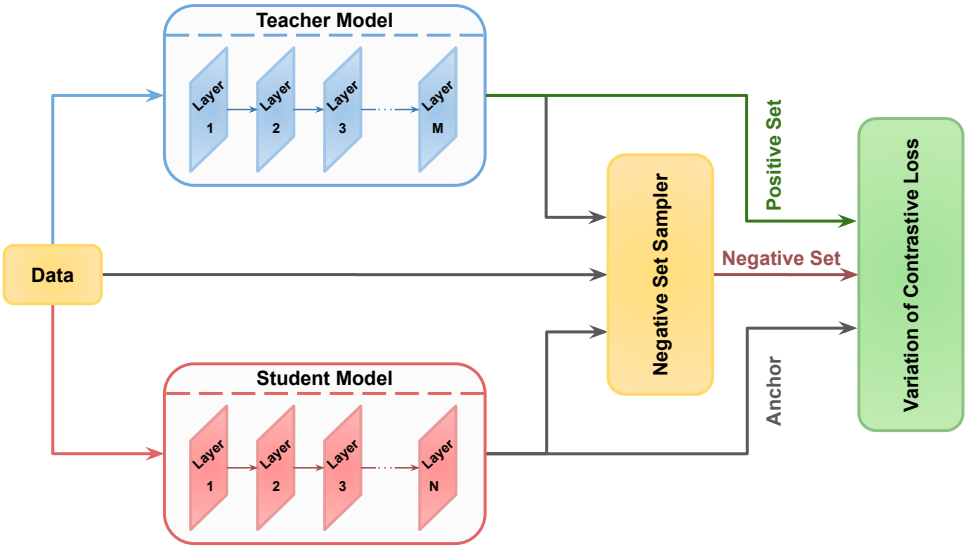


Figure 6: General overview of a contrastive distillation method. Embeddings of the teacher and student are considered as the positive set and anchor, respectively. The negative set can consist of embeddings from the teacher, student, or input data.

Contrastive distillation has gained significant attention for its ability to enhance knowledge transfer through contrastive learning techniques. (Tian et al., 2019) pioneered this approach by employing a contrastive loss to align the features of the teacher and student models, Expanding upon this foundation, WCoRD (Chen et al., 2021b) introduced a method that incorporates both primal and dual forms of the Wasserstein distance. While the dual form ensures global knowledge transfer by maximizing mutual information, the primal form

focuses on local feature alignment within mini-batches, providing a balanced approach to feature alignment and distillation.

In the field of semantic segmentation, Fan et al. (2023) addresses the challenge of dense pixel-level representations by leveraging contrastive loss to align the teacher and student feature maps. Similarly, Liu et al. (2024b) focuses on the domain of single-image super-resolution. Their approach utilizes contrastive loss to distill statistical information from intermediate teacher features into a lightweight student network, enabling high-resolution image reconstruction with minimal computational resources.

The problem of distilling large language-image pretraining models, such as CLIP (Radford et al., 2021), is tackled by Chen et al. (2024g), which develops a novel framework incorporating image feature alignment and educational attention modules. This method effectively aligns multi-modal features through contrastive loss. To address challenges in sentence embedding, DistilCSE (Gao et al., 2021) proposes a method which compresses large contrastive sentence embedding models.

In the domain of medical imaging, CRCKD (Xing et al., 2021) combines class-guided contrastive distillation and categorical relation preserving techniques to enhance intra-class similarity and inter-class divergence. The use of class centroids and relational graphs ensures effective knowledge transfer, even in imbalanced datasets.

In the area of weakly supervised visual grounding, Wang et al. (2021e) uses pseudo labels generated by object detectors. By applying contrastive loss for region-phrase matching, their method eliminates the need for object detection during inference, simplifying the process. Zhu et al. (2021b) tackles the problem of relation-based knowledge transfer with CRCD. By maximizing mutual information between anchor-teacher and anchor-student relation distributions, their framework aligns inter-sample relations through a relation contrastive loss. Most recently, CRD (Zhu et al., 2025) has leveraged contrastive learning to align class-wise semantics by preserving sample-wise similarity for intra-sample alignment, while capturing structural semantics for inter-sample contrast.

In conclusion, the incorporation of contrastive learning into distillation demonstrates exceptional capabilities in feature alignment, relational knowledge preservation, and the integration of local and global information. Proficiency of contrastive distillation in addressing challenges such as multi-modal alignment, dense feature mapping, and efficient model compression underscores its versatility. Moreover, the inherent flexibility of contrastive learning-based distillation allows it to effectively address critical issues, including imbalanced datasets, noisy labels, and computational constraints.

## 5 Modalities

Although most distillation methods have been proposed for computer vision tasks and work with images, KD has been effectively applied to other modalities, such as 3D/multi-view data, text, speech, and video. This section provides a review of distillation methods for each modality, categorizing them based on their respective tasks.

### 5.1 3D Input

Knowledge distillation techniques contribute significantly to enhancing various 3D-related tasks, including object detection, semantic segmentation, shape generation, and shape classification. This section explores the application of KD in 3D data, highlighting innovative approaches and categorizing contributions based on task domains. Figure 7 illustrates the key tasks, domains, and types of 3D data where KD techniques have been successfully applied. Table 6 summarizes distillation methods for each task based on the source of distillation.

#### 5.1.1 3D Object Detection

3D object detection involves identifying and localizing objects within three-dimensional spaces using data such as point clouds, voxel grids, and images from multiple angles. KD enhances this process by transferring knowledge from more complex models, leveraging various approaches and modalities, to simpler ones. This

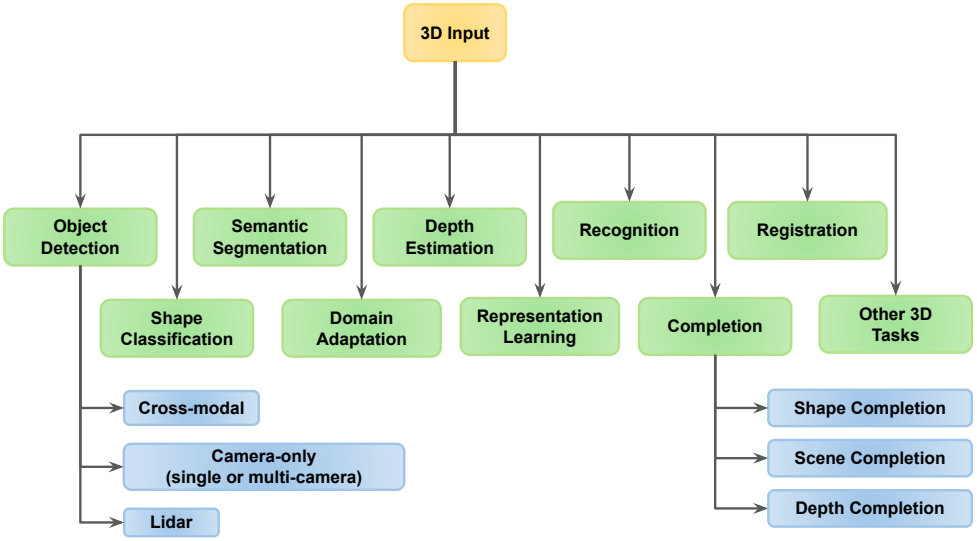


Figure 7: Overview of 3D domains and tasks.

approach improves accuracy and computational efficiency in tasks including autonomous driving and multi-camera detection. Recent advancements in KD-based 3D object detection have focused on three major areas: LiDAR-based, camera-only (single and multi-camera), and cross-modal, each addressing unique challenges in 3D perception.

**LiDAR:** Precise spatial and depth sensing are essential for autonomous systems, , yet challenges such as sparsity and occlusion often degrade performance in 3D object detection. To solve these, several KD frameworks have been proposed (Wei et al., 2022; Yang et al., 2022d; Zhang & Liu, 2023; Li et al., 2023g; Zhang et al., 2023d; Cho et al., 2023; Gambashidze et al., 2024; Bang et al., 2024; Huang et al., 2024e). X-Ray Distillation (Gambashidze et al., 2024) trains a teacher model on object-complete frames that are derived from aggregated LiDAR scans, to transfer knowledge to a student model that improves its ability to handle sparse data and occlusions. Similarly, RadarDistill (Bang et al., 2024) aligns radar and LiDAR features to handle sparsity and noise. Furthermore, RDD(Li et al., 2023g) reduces the mismatch between the teacher and student models by aligning feature representations and refining outputs.

In addition, PointDistiller (Zhang et al., 2023d) focuses on local geometric structures in point clouds, using dynamic graph convolution and reweighted learning to focus on important points and voxels. itKD (Cho et al., 2023) applies feature reduction and mutual refinement to transfer coarse and fine-grained geometric features. Additionally, SRKD (Huang et al., 2024e) bridges domain gaps caused by weather variations by aligning instance features through density and shape similarity and ensuring consistent predictions between the teacher and student models. LiDAR Distillation (Wei et al., 2022) also addresses domain gaps caused by varying LiDAR beam densities by performing distillation from a high-beam LiDAR teacher to a low-beam student.

Moreover, another study (Yang et al., 2022d) introduces two novel techniques: pivotal position logit KD, which focuses on key areas for enhanced distillation, and teacher-guided initialization, which transfers the teacher's feature extraction capabilities to the student through weight inheritance. However, intrinsic challenges such as sparsity, randomness, and varying density limit the effectiveness of normal distillation algorithms. To overcome these limitations, PVD (Zhang & Liu, 2023) leverages the advantages of both point-level and voxel-level knowledge to address these challenges in LiDAR object detection tasks.

**Camera-only (single or multi-camera):** Achieving accurate 3D object detection using only cameras is challenging due to the absence of depth information. To address this, FD3D (Zeng et al., 2023) uses selective masked generative distillation and query-based focal distillation to enhance object-specific learning in per-

Table 6: Summary of 3D distillation methods based on their tasks and sources of distillation.

| Task | Feature-based | Similarity-based | Logit-based |
|---|---|---|---|
| Object Detection | FD3D (Zeng et al., 2023), X3KD (Klingner et al., 2023), Gambashidze et al. (Gambashidze et al., 2024), PointDistiller (Zhang et al., 2023d), itKD (Cho et al., 2023), CRKD (Zhao et al., 2024b), DistillBEV (Wang et al., 2023p), RDD (Li et al., 2023g), RadarDistill (Bang et al., 2024), SRKD (Huang et al., 2024e) | FD3D (Zeng et al., 2023), X3KD (Klingner et al., 2023), Gambashidze et al. (Gambashidze et al., 2024), itKD (Cho et al., 2023), CRKD (Zhao et al., 2024b) | X3KD (Klingner et al., 2023), Gambashidze et al. (Gambashidze et al., 2024), PointDistiller (Zhang et al., 2023d), CRKD (Zhao et al., 2024b), DistillBEV (Wang et al., 2023p), RDD (Li et al., 2023g) |
| Classification | FAD (Lee & Wu, 2023), HSD (Zhou et al., 2024a) | – | FAD (Lee & Wu, 2023), JGEKD (Zheng et al., 2024a), PointViG-Distil (Tian et al., 2023b) |
| Segmentation | Liu et al. (Liu et al., 2021g), Qiu et al. (Qiu et al., 2023), CMDFusion (Cen et al., 2023), Jiang et al. (Jiang et al., 2023a), Smaller3d (Adamyan & Harutyunyan, 2023), Seal (Liu et al., 2024h) | PVD (Hou et al., 2022), Qiu et al. (Qiu et al., 2023) | Genova et al. (Genova et al., 2021), PSD (Zhang et al., 2021f), PVD (Hou et al., 2022), Qiu et al. (Qiu et al., 2023), LGKD (Yang et al., 2023f), Smaller3d (Adamyan & Harutyunyan, 2023), PartDistill (Umam et al., 2024) |
| Depth Estimation | MVP-Net (Chen et al., 2024b), LiRCDepth (Sun et al., 2025) | LiRCDepth (Sun et al., 2025) | MVP-Net (Chen et al., 2024b), KD-MonoRec (Xiao et al., 2024), LiRCDepth (Sun et al., 2025) |
| Representation | PPKT(Liu et al., 2021d), Fu et al.(Fu et al., 2022), SLidR(Sautier et al., 2022), PointCMT(Yan et al., 2022), RECON(Qi et al., 2023), MVNet(Yan et al., 2023), LiDAR2Map(Wang et al., 2023m), C2P(Zhang et al., 2023k), HVDistill(Zhang et al., 2024f), Yao et al.(Yao et al., 2024), Text4Point(Huang et al., 2024c), Diff3F(Dutt et al., 2024) | – | Yu et al.(Yu et al., 2022b), LiDAR2Map(Wang et al., 2023m), I2P-MAE(Zhang et al., 2023g) |
| Domain Adaptation | Cardace et al. (Cardace et al., 2023), Wu et al. (Wu et al., 2023e), SEN (Li et al., 2025a) | – | SEN (Li et al., 2025a) |
| Recognition | DistilVPR (Wang et al., 2024h), PointMCD (Zhang et al., 2023f) | – | – |
| Completion | SCPNet (Xia et al., 2023), RaPD (Fan et al., 2022), Lin et al. (Lin et al., 2024a), VPNet (Wang et al., 2025b), Huang et al. (Huang et al., 2024a) | ADNet (Kim et al., 2024b) | Hwang et al. (Hwang et al., 2021), MonDi (Liu et al., 2022b), Zhou et al. (Zhou et al., 2024b), Zhang et al. (Zhang et al., 2024e), HASSC (Wang et al., 2024i), Huang et al. (Huang et al., 2024a) |
| Registration | Jiang et al. (Jiang et al., 2024a) | – | DiReg (Löwens et al., 2024) |
| Other 3D Tasks | DTC123 (Yi et al., 2024b), CSD (Decatur et al., 2024), Van et al. (Van Vo et al., 2024) | Van et al. (Van Vo et al., 2024) | PCDNet (Huang et al., 2022c) |

spective and Bird's-Eye View (BEV) spaces. Similarly, X3KD (Klingner et al., 2023) integrates cross-task, cross-modal, and cross-stage distillation with techniques such as LiDAR feature alignment and adversarial training for dense supervision.

**Cross-modal:** Transferring spatial and depth knowledge from LiDAR-based models to cost-effective camera-radar or camera-only alternatives enhances multi-sensor systems, improving detection performance while resolving sensor inconsistencies. CRKD (Zhao et al., 2024b) transfers knowledge from LiDAR-camera teacher to a camera-radar student in the BEV space. Similarly, DistillBEV (Wang et al., 2023p) transfers 3D geometric knowledge from a LiDAR-based teacher to a multi-camera BEV student using region decomposition, adaptive scaling, spatial attention, and multi-scale distillation. In addition, CMKD (Hong et al., 2022b) transfers spatial knowledge from LiDAR-based teachers to monocular students, tackling depth estimation challenges using BEV features and teacher predictions. In addition UniDistill (Zhou et al., 2023b) proposes a universal KD framework which supports multiple modality pairs (LiDAR-to-camera, camera-to-LiDAR, fusion-to-camera, and fusion-to-LiDAR). The model employed BEV as a shared representation to align teacher and student detectors across modalities.

### 5.1.2 3D Shape Classification

Point cloud classification is another critical task where the teacher-student framework can be applied. FAD (Lee & Wu, 2023) introduces a novel loss function that combines logit distillation and feature distillation. JGEKD (Tian et al., 2023b) proposes a loss function based on joint graph entropy to address the challenges of non-independent and identically distributed 3D point cloud data in classification tasks. To improve the efficiency Zheng et al. (2024a) introduced an offline distillation framework that incorporates a negative-weight self-distillation approach. Zhou et al. (2024a) employs a self-distillation framework for incomplete point cloud classification.

#### 5.1.3 3D Semantic Segmentation

KD enhances the accuracy and robustness of point cloud segmentation by effectively transferring knowledge from a high-capacity teacher model to a lightweight student model (Genova et al., 2021; Liu et al., 2021g; Zhang et al., 2021f; Hou et al., 2022; Qiu et al., 2023; Yang et al., 2023f; Cen et al., 2023; Jiang et al., 2023a; Adamyan & Harutyunyan, 2023; Karine et al., 2024; Umam et al., 2024; Liu et al., 2024h).

In order to address challenges such as sparsity, randomness and varying density in segmentation, Hou et al. (2022) proposes point-to-voxel KD which distills the probabilistic outputs of the teacher at both the point level (fine-grained details) and voxel level (coarse structural information). Qiu et al. (2023) proposes a multi-to-single KD framework that fuses multi-scan information for hard classes and employs a multilevel distillation strategy, including feature, logit, and instance-aware similarity distillation.

Limited labeled data is another major challenge in the semantic segmentation of large-scale point clouds. Despite the task of annotating point-wise labels for such datasets being expensive and time-consuming, existing methods often rely on fully supervised approaches that require dense annotations. PSD (Zhang et al., 2021f) addresses this issue by introducing perturbed branches and leveraging graph-based consistency in a self-distillation manner. Seal (Liu et al., 2024h) is a self-supervised learning framework that distills knowledge from off-the-shelf vision foundation models for point cloud segmentation. PartDistill (Umam et al., 2024) proposes a method to distill knowledge from a vision-language model to a 3D segmentation network.

#### 5.1.4 3D Domain Adaptation

Unsupervised Domain Adaptation (UDA) struggles to perform effectively when the target domain differs from the labeled source domain, due to factors such as noise, occlusions, or missing elements. Cardace et al. (2023) proposes a self-distillation UDA technique to generate discriminative representations for the target domain and utilizes GNN-based online pseudo-label refinement. Wu et al. (2023e) introduces a cross-modal feature fusion in UDA for semantic segmentation. Li et al. (2025a) proposes a self-ensembling network for domain adaptation in 3D point clouds, which leverages a semi-supervised learning framework for effective knowledge transfer from a labeled source domain to an unlabeled target domain.

#### 5.1.5 3D Depth Estimation

3D depth estimation is crucial for accurately understanding the geometry of a scene, enabling perception systems to reconstruct spatial relationships and object structures. MVP-Net (Chen et al., 2024b) enhances reconstruction by using depth estimation in a multi-view, cross-modal distillation approach for better point cloud upsampling. KD-MonoRec (Xiao et al., 2024) leverages monocular RGB images as input and employs KD along with point cloud optimization to improve depth estimation. LiRCDepth (Sun et al., 2025) utilizes RGB images and radar point clouds and proposes a lightweight depth estimation model via KD.

#### 5.1.6 3D Representation Learning

Representation learning focuses on extracting meaningful representations from data. Collecting a large and high-quality annotated dataset in 3D is a costly and time-consuming process, and pre-training 3D models requires access to large-scale 3D datasets, which is not a trivial task. However, some works, such as DCGLR (Fu et al., 2022), have effectively utilized knowledge distillation (KD) with only 3D data to address this challenge. However, given that numerous large-scale datasets and powerful foundation models are available in the 2D domain, several methods (Liu et al., 2021d; Sautier et al., 2022; Yu et al., 2022b; Yan et al., 2022; Wang et al., 2023m; Zhang et al., 2023g; Yan et al., 2023; Yao et al., 2024; Zhang et al., 2024f; Dutt et al., 2024) have leveraged these capabilities to learn rich representations for 3D point clouds.

Liu et al. (2021d) introduces contrastive learning to transfer 2D semantic knowledge to 3D networks. SLidR (Sautier et al., 2022) introduces a self-supervised 2D-to-3D representation distillation framework that uses superpixel-driven contrastive loss to align image and LiDAR features, enabling effective pre-training of 3D perception models for autonomous driving. Qi et al. (2023) proposes a framework that can be used in either a single-modal or cross-modal setting.

3D representations can be learned with the help of text (Huang et al., 2024c) or through sequences of point clouds, as demonstrated in Zhang et al. (2023k), which develops a self-supervised method for learning 4D point cloud representations using KD.

#### 5.1.7 3D Recognition

Visual place recognition is an important task in computer vision and robotics, aiming to identify previously visited locations based on visual input such as camera image and LiDAR point clouds. PointMCD (Zhang et al., 2023f) and DistilVPR (Wang et al., 2024h) are two innovative cross-modal KD frameworks in this field, with PointMCD employing a multi-view framework, as discussed further in Section 5.2.

#### 5.1.8 3D Completion

Point cloud completion refers to the process of restoring missing geometric details. KD helps transfer learned priors from unpaired data, reducing the need for large annotated datasets and improving completion performance in data-scarce scenarios.

**Shape Completion:** Shape completion aims to reconstruct a complete point cloud from occluded input point cloud. RaPD (Fan et al., 2022) is the first semi-supervised point cloud completion method that utilizes prior distillation. Lin et al. (2024a) uses loss distillation and Zhou et al. (2024b) propose a hierarchical self-distillation point cloud completion method.

**Scene Completion:** LiDAR sensors often capture incomplete point clouds due to occlusions from objects or their surroundings, making large-scale 3D scene interpretation challenging. KD can enhance their performance as mentioned in Xia et al. (2023); Wang et al. (2024i); Zhang et al. (2024e); Wang et al. (2025b).

**Depth Completion:** Depth completion is the process of estimating a dense depth map from sparse or incomplete depth measurements, such as those obtained from LiDAR sensors. KD can aid in predicting missing depth information (Hwang et al., 2021; Liu et al., 2022b; Kim et al., 2024b; Huang et al., 2024a).

#### 5.1.9 3D Registration

KD can enhance point cloud registration by transferring knowledge from a large, complex model to a smaller, efficient one while preserving performance (Löwens et al., 2024; Jiang et al., 2024a).

#### 5.1.10 Other 3D Tasks

KD is applied to various other 3D tasks with different types of distillation. This includes generation (Yi et al., 2024b), stylization (Decatur et al., 2024), sampling (Huang et al., 2022c) and accordance detection, as in (Van Vo et al., 2024), which utilizes all types of KD.

### 5.2 Multi-view Input

In recent years, multi-view learning has emerged as a powerful approach for addressing complex tasks, such as 3D object detection and reconstruction. While leveraging information from multiple perspectives enables models to achieve more accurate and robust results, the diversity of data modalities and the challenge of transferring knowledge across them present significant challenges. To address these, several studies have explored novel KD techniques that allow models to share and refine knowledge across different views or modalities. This section reviews recent advancements in multi-view distillation methods, categorizing them by task and the modality. The existing works can broadly be divided into two main groups: 3D object detection and 3D shape recognition, with other tasks represented by individual studies.

#### 5.2.1 3D Object Detection

3D object detection has been a focal point in multi-view learning, with numerous studies proposing novel distillation techniques to enhance performance (Li et al., 2022c; Chen et al., 2022b; Klingner et al., 2023; Wang et al., 2023p; Jang et al., 2023; Zhao et al., 2024a; Jiang et al., 2024e). To this end, SimDistill (Zhao

Table 7: Summary of Knowledge Distillation in Multi-View Tasks.

| Method | Task | Teacher Modality | Student Modality | Short Description |
|---|---|---|---|---|
| KD-MVS(Ding et al., 2022b) | Multi-view Stereo Depth Reconstruction | Multi-view Images | Multi-view Images | Self-supervised MVS distillation for depth reconstruction. |
| MTS-Net(Tian et al., 2022b) | Multi-view Image Classification | Multi-view Images | Multi-view Images | Multi-view learning and distillation for image classification. |
| BEV-LGKD(Li et al., 2022c) | Multi-view BEV 3D Object Detection | LiDAR | Multi-view RGB Images | LiDAR-guided distillation with foreground masks and depth distillation for BEV perception. |
| BEVDistill(Chen et al., 2022b) | Multi-view 3D Object Detection | LiDAR | Multi-view Images | Cross-modal BEV distillation for unifying image and LiDAR features. |
| STXD(Jang et al., 2023) | Multi-view 3D Object Detection | LiDAR | Multi-view Images | Structural and temporal distillation with cross-correlation regularization and response distillation. |
| PointMCD(Zhang et al., 2023f) | 3D Shape Recognition | Multi-view Images | 3D Point Cloud | Multi-view cross-modal distillation for 3D shape recognition. |
| Acar et al.(Acar et al., 2023) | Vision-based Robotic Manipulation | Multi-view Images | Single-camera | Distillation from multi-camera to single-camera for robust manipulation. |
| Zhang et al.(Zhang et al., 2023e) | Multi-view BEV Detection | Multi-view BEV Images | Multi-view BEV Images | Structured distillation for efficient BEV detection. |
| MVKD(Wang et al., 2023c) | Semantic Segmentation | Multi-view Images | Single-view Images | Multi-view distillation with co-tuning and feature/output distillation losses. |
| MKDT(Lin & Tseng, 2023) | Human Action Recognition | Multi-view Videos | Single-view Videos | Multi-view distillation with hierarchical vision transformer for incomplete data. |
| 3X3KD(Klingner et al., 2023) | Multi-camera 3D Object Detection | LiDAR | Multi-view Images | Cross-modal distillation for multi-camera 3D object detection. |
| DistillBEV(Wang et al., 2023p) | 3D Object Detection for Autonomous Driving | LiDAR | Multi-view BEV Images | Distillation from LiDAR to BEV for enhanced 3D perception. |
| FSD-BEV(Jiang et al., 2024e) | 3D Object Detection for Autonomous Driving | Self-distillation | Self-distillation | Foreground self-distillation for single-model distillation. |
| SimDistill(Zhao et al., 2024a) | Multi-modal 3D Object Detection | LiDAR-camera Fusion | Multi-view Images | Distillation for 3D object detection with modality gap compensation. |
| GMViT(Xu et al., 2024b) | 3D Shape Recognition | Multi-view Images | Multi-view Images | Knowledge distillation with GMViT for efficient 3D shape analysis. |
| MT-MV-KDF(Qiang et al., 2024) | Myocardial Infarction Detection and Localization | Multi-view ECG | Single-view ECG | Multi-task, multi-view distillation with CNN and attention modules. |

et al., 2024a) introduces a LiDAR-camera fusion-based teacher and a simulated fusion-based student for multi-modal learning. By maintaining identical architectures and incorporating a geometry compensation module, the student learns to generate multi-modal features solely from multi-view images.

Klingner et al. (2023) proposes a comprehensive framework for multi-camera 3D object detection by leveraging cross-task and cross-modal distillation. Cross-task distillation from an instance segmentation teacher avoids ambiguous error propagation, while cross-modal feature distillation and adversarial training refine 3D representations using a LiDAR-based teacher. Cross-modal output distillation further enhances detection accuracy.

Wang et al. (2023p) proposes a KD framework for aligning a BEV-based student's features with a LiDAR-based teacher's, addressing the depth and geometry inference issue compared to LiDAR-based methods.

Cross-modal KD often suffers from feature distribution mismatches. FSD scheme (Jiang et al., 2024e) eliminates the need for pre-trained teachers and complex strategies.

BEV-LGKD (Li et al., 2022c) proposes a LiDAR-guided KD framework for multi-view BEV 3D object detection. By transforming LiDAR points into BEV space and generating foreground masks, the method guides RGB-based BEV models without requiring LiDAR at inference. Depth distillation is also incorporated to improve depth estimation, enhancing BEV perception performance.

BEVDistill (Chen et al., 2022b) introduces a cross-modal BEV distillation framework for multi-view 3D object detection by unifying image and LiDAR features in BEV space and adaptively transfers knowledge across non-homogeneous representations in a teacher-student paradigm.

STXD (Jang et al., 2023) proposes a structural and temporal cross-modal distillation framework for multi-view 3D object detection. It reduces redundancy in the student's feature components by regularizing cross-correlation while maximizing cross-modal similarities. Additionally, it encodes temporal relations across frames using similarity maps and employs response distillation to enhance output-level knowledge transfer.

#### 5.2.2 3D Shape Recognition

3D shape recognition is another critical area where multi-view distillation has been applied to bridge the gap between 2D and 3D representations (Zhang et al., 2023f; Xu et al., 2024b). PointMCD (Zhang et al., 2023f) bridges 2D visual and 3D geometric domains through a multi-view cross-modal distillation framework, by distilling knowledge from the teacher to a point encoder student. Group Multi-View Transformer (GMViT) (Xu et al., 2024b) addresses the limitations of view-based 3D shape recognition methods, which often struggle with large model sizes that are unsuitable for memory-limited devices. To overcome this, GMViT introduces a large high-performing model designed to enhance the capabilities of smaller student models

#### 5.2.3 Other Tasks

In addition to 3D object detection and 3D shape recognition, multi-view distillation techniques have been applied to a range of other tasks, including multi-view image classification (Tian et al., 2022b), vision-based robotic manipulation (Acar et al., 2023), multi-view BEV detection (Zhang et al., 2023e), multi-view stereo depth reconstruction (Ding et al., 2022b), semantic segmentation (Wang et al., 2023c), human action recognition (Lin & Tseng, 2023), and medical applications (Qiang et al., 2024). These studies are summarized as follows:

MTS-Net (Tian et al., 2022b) integrates KD with multi-view learning, redefining the roles of the teacher and student models. This end-to-end approach optimizes both view classification and knowledge transfer, with extensions like MTSCNN refining multi-view feature learning for image recognition.

In robotic manipulation, Acar et al. (2023) improves vision-based reinforcement learning by transferring knowledge from a teacher policy trained with multiple camera viewpoints to a student policy using a single viewpoint.

Zhang et al. (2023e) uses a spatio-temporal distillation and BEV response distillation by aligning the student's outputs with those of the teacher, addressing the computational complexity of multi-view BEV detection methods.

Supervised multi-view stereo methods face challenges due to the scarcity of ground-truth depth data. The self-supervised KD-MVS framework (Ding et al., 2022b) uses a teacher trained with photometric and featuremetric consistency to probabilistically transfer knowledge to a student.

MVKD (Wang et al., 2023c) introduces a multi-view KD framework for efficient semantic segmentation. The framework aggregates knowledge from multiple teacher models and transfers it to a student model. To ensure consistency among the multi-view knowledge, MVKD employs a multi-view co-tuning strategy. Additionally, it proposes multi-view feature distillation loss and multi-view output distillation loss to effectively transfer knowledge from multiple teachers to the student.

MKDT (Lin & Tseng, 2023) introduces a multi-view KD transformer framework for human action recognition, addressing the challenge of incomplete multi-view data. The framework consists of a teacher network and a student network, both utilizing a hierarchical vision transformer with shifted windows to effectively capture spatio-temporal information.

MT-MV-KDF (Qiang et al., 2024) introduces a novel multi-task multi-view KD framework for myocardial infarction detection and localization. Multi-view learning extracts ECG feature representations from different views, while multi-task learning captures similarities and differences across related tasks.

In summary, multi-view distillation techniques have shown significant potential across a broad range of tasks, from 3D object detection and shape recognition to specialized applications such as robotic manipulation and medical diagnostics. The variety of approaches, as highlighted in Table 7, emphasizes the adaptability of these methods to various modalities and challenges.

### 5.3 Text Input

Knowledge distillation has been widely applied across various NLP tasks. This technique is particularly valuable in NLP due to the large scale of state-of-the-art models, namely BERT and GPT. While these

models are highly accurate, they can be impractical for deployment in resource-constrained environments. By distilling the rich representations learned by these large models into lightweight versions, KD helps maintain high accuracy in tasks such as neural machine translation, question answering, text generation, event detection, document retrieval, text recognition, named entity recognition, text summarization, natural language understanding, sentiment analysis, and text classification.

**Neural Machine Translation (NMT)**: While NMT has significantly outperformed traditional statistical methods, large models like transformers are computationally intensive. KD addresses this challenge by distilling the teacher's output to the student, achieving comparable translation quality with reduced computational costs. This enables the development of more efficient and scalable NMT solutions (Kim & Rush, 2016; Tan et al., 2019; Wang et al., 2021a; Liang et al., 2022a; Jin et al., 2024)

**Question Answering (QA)**: In QA, KD transfers both answer predictions and intermediate representations, such as attention distributions, from a large teacher model to a smaller student model. This process helps maintain high accuracy while minimizing computational costs, making QA systems more efficient and suitable for real-world applications (Hu et al., 2018; Izacard & Grave, 2020; Yang et al., 2020; Liu et al., 2021b; Yin et al., 2024a).

**Text Generation:** In text generation, KD enables the transfer of a teacher's generative capabilities to a smaller student model, training it to mimic the vocabulary distributions and language patterns of the teacher. Chen et al. (2019c) focuses on leveraging BERT's contextual understanding for efficient text generation, while Haidar & Rezagholizadeh (2019) enhances fluency and diversity by combining distillation with GANs.

**Event Detection:** KD is applied in event detection to distill knowledge from larger, more complex models to smaller, more efficient ones. Ren et al. (2021) improves event detection across languages by distilling knowledge from high-resource to low-resource models, while Yu et al. (2021) uses distillation to retain learned knowledge while adapting to new events. Meanwhile, Liu et al. (2019a) employs adversarial imitation to enhance detection accuracy.

**Document Retrieval:** In document retrieval, KD transfers the ranking and matching capabilities of large teacher models to smaller student models, typically replicating relevance scores and interaction features (Shakeri et al., 2019; Chen et al., 2021c).

**Text Recognition:** KD in text recognition facilitates the transfer of both visual and contextual features from a teacher to a student, allowing the student to maintain high recognition accuracy with reduced computational cost (Bhunia et al., 2021; Wang & Du, 2021).

**Named Entity Recognition (NER):** In NER, KD helps a smaller student model learn from the teacher's contextual understanding and classification abilities (Liang et al., 2021; Ge et al., 2024).

**Text Summarization:** KD in text summarization transfers summarization capabilities from teacher to student, training the student to replicate output sequences and probability distributions (Chen et al., 2017; Liu et al., 2020b; Nguyen & Luu, 2022).

**Natural Language Understanding (NLU):** KD in NLU transfers the rich linguistic understanding of the teacher models to the student. In NLU, the student learns to replicate the teacher's output probabilities and internal representations across various tasks, ensuring high performance while reducing computational costs (Liu et al., 2019c; Tang et al., 2019; FitzGerald et al., 2022).

**Sentiment Analysis:** In sentiment analysis, KD distills the sentiment classification capabilities of the teacher to the student, training it to mimic the teacher's probability distributions and decision patterns (Lehečka et al., 2020; Malki, 2023).

**Text Classification**: In text classification, KD transfers classification capabilities from the teacher to the student, enabling the student to mimic the teacher's feature representations and output probabilities (Xu & Yang, 2017; Li & Li, 2021).

### 5.4 Speech Input

Deep neural acoustic models have achieved remarkable success in speech recognition tasks; however, their complexity poses challenges for real-time deployment on devices with limited computational resources. As the demand for efficient and rapid processing increases on embedded platforms, conventional large-scale models often increases on embedded platforms. To this end, KD has garnered significant attention in various speech-related tasks, including speech recognition, speech enhancement, speaker recognition and verification, speech translation, speech synthesis, speech separation, spoken language identification and understanding, deepfake speech and spoofing detection, audio classification and tagging, spoken question answering and conversational AI, and audio captioning and retrieval.

**Speech Recognition (ASR):** KD is important in enhancing ASR models by transferring expertise from high-precision teacher models to student models, supporting rapid inference with minimal loss in accuracy. This also improves domain adaptation, making ASR effective in poor acoustic environments or in expert domains, including medical transcription (Chebotar & Waters, 2016; Markov & Matsui, 2016; Takashima et al., 2018; Mun'im et al., 2019; Kurata & Saon, 2020; Yoon et al., 2021; You et al., 2021a; Yoon et al., 2022; Zhao et al., 2022c; Kim et al., 2023b; Park et al., 2023; Zhang et al., 2023j).

**Speech Enhancement:** KD in speech enhancement enables denoising through high-fidelity speech representations and model compression (Watanabe et al., 2017; Wu et al., 2019; Kim & Kim, 2021; Shin et al., 2023; Park et al., 2024).

**Speaker Recognition and Verification:** In speaker recognition and verification, KD helps reduce model complexity while preserving speaker identity features, allowing for fast and secure authentication. It also enhances robustness against adversarial attacks and cross-domain variations, improving performance in noisy or multilingual environments (Mingote et al., 2020; Peng et al., 2022; Liu et al., 2022a; Jin et al., 2023b; Mingote et al., 2023; Truong et al., 2024; Hoang et al., 2024).

**Speech Translation:** KD improves the efficiency of multilingual translation models by enabling knowledge sharing between high-resource and low-resource languages (Liu et al., 2019e; Gaido et al., 2020; Inaguma et al., 2021; Lei et al., 2023). This enhances translation fluency while maintaining low latency for real-time applications.

**Speech Synthesis (Text-to-Speech):** In speech synthesis, KD reduces the complexity of deep generative models while retaining natural-sounding speech output (Oord et al., 2018; Xu et al., 2020a; Li et al., 2024f). It also helps distill expressive features, making synthetic voices more human-like with lower computation costs.

**Speech Separation:** KD compresses complex separation models into smaller networks while preserving speaker distinction (Chen et al., 2018; Zhang et al., 2021b). It enables real-time speaker and cocktail-party effect reduction on low-power devices.

**Spoken Language Identification and Understanding:** In this task, KD enables language adaption in low-resource languages through the transfer of information between high-resource multilingual teacher models (Shen et al., 2018; 2019c; 2020; Kim et al., 2021b; Cappellazzo et al., 2022; Mao & Zhang, 2023; Cappellazzo et al., 2023). As a result, such mechanism empowers lighter, accelerated models to achieve high performance in language identification and speech dialogue comprehension.

**Deepfake Speech and Spoofing Detection:** With growing concerns about synthetic voice fraud, the need for reliable approaches to deepfake detection has increased. KD strengthens anti-spoofing by transferring knowledge between DNNs trained on a range of deepfake speech datasets and lightweight detection architectures (Liao et al., 2022; Xue et al., 2023; Ren et al., 2023c;b; Lu et al., 2024; Wani & Amerini, 2025).

**Audio Classification and Tagging:** KD strengthens feature learning relevant to sound event detection, enhancing model efficiency and allowing deployment on embedded platforms such as IoT and edge AI platforms (Chang et al., 2020; Yin et al., 2021; Schmid et al., 2023; Liang et al., 2022b; Dinkel et al., 2023; Tang et al., 2024).

**Spoken Question Answering and Conversational AI:** KD enables efficient question-answer through the distillation of contextual information in deep, lengthy transformer models, mapping them onto efficient, quick-answer-producing models (You et al., 2020a;b; 2021a;b).

**Audio Captioning and Retrieval:** In this task, KD enables deep audio-text embedding models to be compressed while maintaining generalization, making them suitable for efficient multimedia search architectures (Xu et al., 2024c; Primus & Widmer, 2024).

Table 8: Summary of video knowledge distillation methods based on their sources of distillation.

| Category | Method |
|---|---|
| **Feature-based** | Wang et al.(Wang et al., 2024c), Liu et al.(Liu et al., 2019b), V2I-DETR (Jiang et al., 2024c), MaskAgain(Li et al., 2023e) |
| **Similarity-based** | RSKD(Wang & Tang, 2023), Bhardwaj et al.(Bhardwaj et al., 2019), OOKD (Kim et al., 2024a), DL-DKD(Dong et al., 2023a) |
| **Logit-based** | Camarena et al.(Camarena et al., 2024), Wu et al.(Wu & Chiu, 2020), PKD (Soufleri et al., 2024), MobileVOS(Miles et al., 2023), V2I-DETR(Jiang et al., 2024c), DTO(Monti et al., 2022), MVD (Wang et al., 2023l), RankDVQA-mini(Feng et al., 2024a), VideoAdviser (Wang et al., 2023o), Afouras et al.(Afouras et al., 2020), Perez et al.(Perez et al., 2020) |

### 5.5 Video Input

Knowledge distillation has been widely applied in image and language tasks, but its potential in video-based applications is gaining increasing attention. Recent studies have explored KD for video action recognition (Liu et al., 2019b; Wu & Chiu, 2020; Camarena et al., 2024; Wang et al., 2024c; Soufleri et al., 2024), video classification (Bhardwaj et al., 2019; Wang & Tang, 2023), video object segmentation (Miles et al., 2023; Jiang et al., 2024c), video instance segmentation (Kim et al., 2024a), Partially Relevant Video Retrieval (PRVR) (Dong et al., 2023a), trajectory forecasting (Monti et al., 2022), and representation learning through masked video modeling (Li et al., 2023e; Wang et al., 2023l). Table 8 summarizes video distillation methods.

The main challenge in deep learning models, particularly in applications such as video quality estimation (Feng et al., 2024a) and attention prediction (Fu et al., 2020), is the high computational cost and complexity of large models, which restrict their execution on edge devices. KD effectively addresses this by enabling a smaller, more efficient model (student) to learn from a larger, complex model (teacher) with minimal performance reduction while significantly decreasing computational and memory requirements. MobileVOS (Miles et al., 2023) specifically focuses on semi-supervised video object segmentation on resource-constrained devices, for example mobile phones, and proposes a pixel-wise representation distillation loss to efficiently transfer structural information from the teacher to the student while ensuring feature consistency across frames. In Kim et al. (2024a), the authors propose a real-time approach that distills similarity information from an offline teacher model, which processes entire video sequences, to an online model, which processes individual frames. Mullapudi et al. (2019) introduces an online approach that applies feature distillation on live videos for low-cost semantic segmentation.

Furthermore, KD is applicable in multi-modal learning, enabling cross-modal knowledge transfer, such as video-to-image (Jiang et al., 2024c), video-to-text (Wang et al., 2023o), text-to-video (Dong et al., 2023a), audio-to-video (Afouras et al., 2020), and video-to-audio (Perez et al., 2020). Notably, (Pan et al., 2020) leverages KD in a spatio-temporal graph model for the video captioning task, making the model more robust against spurious correlations. V2I-DETR enhances medical video lesion detection by distilling temporal knowledge from a video-based teacher model to an image-based student model, achieving real-time inference speed with high accuracy (Jiang et al., 2024c).

# 6 Applications

Knowledge distillation can be applied across various fields and has important applications, including distillation in LLMs, distillation in foundation models and in vision transformers, distillation in self-supervised learning, distillation in diffusion models, and distillation in visual recognition tasks. The following provides a detailed explanation of each application.

## 6.1 Distillation in Large Language Models

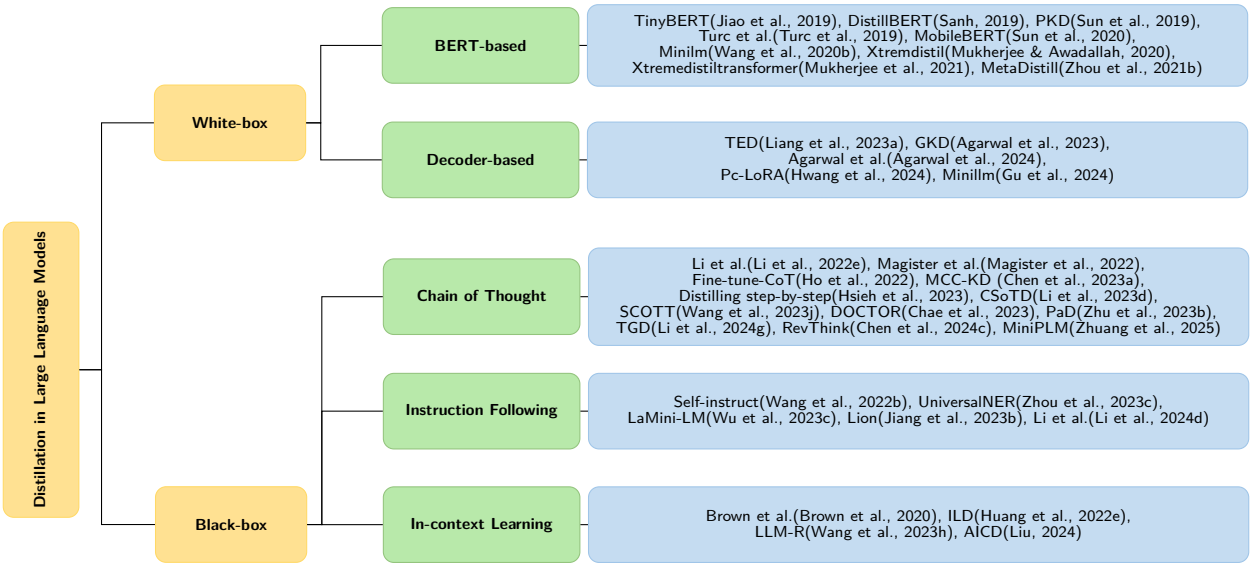


Figure 8: Classification of distillation methods for large language models.

LLMs have recently revolutionized the field of NLP, with significant applications across various domains in both academia and industry. However, these models pose critical challenges due to their large number of parameters, which limits their deployment in real-time scenarios. Additionally, these giant models, with a considerable number of attention blocks, have been shown to be overparameterized (Kaplan et al., 2020; Fischer et al., 2024). Therefore, KD plays a crucial role in compressing LLMs into Small Language Models (SLMs). Existing KD methods for LLMs are mainly classified into two major categories: white-box distillation and black-box distillation. In white-box KD, logits and intermediate outputs of the LLMs are accessible, enabling the student to benefit from the internal information of the teacher. This information can be either logits (Hinton, 2015) or intermediate features (Romero et al., 2014), such as outputs of each attention block or attention scores. On the other hand, in black-box distillation, only the predictions of the teacher are available for distillation, which is common in many closed-source LLMs like GPT-4 (Achiam et al., 2023) and Gemini (Team et al., 2023), limiting the distillation process to only the predictions. Black-box methods are categorized into In-Context Learning (ICL) (Huang et al., 2022e), Chain of Thought (CoT) (Li et al., 2022e), and Instruction Following (IF) (Wang et al., 2022b). Figure 8 summarizes the black-box and white-box distillation methods for LLMs.

Most white-box KD methods focus on the BERT (Kenton & Toutanova, 2019) model (Turc et al., 2019; Jiao et al., 2019; Sanh, 2019; Sun et al., 2019; 2020; Wang et al., 2020b; Mukherjee & Awadallah, 2020; Mukherjee et al., 2021). DistilBERT (Sanh, 2019) was one of the works that initialized the student with a pre-trained teacher and minimized the soft probabilities between the student and the teacher during the pre-training phase. In a concurrent work, MiniLM (Wang et al., 2020b) proposed distilling the self-attention module of the last transformer layer of the teacher. MobileBERT (Sun et al., 2020) first trains a special teacher and then uses feature and attention transfer to distill knowledge to the student. TinyBERT (Jiao et al.,

2019) introduced a KD method for both pre-training and task-specific training of a smaller version of BERT by distilling the logits, attention matrices, hidden states, and the output of the teacher's embedding layer. PKD (Sun et al., 2019) introduces two strategies for incremental distillation, while Mukherjee & Awadallah (2020) and Mukherjee et al. (2021) propose architecture-agnostic and task-agnostic distillation methods, respectively.

Recently, several new white-box KD methods have been proposed (Zhou et al., 2021b; Agarwal et al., 2023; Liang et al., 2023a; Hwang et al., 2024; Agarwal et al., 2024; Gu et al., 2024; Anshumann et al., 2025), primarily designed for decoder-based LLMs. MetaDistill (Zhou et al., 2021b) utilizes a feedback mechanism within a meta-learning framework, GKD (Agarwal et al., 2023) trains the student on its self-generated output sequences, and PC-LoRA (Hwang et al., 2024) concurrently conducts progressive model compression and fine-tuning. One of the most notable recent works is MiniLLM (Gu et al., 2024), which suggests using the reverse of the Kullback-Leibler Divergence to prevent student models from overestimating the low-probability distribution of the teacher. CKA(Dasgupta & Cohn, 2025) matches hidden states of different dimensionality, allowing for smaller students and higher compression ratios, and Gong et al. (2025) Alignings feature dynamics for effective KD.

Furthermore, white-box KD has been successfully employed in industry to train small language models in pre-training stage (Liu et al., 2023b; Sreenivas et al., 2024; Muralidharan et al., 2024; Wang et al., 2024b; Team et al., 2025; Yang et al., 2025). Liu et al. (2023b) proposes a data-free, quantization-aware training approach in which a quantized version of the teacher is trained as a student using data-free distillation. Pruning-aware distillation has also been adopted by NVIDIA (Muralidharan et al., 2024; Sreenivas et al., 2024). Muralidharan et al. (2024) prunes the teacher's neurons and uses a small portion of the original data for post-training the student via KD. In a related approach, Sreenivas et al. (2024) additionally introduces a teacher correction step prior to distillation. Gemma3 (Team et al., 2025) and Qwen3 (Yang et al., 2025) employ logit sampling and weak-to-strong distillation, respectively.

Black-box KD has recently garnered increased attention due to the rise of closed-source LLMs. The first category of black-box KD is ICL, initially introduced in GPT-3 (Brown et al., 2020). In ICL, the teacher receives a task description, a few task examples, and a query, where the teacher predicts a response that the student aims to mimic. Huang et al. (2022e) combines in-context learning objectives with language modeling objectives to distill both the ability to interpret in-context examples and task knowledge into smaller models. AICD (Liu, 2024) explores the simultaneous distillation of in-context learning and reasoning, leveraging the autoregressive nature of LLMs to optimize the likelihood of all rationales in context. Wang et al. (2023h) introduces a novel framework to iteratively train dense retrievers capable of identifying high-quality in-context examples for LLMs.

The second category of black-box KD is IF, which seeks to enhance the zero-shot capabilities of LLMs through instruction prompts. Specifically, in IF, the teacher generates task-specific instructions on which the student network is fine-tuned (Jiang et al., 2023b; Zhou et al., 2023c; Li et al., 2024d; Wang et al., 2022b; Wu et al., 2023c). Self-Instruct (Wang et al., 2022b) generates instructions, input, and output samples, filters out invalid samples, and then fine-tunes the student model. UniversalNer (Zhou et al., 2023c) explores mission-focused instruction tuning to train student models capable of excelling in a broad application class. Lion (Jiang et al., 2023b) utilizes an adversarial framework to generate challenging instructions for students, while LaMini-LM (Wu et al., 2023c) compiles a large dataset containing 2.58 million instructions based on both existing and newly generated instructions.

The last and most popular black-box KD approach is COT, where the teacher model generates rationales alongside predictions, providing the student network with intermediate inference steps. In COT, the student is expected to generate predictions and reasonings similar to the teacher, similar to feature learning where the student mimics the teacher's intermediate steps for better predictions. This concept was first introduced by Li et al. (2022e) and has been expanded upon in subsequent works (Ho et al., 2022; Hsieh et al., 2023; Chen et al., 2023a; Magister et al., 2022; Li et al., 2023d; Wang et al., 2023j; Zhu et al., 2023b; Chae et al., 2023; Li et al., 2024g; Chen et al., 2024c; Zhuang et al., 2025). Magister et al. (2022) explores the trade-off between model and dataset size, demonstrating that the reasoning ability of LLMs can be transferred to SLMs. MCC-KD (Chen et al., 2023a) generates multiple reasonings for each sample and enforces consistency

and diversity among them. Hsieh et al. (2023) employs a multi-task learning framework to train the student with the teacher's reasonings. In contrast to other approaches that overlook reasonings with incorrect labels, Li et al. (2024g) utilizes both positive and negative data. Most recently, Chen et al. (2024c) demonstrated that generating backward questions and reasoning, and training the student on both forward and backward reasoning, significantly enhances the generalization ability of the student.

In summary, white-box KD methods possess greater efficacy in transferring knowledge by granting the student access to the internal information of the teacher. However, they typically come with high computational costs when distilling knowledge during student training. Moreover, most powerful LLMs are currently not open-source, rendering the use of white-box KD infeasible for these models. On the other hand, black-box methods enable KD from any model. Nevertheless, as black-box methods lack access to the internal workings of the teacher and solely rely on training the student with data generated by the teacher, they may not encompass all possibilities and could exhibit lower generalizability. One potential strategy for leveraging white-box distillation could involve initially distilling knowledge from a closed-source model to an open-source one using black-box KD, and subsequently applying a white-box method for further distillation.

### 6.2 Foundation Model Distillation

Foundation Models (FMs) are large-scale, pre-trained architectures designed to generalize across multiple downstream tasks. KD serves as a critical technique in optimizing these models by transferring their knowledge into smaller, task-specific, or efficient student models. KD addresses challenges such as computational costs and deployment on resource-constrained devices, making it indispensable for scaling FMs' applications.

Prominent FMs, including VLMs such as CLIP (Radford et al., 2021), conversational agents such as ChatGPT (Brown et al., 2020; Radford, 2018; Baktash & Dawodi, 2023), and generative models such as DALL-E (Ramesh et al., 2021), illustrate the diverse domains from which knowledge can be distilled using KD. KD is utilized to distill the complex, multi-modal representations from these models into simpler ones optimized for specific tasks. In addition, vision-specific models such as SAM (Kirillov et al., 2023) and DINO (Caron et al., 2021) serve as valuable sources of knowledge that can be distilled for various applications, including segmentation and unsupervised object discovery. Transformer-based models, including ViT (Dosovitskiy, 2020), DETR (Carion et al., 2020), Swin Transformer (Liu et al., 2021e), and DeIT (Touvron et al., 2021) also provide foundational knowledge that can be transferred through KD to enhance performance in vision-related tasks.

To provide a comprehensive overview of these FMs, this section explores KD for each model individually. Figure 9 presents a structured overview of the topics covered in this section, outlining the different types of FMs and their respective applications. To complement this, Figure 10 illustrates the architectures of these prominent models, highlighting key points where data is commonly extracted and distilled for downstream tasks.

#### 6.2.1 Vision-language Models and Knowledge Distillation

VLMs bridge visual and textual data through joint representation learning, offering capabilities to align these data types effectively. KD plays a central role in refining these models (Chen et al., 2024i; Wang et al., 2023h; Minderer et al., 2022), enabling their general-purpose knowledge to be adapted for more compact or task-specific models, enhancing scalability and reducing computational overhead.

CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021), one of the most significant VLMs, exemplifies this integration. Trained on paired image and text data, CLIP achieves remarkable generalization by aligning global visual and textual representations. Through KD, these capabilities are transferred into student models designed for downstream tasks, including several key areas:

**Open-vocabulary Tasks** In conventional vision tasks such as semantic segmentation and object detection, a dataset-focused methodology is traditionally employed, where the most successful techniques depend on a training dataset that has been manually annotated for a predefined and narrow range of categories. However, the rise of advanced VLMs, namely CLIP, is driving a shift towards an open-world paradigm. These models

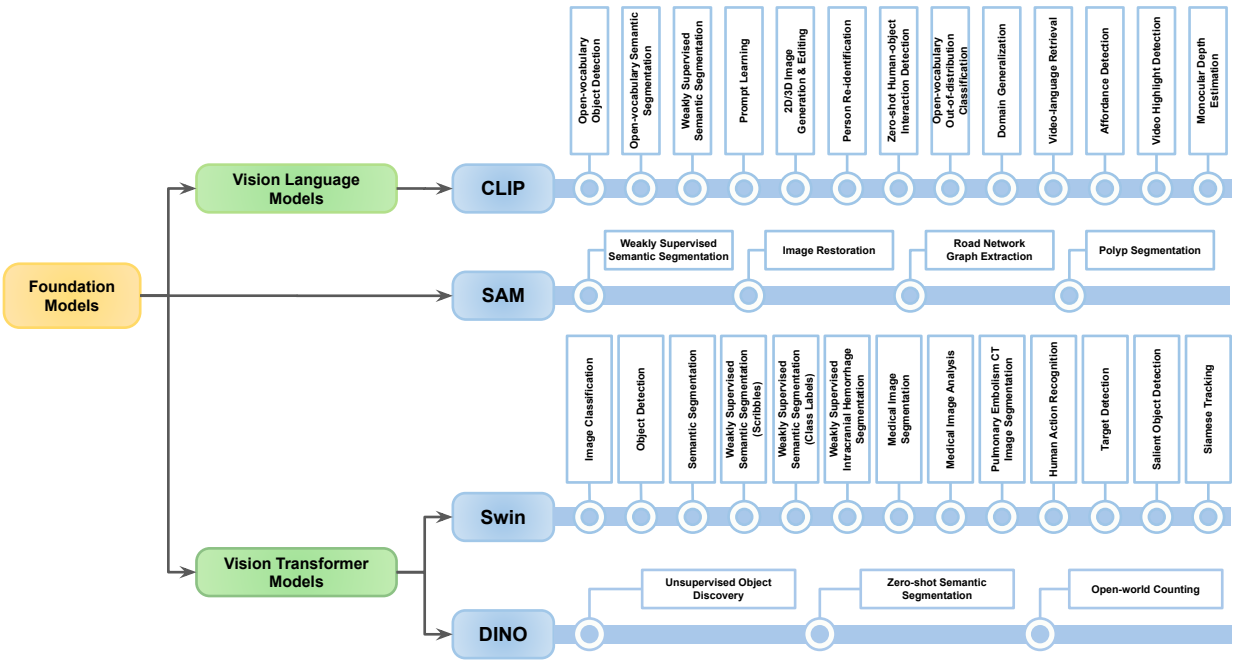


Figure 9: Categorization of foundation models based on their type, and the tasks for which their knowledge is distilled.

are trained through a simple yet scalable process that aligns image-text pairs, leveraging broad, loosely descriptive captions that can be collected in vast amounts with little human intervention. In object detection, KD techniques (Gu et al., 2021b; Ma et al., 2022; Bangalath et al., 2022; Zang et al., 2022; Xie & Zheng, 2022; Kim et al., 2023a; Wu et al., 2023d; Wang et al., 2023i; Du et al., 2022; Feng et al., 2022; Gao et al., 2022b; Huynh et al., 2022; Long et al., 2023; Zhao et al., 2022f; Lin et al., 2022a; Kuo et al., 2022; Zhou et al., 2022d; Yao et al., 2022) distill CLIP's global visual-textual knowledge into models capable of detecting objects outside fixed vocabularies, while in semantic segmentation (Lüddecke & Ecker, 2022; Ding et al., 2022a; Li et al., 2022a; Zabari & Hoshen, 2021; Zhou et al., 2022a; 2023d; Ma et al., 2022; Liang et al., 2023b; Xu et al., 2022a; Ghiasi et al., 2022; Shin et al., 2022a; Qin et al., 2023; Wysoczańska et al., 2024a; Chen et al., 2023b), and part segmentation (Choi et al., 2024) KD aids in transferring CLIP's textual alignment to dense pixel-level tasks, enabling fine-grained scene understanding. Furthermore, Wei et al. (2025) has explored open-vocabulary customization from CLIP through Data-Free Knowledge Distillation (DFKD), introducing a framework that enables CLIP model adaptation without requiring the original training data.

**Weakly Supervised Semantic Segmentation** The utilization of CLIP has recently been extensively explored for weakly supervised semantic segmentation (Xie et al., 2022; Lin et al., 2023b; Xu et al., 2023b; Deng et al., 2023; Murugesan et al., 2024; Zhang et al., 2024a; Zhu et al., 2024a; Jang et al., 2024; Zhou et al., 2025) where it enables the generation of segmentation masks with minimal labeled data, by leveraging its ability to align textual and visual representations within a shared embedding space. Through this approach, text prompts representing class labels guide the identification of relevant regions in an image by querying CLIP's pre-trained image-text alignment capabilities. The visual features extracted by CLIP's image encoder are matched to the semantic meaning encapsulated in text embeddings, producing coarse class activation maps that localize objects and regions associated with the given labels.

**Prompt Learning** Prompt learning is a method that enables the use of a large pre-trained model, such as CLIP, for specific tasks without having to retrain the entire model. This approach suggests adjusting

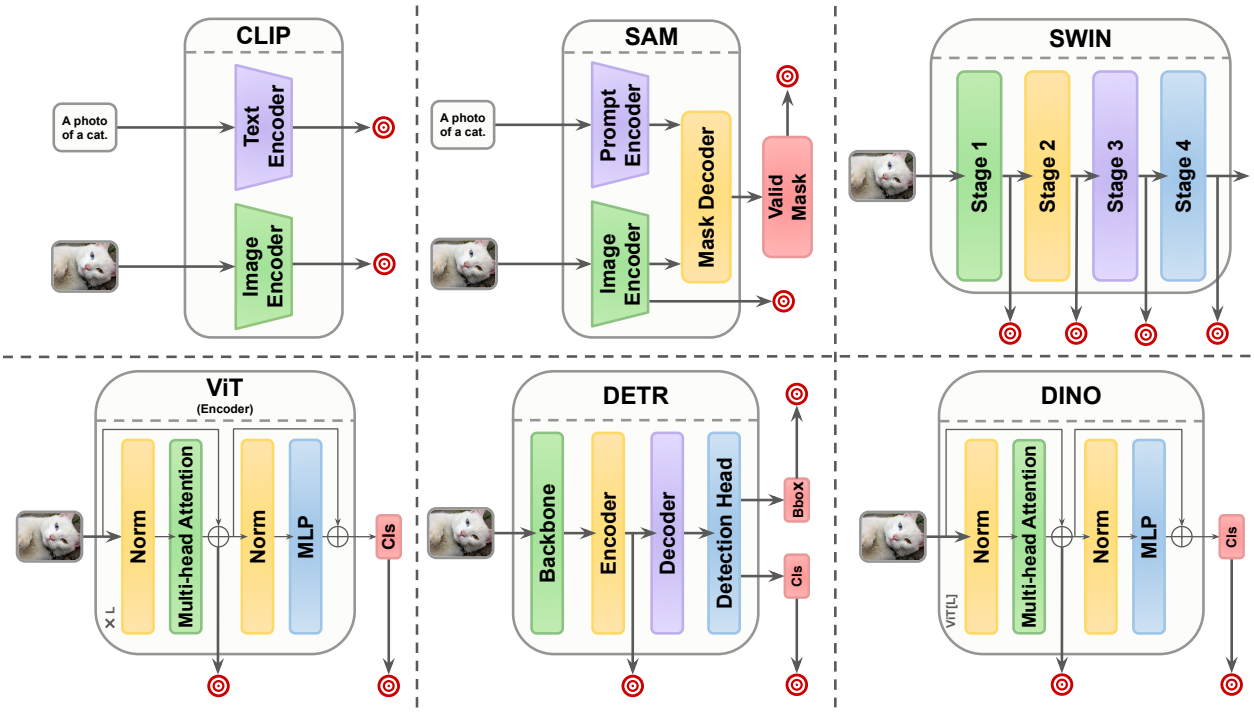


Figure 10: Architectural overview of prominent foundation models, indicating key points for data extraction and distillation to downstream tasks.

the model's representations for particular tasks by using learnable soft prompts, either text-based or visual, rather than relying on pre-designed hard prompts. Several methods (Zhou et al., 2022c;b; Li et al., 2024h; Ge et al., 2023; Rao et al., 2022) have been proposed to implement this approach.

**Generation and Editing** Applications such as generation and editing of 2D (Abdal et al., 2022; Gal et al., 2022; Patashnik et al., 2021) and 3D images (Hyung et al., 2023; Canfes et al., 2023; Hong et al., 2022a; Jain et al., 2022; Kobayashi et al., 2022; Michel et al., 2022; Sanghi et al., 2022; Wang et al., 2022a), where text-guided modifications are required, also benefit significantly from CLIP-guided KD.

**Other Applications** The range of tasks addressed by KD in VLMs, particularly CLIP, is broad, including person re-identification (Liu et al., 2024c), zero-shot human-object interaction (HOI) detection (Wan & Tuytelaars, 2024), open-vocabulary out-of-distribution classification (Li et al., 2023f), domain generalization (Huang et al., 2023c), video-language retrieval (Pei et al., 2023), affordance detection (Cuttano et al., 2024), video highlight detection (Han et al., 2024a), and monocular depth estimation (Son & Lee, 2024). These applications demonstrate how KD transforms large-scale VLMs into accessible, efficient tools for specialized needs, with some works (Yang et al., 2023a; Nair, 2024; Wu et al., 2023b; Dong et al., 2023b) tackling multiple tasks simultaneously.

### 6.2.2 Segment Anything (SAM)

The Segment Anything Model (SAM) (Kirillov et al., 2023) is a versatile and general-purpose model designed for efficient and accurate object segmentation across diverse image domains. Developed with a focus on flexibility, SAM employs a powerful vision-transformer architecture that can segment any object in an image based on prompts, such as points, boxes, or masks. Its pre-trained nature enables SAM to generalize well across tasks without task-specific fine-tuning, making it a valuable resource for many applications. Some

works (Zhang et al., 2024e; Hetang et al., 2024; Rahman et al., 2024) leverage SAM's pre-trained knowledge by distilling its segmentation capabilities into lighter, task-specific models, enabling the transfer of its robust segmentation performance to applications with constrained data annotations, computational resources or domain-specific requirements.

**SAM for Weakly Supervised Semantic Segmentation** SAM is utilized for weakly supervised semantic segmentation (Kweon & Yoon, 2024; Jiang & Yang, 2023; Chen & Sun, 2023; Chen et al., 2023c; Sun et al., 2023a) by leveraging its pre-trained capabilities to generate high-quality pseudo-labels from minimal or weak supervision signals, such as image-level tags, points, or bounding boxes. These pseudo-labels serve as a foundation for training more efficient models, bridging the gap between limited supervision and fully labeled data. The robustness of SAM's outputs allows semantic segmentation models to learn complex patterns and object boundaries without extensive manual annotation. This approach significantly reduces labeling costs while maintaining competitive performance.

#### 6.2.3 Multi-model Applications

In some instances, research has explored synergistic use of multiple FMs, combining their strengths through KD to enhance performance in segmentation, detection, and multi-modal applications. This collaborative use of KD fosters the development of unified models that inherit the complementary strengths of various teacher architectures. General studies have demonstrated the effectiveness of integrating diverse FMs across tasks (Yin & Jiang, 2024; Ranzinger et al., 2024; Sun et al., 2023c; Ye et al., 2023a; Shang et al., 2024; Gao et al., 2024; Englert et al., 2024; Rai et al., 2024), highlighting improvements in both efficiency and accuracy. One common approach involves combining models with complementary strengths. For instance, (SAM + CLIP) (Wang et al., 2024d; Aleem et al., 2024; Yang & Gong, 2024) leverages SAM's robust image segmentation alongside CLIP's semantic understanding, enhancing performance in various vision tasks. Similarly, (DINO + CLIP) (Wysoczańska et al., 2024b; Kerr et al., 2023) has been explored, taking advantage of DINO's self-supervised learning capabilities and CLIP's strong vision-language alignment. Extending this idea, (DETR + CLIP) has been proposed in Jiang et al. (2024b) for open-vocabulary object detection, where DETR serves as the detection model while CLIP provides text-based prompts. Additionally, research (Ren et al., 2024; Das et al., 2024) has investigated using the complementary strengths of (SAM + GDINO) for object detection and segmentation.

#### 6.2.4 Vision Transformers

Vision Transformers (ViTs) leverage the self-attention mechanism to capture long-range dependencies in images, enabling them to model complex and global visual features. This capability makes them highly effective for tasks such as image classification, segmentation, and detection. However, the high computational cost of training and deploying ViTs highlights the need for distilling the knowledge of pre-trained transformer models for more specialized tasks. The following section discusses prominent ViTs and their distillation methods.

**Vision Transformer** ViT (Dosovitskiy, 2020) is a deep learning model that extends the transformer architecture (Vaswani, 2017), originally developed for natural language processing, to image analysis. Several works (Lin et al., 2023a; Ren et al., 2023a; Shang et al., 2023a; Kang et al., 2023b; Bai et al., 2023; Zhao et al., 2023a; Yang et al., 2024) have explored the application of KD to ViT, investigating a range of strategies aimed at achieving diverse objectives, as summarized in Table 9.

**Distillation with No Labels (DINO)** DINO (Caron et al., 2021) is a self-supervised vision transformer that learns rich visual representations from unlabeled data. Due to its strong ability to capture meaningful features, many works leverage Dino's pretrained features for object localization for various downstream tasks, such as unsupervised object discovery (Kara et al., 2024; Siméoni et al., 2021; 2023; Wang et al., 2023n; 2022b; Liu et al., 2024f), open-world counting (Amini-Naieni et al., 2024), and zero-shot semantic segmentation (Karazija et al., 2023).

Table 9: Summary of Methods Applying Knowledge Distillation to the Vision Transformer Model.

| Method | Contribution |
|---|---|
| SMKD (Lin et al., 2023a) | Explores the tendency of ViTs to overfit and degrade in performance under few-shot learning, due to the lack of CNN-like inductive biases, and proposes a KD method to address this issue. |
| TinyMIM (Ren et al., 2023a) | Explores KD techniques for transferring the benefits of large MIM (Masked Image Modeling) pre-trained models to smaller ones. |
| Incrementer (Shang et al., 2023a) | Introduces a class-incremental semantic segmentation framework that employs ViT instead of traditional CNNs, emphasizing the need for global context modeling and incorporating convolutional layers to support new class predictions in incremental learning settings. |
| CST (Kang et al., 2023b) | Uses self-distillation by leveraging attention maps from a frozen ViT to generate pseudo-labels for few-shot, weakly supervised tasks. |
| DMAE (Bai et al., 2023) | Suggests distilling intermediate features, rather than logits, from a large ViT model to a smaller one. |
| CSKD (Zhao et al., 2023a) | Presents a method for enhancing ViT performance through KD from CNN models. |
| ViTKD (Yang et al., 2024) | Introduces a method for distilling knowledge between ViT feature maps instead of logits, using DeiT (Touvron et al., 2021) as the base model. |

**DEtection TRansformer (DETR)** DETR (Carion et al., 2020) is an end-to-end object detection model built on ViT, capable of directly predicting bounding boxes and class labels within a unified framework, thereby eliminating the need for anchors or complex post-processing. Its ability to effectively capture global context and model interactions between objects has positioned it as a leading backbone in object detection tasks. However, DETR's large model size and substantial computational demands pose significant challenges for deployment in real-world scenarios with limited computational budgets. To overcome these limitations, recent works (Chang et al., 2023; Wang et al., 2024m) have explored compressing DETR through KD, aiming to retain its strong performance while reducing computational overhead for real-time and resource-limited applications.

**Swin Transformer** Swin Transformer (Liu et al., 2021e) is a powerful architecture that effectively combines the strengths of CNNs and ViTs through its hierarchical design and efficient window-based attention mechanism. By employing local window attention within each window and shifting the windows between layers, Swin captures both local and global context, enabling it to excel in modeling fine-grained details as well as long-range dependencies. This rich feature representation has made Swin a preferred choice in various computer vision tasks. Recent works (Ahmadi & Kasaei, 2024; Zhang et al., 2021a; Cheng et al., 2021; Liu et al., 2021f; Lin et al., 2022b; Shi et al., 2022; Li et al., 2022f; Tang et al., 2022b; Rasoulian et al., 2023; Heidari et al., 2023; Chen & Mo, 2023; Wang et al., 2023f; Chen et al., 2024h) have leveraged Swin Transformer as an encoder, utilizing its features across different vision tasks, as summarized in Table 10.

Table 10: Summary of Methods Applying Knowledge Distillation to the Swin Transformer Model.

| Method | Task |
|---|---|
| MaskFormer (Cheng et al., 2021) | Semantic Segmentation |
| DFR (Zhang et al., 2021a) | Weakly Supervised Semantic Segmentation (Scribbles) |
| SwinNet (Liu et al., 2021f) | Salient Object Detection |
| SwinTrack (Lin et al., 2022b) | Siamese Tracking |
| SwinF (Li et al., 2022f) | Target Detection |
| Swin-UNETR (Tang et al., 2022b) | Medical Image Analysis |
| Ssformer (Shi et al., 2022) | Semantic Segmentation |
| HGI-SAM (Rasoulian et al., 2023) | Weakly Supervised Intracranial Hemorrhage Segmentation |
| HiFormer (Heidari et al., 2023) | Medical Image Segmentation |
| Swin-Fusion (Chen & Mo, 2023) | Human Action Recognition |
| P.Swin (Wang et al., 2023f) | Image Classification and Object Detection |
| SCUNet++ (Chen et al., 2024h) | Pulmonary Embolism CT Image Segmentation |
| SWTformer (Ahmadi & Kasaei, 2024) | Weakly Supervised Semantic Segmentation (Class Labels) |

### 6.3 Distillation in Self-supervised learning

Self-Supervised Learning (SSL) is an important field due to its independence from labeled data, and self-distillation is the basis of the most prominent works in SSL, especially in the realm of computer vision. The most prominent work in SSL was based on self-distillation (Chen et al., 2020c). SimCLR (Chen et al., 2020c) proposed a shared network that receives two different augmentations of an image and minimizes their extracted embeddings. Following SimCLR, different extensions have been proposed (He et al., 2020; Chen et al., 2020d; Grill et al., 2020; Misra & Maaten, 2020; Caron et al., 2020; Xu et al., 2021). MOCO (He et al., 2020; Chen et al., 2020d) and BYOL (Grill et al., 2020) are among successful methods where they propose to use the same network but with a different training process. The student is trained using backpropagation, and the teacher is an EMA of the student. SwAP (Caron et al., 2020) uses some prototypes and assigns a code to the augmentations of the same image. The student is then trained to predict the same class ID as the teacher. DINO (Caron et al., 2021) and DINOv2 (Oquab et al., 2023) are the most recent important applications of KD in SSL, where global and local patches of an image pass through similar teacher and student networks. The teacher is updated using EMA, and the representations of the teacher and the student are forced to be similar. This local-to-global approach makes it ideal for segmentation tasks.

Recently, more complex forms of SSL based on KD have been proposed (Misra & Maaten, 2020; Xu et al., 2021; Wang et al., 2023l; Song et al., 2023a; Gao et al., 2022c; Gu et al., 2021a). BINGO (Xu et al., 2021) uses a pre-trained SSL method to group the images of a dataset into a bag of samples. Then, in each epoch, samples are taken from the bag and the inter and intra-sample losses are minimized. Wang et al. (2023l) trains two SSL teachers for images and videos using masked modeling and follows a scenario similar to SimCLR, minimizing the spatial features with the image teacher and the spatio-temporal features with the video teacher. Song et al. (2023a) uses a parallel SSL approach and defines consistency losses between them, and Gao et al. (2022c) integrates a known KD approach into an SSL framework using an SSL pre-trained teacher.

Apart from the mentioned papers, some studies attempt to distill the knowledge of a pre-trained SSL network into a smaller network (Abbasi Koohpayegani et al., 2020; Tejankar et al., 2021; Navaneet et al., 2022; Fang et al., 2021; Dadashzadeh et al., 2022). The gap between two similar networks, one trained fully-supervised and one trained self-supervised, is shown to decrease with a bigger model size (Abbasi Koohpayegani et al., 2020). Navaneet et al. (2022) shows the impact of an MLP layer after representations of the student, while Abbasi Koohpayegani et al. (2020) proposes a similarity-based distillation method for distilling the knowledge of a pre-trained network, which can outperform the corresponding fully-supervised trained network. In a similar approach, SEED (Fang et al., 2021) uses a queue of samples and minimizes the distance of the student's representation with anchors in the queue and the corresponding distances between the anchors and the teacher's representation. Dadashzadeh et al. (2022) proposes a novel approach to complement self-supervised pretraining via an auxiliary pretraining phase for small unlabeled datasets.

### 6.4 Diffusion Distillation

Diffusion models (Ho et al., 2020; Song et al., 2020) generate high-quality images, text, and 3D data through an iterative denoising process that can require hundreds or thousands of steps. This is computationally expensive, limiting deployment on low-resource hardware. KD mitigates this by transferring soft outputs and internal representations from a large teacher model to a smaller student model. The distilled model then achieves the same quality with far fewer sampling steps, requiring significantly less inference time and computation (Luo, 2023).

Several strategies have been developed to accelerate diffusion models. Feature-level methods have the student mimic the teacher's internal representations (Huang et al., 2023b; Hsiao et al., 2024; Feng et al., 2024b; Zhang et al., 2024b). Sampling process distillation reduces the iterative denoising by either merging multiple steps into one or matching a one-step student's output to that of the full teacher (Salimans & Ho, 2022; Yin et al., 2024c; Zhou et al., 2024e). Adversarial and data-free approaches align outputs using GAN losses or bootstrapping without requiring large synthetic datasets (Sauer et al., 2024; Mekonnen et al., 2024; Xie et al., 2024b). Consistency models reformulate the denoising into a direct, often single-step transformation,

cutting down inference time (Song et al., 2023c; Song & Dhariwal, 2023; Geng et al., 2024; Luo et al., 2023a; Xie et al., 2024a; Lu & Song, 2024). In the following sections, each category is explained in detail.

**Feature-level:** Feature-level distillation accelerates diffusion model efficiency by bringing the student representations closer to the teacher and reducing inference steps. Recent methods reinforce feature alignment through denoising (Huang et al., 2023b), external guidance (Hsiao et al., 2024), and cross-sample consistency (Feng et al., 2024b), which facilitate convergence and improved sampling efficiency. Furthermore, combining feature distillation with model compression reduces model size and inference time (Zhang et al., 2024b). These techniques accelerate generation without compromising output quality, thereby making diffusion models more realistic.

**Sampling Process:** Distillation of the sampling procedure is performed to ease the naturally iterative denoising process characteristic of diffusion models. Progressive distillation (Salimans & Ho, 2022) achieves this by consolidating two steps into a single step; through repeated application of this simplification, the duration of the sampling procedure can be drastically shortened. Distribution matching distillation (Yin et al., 2024c) attempts to condense the entire iterative procedure into a single step. This alignment is accomplished by matching the output distribution of a one-step student model directly with that of the combined multi-step teacher model.

**Adversarial and Data-free:** Adversarial methods (Sauer et al., 2024; Mekonnen et al., 2024) leverage GAN-based losses to enforce the student's output distribution to closely match that of the teacher. Alternatively, data-free methods like EM distillation (Xie et al., 2024b), employ bootstrapping between successive denoising steps, eliminating the need for large synthetic datasets during training.

**Consistency Models:** Consistency models (Song et al., 2023c; Song & Dhariwal, 2023; Geng et al., 2024) reformulate the denoising process as a direct, often single transformation from noisy data to its clean version. Their latent-space variants (Luo et al., 2023a; Xie et al., 2024a), along with improvements focusing on stability and scalability (Lu & Song, 2024), achieve significant speedups in inference times. However, training these models may require careful tuning to prevent quality collapse.

Furthermore, recent developments in the area of distillation techniques expanded their accessibility to multi-modal and multi-model settings, thus improving transferability across tasks and diverse architectures. ProlificDreamer (Wang et al., 2023q) and DREAMFUSION (Poole et al., 2022) utilize the methods in the realm of text-to-3D translation, and general-purpose systems like Diff-Instruct (Luo et al., 2023b) can facilitate generalization for the task.

In summary, various approaches offer distinct benefits: Sampling process distillation drastically accelerates generation but tends to require cumbersome training in order to maintain sample quality. Adversarial and data-free methods provide more flexibility, though sometimes at the cost of training stability. Consistency models enable low-latency generation, provided they are well-calibrated. Furthermore, combining cross-modal and universal approaches significantly expands the applicability of these distillation techniques across a wide range of generation tasks.

### 6.5 Visual Recognition Distillation

Knowledge distillation has been widely applied to computer vision tasks, enabling the training of compact student models replicating the behavior of larger, resource-intensive teacher models. Key applications of KD in visual recognition are summarized in Table 11, based on two main factors: the type of knowledge transferred (feature-based, logit-based, similarity-based, and combinations) and the distillation scheme employed (offline, online, and self-distillation). Several important visual recognition tasks are covered in this table, including depth estimation, face recognition, image segmentation, medical imaging, object detection, object tracking, pose estimation, super resolution, action recognition, and image retrieval.

Table 11: Summary of knowledge distillation methods in visual recognition tasks based on their sources and schemes of distillation.

| Task | Knowledge | Scheme | Reference |
|---|---|---|---|
| Depth Estimation | feature | offline | (López-Cifuentes et al., 2023) |
| | | self-distillation | (Zhou & Dong, 2022) |
| | logit | offline | (Hu et al., 2023b), (Xu et al., 2023c), (Shi et al., 2023), (Ka et al., 2024), (Wang & Liu, 2024) |
| | | self-distillation | (Peng et al., 2021), (Zhao et al., 2023b), (Marsal et al., 2024), (Hu et al., 2024a), (Boutros et al., 2024), (Dong et al., 2024c) |
| | similarity | offline | (Chen et al., 2024e) |
| | feature + logit | offline | (Wu et al., 2023f) |
| | | online | (Shao et al., 2024b) |
| Face Recognition | feature | offline | (Shin et al., 2022b), (Boutros et al., 2024), (Otroshi-Shahreza et al., 2023), (Neto et al., 2024), (Caldeira et al., 2024), (Li et al., 2023b), (Huang et al., 2024f), (Babnik et al., 2024) |
| | logit | offline | (Jung et al., 2021), (Liu et al., 2021a), (Xu et al., 2023a) |
| | similarity | offline | (Huang et al., 2022d) |
| | | self-distillation | (Di et al., 2024) |

Table 11 – Continued

| Task | Knowledge | Scheme | Reference |
|---|---|---|---|
| Image Segmentation | feature | offline | (Ji et al., 2022), (Yang et al., 2022f), (Kang et al., 2023a), (Shang et al., 2023a), (Yuan et al., 2024a), (Liu et al., 2024g), (Shu et al., 2021) |
| | | self-distillation | (Wu et al., 2024b) |
| | logit | offline | (Amirkhani et al., 2021), (Xu et al., 2023c), (Li et al., 2024c) |
| | | online | (Shen et al., 2023), (Berrada et al., 2024) |
| | similarity | offline | (Yang et al., 2022c), (Mansourian et al., 2025) |
| | | online | (Gao et al., 2022a) |
| | feature + logit | offline | (Tian et al., 2022c), (Phan et al., 2022), (Xiao et al., 2023a), (Zheng et al., 2025) |
| | | online | (Zhu et al., 2023a) |
| | feature + similarity | offline | (Feng et al., 2021), (Liu et al., 2024d) |
| | | online | (Wang et al., 2024p) |
| Medical Imaging | feature | offline | (Qin et al., 2021), (Li et al., 2022d), (Ghosh et al., 2023), (Wang et al., 2023g), (Cao et al., 2024), (Nasser et al., 2024), (Chen et al., 2024d), (Zhou et al., 2024c) |
| | | online | (Dong et al., 2025) |
| | | self-distillation | (Ye et al., 2022), (Yi et al., 2024a) |
| | logit | offline | (Feng et al., 2024c), (Wang et al., 2024l) |
| | | online | (Xie et al., 2023c) |
| | | self-distillation | (Zhong et al., 2023), (Huang et al., 2024b) |
| | similarity | online | (Xing et al., 2021) |
| | | self-distillation | (You et al., 2022), (Sharma et al., 2024) |
| | feature + logit | offline | (El-Assiouti et al., 2024) |
| | feature + similarity | offline | (Liu et al., 2022c), (Qi et al., 2024) |
| Object Detection | feature | offline | (Dai et al., 2021b), (Guo et al., 2021a), (Guo et al., 2021b), (Kang et al., 2021), (Yang et al., 2022a), (Jang et al., 2022), (Yang et al., 2022f), (Lao et al., 2023), (Lan & Tian, 2024), (Wang et al., 2023i), (Liu et al., 2024i) |
| | | self-distillation | (Wu et al., 2023a), (Jia et al., 2024), (Wu et al., 2024b), (Zheng et al., 2023) |
| | logit | offline | (Yang et al., 2023e), (Jin et al., 2023a), (Li et al., 2023c) |
| | | self-distillation | (Zhou et al., 2023a) |
| | | online | (Nguyen et al., 2022) |
| | similarity | offline | (Ni et al., 2023) |
| | feature + logit | offline | (Tang et al., 2022a), (Du et al., 2021) |
| | | online | (Zhang & Ma, 2021), (Yao et al., 2021), (Yang et al., 2022e), (Li et al., 2022b), (Zhang & Ma, 2023), (Zhu et al., 2023c) |
| | feature + similarity | offline | (Zhang & Ma, 2021), (Yao et al., 2021), (Yang et al., 2022e), (Li et al., 2022b), (Zhang & Ma, 2023), (Zhu et al., 2023c) |
| | | self-distillation | (Zhang et al., 2022c) |
| | logit + similarity | offline | (Wang et al., 2024f) |

Table 11 – Continued

| Task | Knowledge | Scheme | Reference |
|---|---|---|---|
| Object Tracking | feature | offline | (Hou et al., 2024), (Zhang et al., 2024h) |
| | logit | offline | (Valverde et al., 2021) |
| | feature + similarity | offline | (Wang et al., 2024k) |
| Pose Estimation | feature | offline | (Nie et al., 2019) |
| | | self-distillation | (Chen et al., 2024f) |
| | logit | offline | (Guo et al., 2023b), (Ye et al., 2023b), (Wang et al., 2021c), (Aghli & Ribeiro, 2021), (Xu et al., 2022c), (Zhang et al., 2019a) |
| | | online | (Li et al., 2021b) |
| | | self-distillation | (Li & Lee, 2021) |
| | feature + logit | offline | (Fang et al., 2023b), (GUAN et al., 2023) |
| | | self-distillation | (Yang et al., 2023h) |
| Super Resolution | feature | offline | (Jiang et al., 2024d), (Liu et al., 2024b) |
| | | self-distillation | (Wang et al., 2024e), (Wang & Zhang, 2024) |
| | logit | offline | (Noroozi et al., 2024) |
| | | self-distillation | (Park, 2024) |
| | relation | offline | (Angarano et al., 2022), (Zhang et al., 2024j) |
| | feature + similarity | offline | (Xie et al., 2023a), (Fang et al., 2023a) |
| | | self-distillation | (Wang et al., 2021f) |
| | feature + logit | offline | (Zhang et al., 2021g), (Yu et al., 2023), (Wang et al., 2024n) |
| Action Recognition | feature | offline | (Zhao et al., 2022d), (Wang et al., 2024c), (Lee et al., 2023b), (Huang et al., 2024d), (Hong et al., 2021) |
| | | online | (Guo et al., 2023a) |
| | logit | offline | (Radevski et al., 2023) |
| | feature + similarity | offline | (Dai et al., 2021a), (Bian et al., 2021) |
| Image Retrieval | feature | offline | (Jin et al., 2020), (Porrello et al., 2020), (Budnik & Avrithis, 2021) |
| | similarity | offline | (Wu et al., 2022), (Xu et al., 2024a), (Tian et al., 2021), (Xie et al., 2024d) |
| | feature + logit | offline | (Xie et al., 2023b) |
| | feature + similarity | offline | (Suma & Tolias, 2023), (Xie et al., 2024c) |

## 7 Performance Comparison

KD enables the compression of a deep network into a smaller model while preserving strong performance. Due to the vast number of methods proposed in this field, a comprehensive comparison across existing methods in different settings is necessary. However, the complicated nature of KD, encompassing various schemes, algorithms, sources, and teacher-student architectures, makes it challenging to conduct a comprehensive and fair comparison of all existing methods. For example, the exact pre-trained teacher model along with factors such as the number of epochs and batch size, can influence the results of the distillation. These factors are highly dependent on the hardware resources employed in each method, which is the determining factor in the final results.

In this work, existing methods are compared for two well-known tasks: classification and semantic segmentation segmentation, as almost all the proposed distillation methods evaluate their results on these two tasks. For classification, the ImageNet and CIFAR-100 datasets, and for semantic segmentation, the Cityscapes and PascalVOC datasets are used, as these two datasets are widely employed in existing methods.

Table 12: Performance comparison of different knowledge distillation methods on the CIFAR-100 classification dataset (* denotes that results are not reported from the original paper).

| Method | Knowledge | Teacher (baseline) | Student (baseline) | Accuracy | Code |
|---|---|---|---|---|---|
| FitNet (Romero et al., 2014) | Feature | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 72.24 (+0.26)* | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 73.54 (+1.72) | |
| KD (Hinton, 2015) | Logit | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 73.54 (+1.56)* | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 74.45 (+2.63) | |
| AT (Zagoruyko & Komodakis, 2016) | Feature | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 72.77 (+1.79)* | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 72.73 (+0.91) | |
| CCKD (Peng et al., 2019) | Similarity | ResNet-110 (–) | ResNet-20 (68.40) | 72.40 (+4.00) | – |
| SPKD (Tung & Mori, 2019) | Similarity | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 72.43 (+1.45)* | – |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 74.56 (+2.74) | |
| CRD (Tian et al., 2019) | Feature | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 74.14 (+2.16) | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 75.65 (+3.83) | |
| OFD (Heo et al., 2019a) | Feature | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 74.33 (+2.35) | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 76.82 (+5) | |
| Chen et al. (Chen et al., 2020f) | Similarity + Logit | ResNet-101 (71.77) | ResNet-18 (65.64) | 69.14 (+3.49) | Link |
| TAKD (Mirzadeh et al., 2020) | Logit | ResNet-110 (–) | ResNet-8 (–) | 61.82 | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 74.82 (+3) | |
| SphericalKD (Guo et al., 2020a) | Logit | ResNet-32×4 (79.42) | ResNet-8×4 (72.50) | 76.40 (+3.90) | Link |
| RKD (Chen et al., 2021d) | Feature | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 75.09 (+3.11) | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 77.78 (+5.96) | |
| SemCKD (Chen et al., 2021a) | Feature + Logit | WRN-40-2 (75.61) | MobileNetV2 (65.43) | 69.67 (+4.24) | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (72.60) | 77.62 (+5.02) | |
| ICKD (Liu et al., 2021c) | Similarity + Logit | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 74.63 (+2.65) | Link |
| DKD (Zhao et al., 2022a) | Logit | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 74.81 (+2.83) | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 77.07 (+5.25) | |
| DistKD (Huang et al., 2022a) | Similarity + Logit | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 74.73 (+2.75) | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 77.35 (+5.53) | |
| Norm (Liu et al., 2023a) | Feature | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 74.82 (+2.84) | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 78.07 (+6.25) | |
| SFKD (Yuan et al., 2024b) | Logit | ResNet-32×4 (79.55) | ResNet-8×4 (72.50) | 76.84 (+4.34) | – |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 76.96 (+5.14) | |
| NormKD (Chi et al., 2023) | Logit | ResNet-32×4 (79.42) | ResNet-8×4 (72.50) | 76.57 (+4.07) | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 76.01 (+4.19) | |
| CRLD (Zhang et al., 2024g) | Similarity + Logit | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 75.58 (+3.60) | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 78.27 (+6.45) | |
| NTCE-KD (Li et al., 2024a) | Logit | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 76.44 (+4.46) | – |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 79.43 (+7.61) | |
| TTM (Zheng & Yang, 2024) | Logit | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 74.58 (+2.60) | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 76.57 (+4.75) | |
| Miles et al. (Miles & Mikolajczyk, 2024) | Logit | WRN-40-2 (76.46) | WRN-16-2 (73.64) | 76.14 (+2.50) | Link |
| | | ResNet32×4 (79.63) | ShuffleNetV2 (73.12) | 79.06 (+5.94) | |
| SDD (Wei et al., 2024b) | Logit | ResNet-32×4 (79.42) | MobileNetV2 (64.6) | 68.84 (+4.24) | Link |
| | | ResNet50 (79.34) | MobileNetV2 (64.60) | 71.46 (+6.86) | |
| Sun et al. (2024a) | Logit | WRN-40-2 (75.61) | ResNet-8×4 (72.50) | 77.11 (+3.14) | Link |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 75.56 (+3.74) | |
| InDistill (Sarridis et al., 2025) | Feature | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 75.09 (+3.11) | Link |
| IKD (Wang et al., 2025a) | Logit | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 74.84 (+2.86) | – |
| | | ResNet50 (79.34) | MobileNetV2 (64.60) | 70.49 (+5.89) | |
| ADG-KD (Li et al., 2025b) | Logit | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 75.84 (+3.86) | – |
| | | ResNet32×4 (79.42) | ShuffleNetV2 (71.82) | 78.96 (+7.14) | |
| TeKAP (Hossain et al., 2025) | Feature + Logit | WRN-40-2 (75.61) | WRN-40-1 (71.98) | 74.41 (+2.43) | Link |
| | | ResNet32×4 (74.64) | ShuffleNetV2 (70.36) | 77.38 (+7.02) | |

ImageNet is a large-scale dataset designed for visual object recognition tasks, consisting of over 1.2 million training images, 50,000 validation images, and 100,000 testing images across 1,000 object classes. CIFAR-100 is composed of $32 \times 32$ images taken from 100 classes, with 50,000/10,000 images for training/testing, and each class has the same number of training and testing images. Cityscapes is designed for understanding urban scenes and includes 2,975/500/1,525 images for training/validation/testing in 19 classes. PascalVOC dataset comprises 1464/1449/1456 images for train/val/test, with 21 classes.

Table 13: Performance comparison of different knowledge distillation methods on the ImageNet classification dataset (* denotes that results are not reported from the original paper).

| Method | Knowledge | Teacher (baseline) | Student (baseline) | Accuracy | Code |
|---|---|---|---|---|---|
| FitNet (Romero et al., 2014) | Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 70.31 (+0.56)* | Link |
| KD (Hinton, 2015) | Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 70.62 (+0.87)* | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (68.87) | 68.58 (-0.29)* | |
| AT (Zagoruyko & Komodakis, 2016) | Feature | ResNet-34 (73.31) | ResNet-18 (69.75) | 70.30 (+0.55)* | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (68.87) | 69.56 (+0.69)* | |
| CRD (Tian et al., 2019) | Feature | ResNet-34 (73.31) | ResNet-18 (69.75) | 71.17 (+1.42) | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (68.87) | 71.37 (+2.50) | |
| OFD (Heo et al., 2019a) | Feature | ResNet-34 (73.31) | ResNet-18 (69.75) | 70.81 (+1.06) | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (68.87) | 71.25 (+2.38) | |
| SphericalKD (Guo et al., 2020a) | Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 72.30 (+2.55) | Link |
| RKD (Chen et al., 2021d) | Feature + Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 71.61 (+1.86) | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (68.87) | 72.56 (+3.69) | |
| SemCKD (Chen et al., 2021a) | Feature + Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 70.87 (+1.12) | Link |
| ICKD (Liu et al., 2021c) | Similarity + Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 72.19 (+2.44) | Link |
| DKD (Zhao et al., 2022a) | Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 72.05 (+2.30) | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (68.87) | 72.05 (+3.18) | |
| DistKD (Huang et al., 2022a) | Similarity + Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 72.07 (+2.32) | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (70.13) | 73.24 (+3.11) | |
| CAT-KD (Guo et al., 2023d) | Feature | ResNet-34 (73.31) | ResNet-18 (69.75) | 71.26 (+1.51) | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (70.13) | 72.24 (+3.37) | |
| NKD (Yang et al., 2023g) | Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 71.96 (+2.21) | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (70.13) | 72.58 (+3.71) | |
| MLLD (Jin et al., 2023a) | Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 71.90 (+2.15) | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (70.13) | 73.01 (+4.14) | |
| Norm (Liu et al., 2023a) | Feature | ResNet-34 (73.31) | ResNet-18 (69.75) | 72.14 (+2.39) | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (68.87) | 74.26 (+5.39) | |
| SFKD (Yuan et al., 2024b) | Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 71.86 (+2.11) | – |
| | | ResNet-50 (76.16) | MobileNetV1 (68.87) | 72.24 (+3.37) | |
| NormKD (Chi et al., 2023) | Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 72.03 (+2.28) | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (68.87) | 72.79 (+3.92) | |
| CRLD (Zhang et al., 2024g) | Similarity + Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 72.37 (+2.62) | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (70.13) | 73.53 (+4.66) | |
| NTCE-KD (Li et al., 2024a) | Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 73.12 (+3.37) | – |
| | | ResNet-50 (76.16) | MobileNetV1 (70.13) | 74.90 (+6.03) | |
| TTM (Zheng & Yang, 2024) | Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 72.19 (+2.44) | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (70.13) | 773.09 (+4.22) | |
| SDD (Wei et al., 2024b) | Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 72.33 (+2.58) | Link |
| Sun et al. (2024a) | Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 72.08 (+2.33) | Link |
| | | ResNet-50 (76.16) | MobileNetV1 (68.87) | 73.22 (+4.35) | |
| InDistill (Sarridis et al., 2025) | Feature | ResNet-34 (73.31) | ResNet-18 (69.75) | 71.34 (+1.59) | Link |
| IKD (Wang et al., 2025a) | Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 71.82 (+2.07) | – |
| ADG-KD (Li et al., 2025b) | Logit | ResNet-34 (73.31) | ResNet-18 (69.75) | 72.22 (+2.47) | – |
| | | ResNet-50 (76.16) | MobileNetV1 (68.87) | 73.43 (+4.56) | |

Furthermore, due to the importance of distillation in LLMs nowadays, compare existing methods for distillation in LLMs are compared. This includes comparing black-box and white-box methods. However, datasets and models used in LLMs vary significantly across different papers, specifically in white-box methods. Results for the datasets and models that are more commonly used in existing methods are reported, making the comparison easier and fair. The datasets used are: GLUE, Multilingual NER, DollyEval, IMDB, GSM8K, CrossFit, CommonsenseQA, SVAMP, Universal NER Benchmark, OpenBookQA, BBH, StrategyQA, and MATH.

Tables 12 and 13 summarize various knowledge distillation methods applied to classification tasks on the CIFAR-100 and ImageNet datasets. For each method, the source of knowledge, teacher and student architectures, and post-distillation accuracy are reported. In most cases, two scenarios are considered: one where the teacher and student share similar architectures, and another where their architectures differ.

Table 14: Performance comparison of different knowledge distillation methods on the PascalVoc segmentation dataset (* denotes that results are not reported from the original paper).

| Method | Knowledge | Teacher (baseline) | Student (baseline) | Accuracy | Code |
|---|---|---|---|---|---|
| FitNet (Romero et al., 2014) | Logit | DeepLab-R50 (76.21) | DeepLab-MV2 (70.57) | 71.30 (+0.73)* | Link |
| KD (Hinton, 2015) | Logit | DeepLab-R101 (77.85) | DeepLab-R18 (67.50) | 69.13 (+1.63)* | Link |
| | | DeepLab-R101 (77.85) | DeepLab-MV2 (63.92) | 66.39 (+2.47) | |
| AT (Zagoruyko & Komodakis, 2016) | Feature | DeepLab-R101 (77.85) | DeepLab-R18 (67.50) | 68.95 (+1.09)* | Link |
| | | DeepLab-R101 (77.85) | DeepLab-MV2 (63.92) | 66.27 (+2.35) | |
| SKD (Liu et al., 2019d) | Feature + Logit | PSPNet-R101 (78.52) | PSPNet-R18 (71.35) | 72.01 (+0.66) | Link |
| | | PSPNet-R101 (77.67) | DeepLab-R18 (71.98) | 73.03 (+1.05) | |
| He et al. (2019) | Feature + Similarity | DeepLab-R50 (76.21) | DeepLab-MV2 (70.57) | 72.50 (+1.93) | – |
| IFVD (Wang et al., 2020d) | Similarity + Logit | PSPNet-R101 (78.52) | PSPNet-R18 (71.35) | 73.49 (+2.14) | Link |
| | | PSPNet-R101 (77.67) | DeepLab-R18 (71.98) | 72.87 (+0.89) | |
| CWD (Shu et al., 2021) | Feature | PSPNet-R101 (78.52) | PSPNet-R18 (71.35) | 73.49 (+2.14) | Link |
| | | PSPNet-R101 (77.67) | DeepLab-R18 (71.98) | 73.78 (+1.79) | |
| DSD (Feng et al., 2021) | Similarity | DeepLab-R101 (79.10) | DeepLab-R18 (71.9) | 73.70 (+1.80) | – |
| CIRKD (Yang et al., 2022c) | Similarity + Logit | DeepLab-R101 (77.67) | DeepLab-R18 (73.21) | 74.50 (+1.29) | Link |
| | | DeepLab-R101 (77.67) | PSPNet-R18 (73.33) | 74.78 (+1.45) | |
| DIST (Huang et al., 2022a) | Similarity + Logit | DeepLab-R101 (77.85) | DeepLab-R18 (67.50) | 69.84 (+2.34) | Link |
| | | DeepLab-R101 (77.85) | PSPNet-R18(67.40) | 68.93 (+1.53) | |
| IDD (Zhang et al., 2022d) | Similarity | PSPNet-R101 (77.86) | PSPNet-R18 (70.80) | 76.70 (+5.9) | – |
| MGD (Yang et al., 2022f) | Feature | PSPNet-R101 (78.52) | PSPNet-R18 (71.35) | 74.98 (+3.63)* | Link |
| FAKD (Yuan et al., 2024a) | Feature | PSPNet-R101 (78.52) | PSPNet-R18 (70.52) | 73.97 (+3.45) | – |
| | | DeepLab-R101 (78.62) | DeepLab-MV2 (62.56) | 67.62 (+5.06) | |
| BPKD (Liu et al., 2024d) | Logit | PSPNet-R101 (78.52) | PSPNet-R18 (71.35) | 75.74 (+4.39) | Link |
| LAD (Liu et al., 2024g) | Feature + Logit | PSPNet-R101 (78.52) | PSPNet-R18 (71.35) | 75.74 (+4.39) | – |
| | | PSPNet-R101 (78.52) | DeepLab-R18 (71.98) | 76.33 (+4.35) | |
| AttnFD (Mansourian et al., 2024) | Feature | DeepLab-R101 (77.85) | DeepLab-R18 (67.50) | 73.09 (+5.59) | Link |
| | | DeepLab-R101 (77.85) | PSPNet-R18(67.40) | 70.95 (+3.55) | |
| AICSD (Mansourian et al., 2025) | Similarity + Logit | DeepLab-R101 (77.85) | DeepLab-R18 (67.50) | 70.58 (+3.08) | Link |
| | | DeepLab-R101 (77.85) | PSPNet-R18(66.60) | 68.29 (+1.69) | |

In a similar vein, Tables 14 and 15 report the results of various methods for the semantic segmentation task on the PASCAL VOC and Cityscapes datasets. For a fair comparison, the reported results are taken directly from the original papers. However, in cases where results on the specified datasets are not provided in the original work, we report results obtained from our own implementations (denoted by *).

Table 15: Performance comparison of different knowledge distillation methods on the Cityscapes segmentation dataset (* denotes that results are not reported from the original paper).

| Method | Knowledge | Teacher (baseline) | Student (baseline) | Accuracy | Code |
|---|---|---|---|---|---|
| KD (Hinton, 2015) | Logit | DeepLab-R101 (77.66) | DeepLab-R18 (64.09) | 65.21 (+1.12)* | Link |
| | | DeepLab-R101 (77.66) | DeepLab-MV2 (63.05) | 64.03 (+0.98) | |
| AT (Zagoruyko & Komodakis, 2016) | Feature | DeepLab-R101 (77.66) | DeepLab-R18 (64.09) | 65.29 (+1.20)* | Link |
| | | DeepLab-R101 (77.66) | DeepLab-MV2 (63.05) | 63.72 (+0.67) | |
| Xie et al. (2018) | Similarity + Logit | DeepLab-R101 (70.90) | DeepLab-MV2 (67.30) | 71.90 (+4.60) | – |
| SKD (Liu et al., 2019d) | Feature + Logit | PSPNet-R101 (79.76) | PSPNet-R18 (72.65) | 74.23 (+1.58) | Link |
| | | PSPNet-R101 (79.76) | DeepLab-R18 (74.96) | 75.32 (+0.36) | |
| He et al. (2019) | Feature + Similarity | DeepLab-R101 (71.40) | DeepLab-MV2 (70.20) | 72.70 (+2.50) | – |
| IFVD (Wang et al., 2020d) | Similarity + Logit | PSPNet-R101 (79.76) | PSPNet-R18 (72.65) | 74.55 (+1.90) | Link |
| | | PSPNet-R101 (79.76) | DeepLab-R18 (74.96) | 76.01 (+1.05) | |
| CWD (Shu et al., 2021) | Feature | PSPNet-R101 (79.76) | PSPNet-R18 (72.65) | 75.91 (+3.26) | Link |
| | | PSPNet-R101 (79.76) | DeepLab-R18 (74.96) | 77.13 (+2.17) | |
| DSD (Feng et al., 2021) | Similarity | DeepLab-R101 (78.23) | DeepLab-R18 (69.42) | 72.24 (+2.82) | – |
| CIRKD (Yang et al., 2022c) | Similarity + Logit | DeepLab-R101 (78.07) | DeepLab-R18 (74.21) | 76.38 (+2.27) | Link |
| | | DeepLab-R101 (78.07) | PSPNet-R18 (72.55) | 74.73 (+2.18) | |
| DIST (Huang et al., 2022a) | Similarity + Logit | DeepLab-R101 (78.07) | DeepLab-R18 (74.21) | 77.10 (+2.89) | Link |
| | | DeepLab-R101 (78.07) | PSPNet-R18 (72.55) | 76.31 (+3.76) | |
| IDD (Zhang et al., 2022d) | Similarity | PSPNet-R101 (78.50) | PSPNet-R18 (70.09) | 77.59 (+7.5) | – |
| | | PSPNet-R101 (78.50) | ESPNet (61.40) | 68.87 (+7.47) | |
| MGD (Yang et al., 2022f) | Feature | PSPNet-R101 (78.34) | PSPNet-R18 (69.85) | 73.63 (+3.78) | Link |
| | | PSPNet-R101 (78.34) | DeepLab-R18 (73.20) | 76.31 (+3.11) | |
| MLP (Liu et al., 2023c) | Feature | PSPNe-R101 (78.34) | PSPNet-R18 (69.85) | 74.51 (+4.66) | – |
| | | PSPNet-R101 (78.34) | DeepLab-R18 (73.20) | 76.55 (+3.35) | |
| DiffKD (Huang et al., 2023b) | Feature | DeepLab-R101 (78.07) | DeepLab-R18 (74.21) | 77.78 (+3.57) | Link |
| | | DeepLab-R101 (78.07) | PSPNet-R18 (72.55) | 75.83 (+3.28) | |
| FAKD (Yuan et al., 2024a) | Feature | PSPNet-R101 (79.74) | PSPNet-R18 (68.99) | 74.75 (+5.76) | – |
| | | DeepLab-R101 (80.98) | DeepLab-MV2 (70.49) | 74.08 (+3.59) | |
| BPKD (Liu et al., 2024d) | Logit | PSPNet-R101 (79.74) | PSPNet-R18 (74.23) | 77.57 (+3.34) | Link |
| | | DeepLab-R101 (80.98) | DeepLab-MV2 (75.29) | 78.59 (+3.30) | |
| LAD (Liu et al., 2024g) | Feature + Logit | PSPNet-R101 (79.76) | PSPNet-R18 (72.65) | 76.86 (+4.21) | – |
| | | PSPNet-R101 (79.76) | DeepLab-R18 (74.96) | 77.23 (+2.27) | |
| FreeKD (Zhang et al., 2024i) | Feature | PSPNet-R101 (78.34) | PSPNet-R18 (69.85) | 74.40 (+4.55) | Link |
| | | PSPNet-R101 (78.34) | DeepLab-R18 (73.20) | 76.45 (+3.25) | |
| AttnFD (Mansourian et al., 2024) | Feature | DeepLab-R101 (77.66) | DeepLab-R18 (64.09) | 73.04 (+8.96) | Link |
| | | DeepLab-R101 (77.66) | PSPNet-R18 (65.72) | 68.86 (+3.14) | |
| AICSD (Mansourian et al., 2025) | Similarity + Logit | DeepLab-R101 (77.66) | DeepLab-R18 (64.09) | 70.96 (+6.88) | Link |
| | | DeepLab-R101 (77.66) | DeepLab-MV2 (63.05) | 69.51 (+6.56) | |

Along the same line, Table 16 presents a comparison of distillation methods for LLMs. It shows the compression rate and the improvement over the teacher model for each method. As can be seen, distillation in LLMs is more crucial than in other fields, as it allows for compressing large models with a high compression rate while still maintaining comparable performance. It should be noted that in cases where the exact performance of the teacher model is not reported, performance of the student model before and after distillation is reported (denoted by *).

Table 16: Performance comparison of different Black-box and White-box distillation methods for LLMs (* denotes that the comparison is between the student model before and after distillation).

| Method | Distillation Type | Teacher Model | Compression Rate | Benchmark | Comparison with Teacher | | Code |
|---|---|---|---|---|---|---|---|
| **White-box Distillation** | | | | | | | |
| DistillBERT (Sanh, 2019) | Logit | $BERT_{base}$ | 1.66 | GLUE | 77.00 / 79.50 | 97.0% | – |
| TinyBERT (Jiao et al., 2019) | Feature | $BERT_{base}$ | 7.50 | GLUE | 77.00 / 79.50 | 97.0% | Link |
| PKD (Sun et al., 2019) | Logit + Feature | $BERT_{base}$ | 2.00 | GLUE | 77.70 / 84.90 | 92.0% | Link |
| PD (Turc et al., 2019) | Logit | $BERT_{base}$ | 2.00 | GLUE | 82.10 / 81.70 | 100.5% | Link |
| Xtremedistil (Mukherjee & Awadallah, 2020) | Feature | $mBERT_{base}$ | 35.00 | Multilingual NER | 88.60 / 92.70 | 95.0% | Link |
| MobileBERT (Sun et al., 2020) | Feature | $BERT_{base}$ | 4.30 | GLUE | 77.70 / 78.30 | 99.0% | Link |
| MINILM (Wang et al., 2020b) | Feature | $BERT_{base}$ | 1.65 | GLUE | 80.40 / 81.50 | 98.0% | – |
| xtremedistiltransformers (Mukherjee et al., 2021) | Feature | $BERT_{base}$ | 7.70 | GLUE | 81.70 / 83.70 | 97.6% | Link |
| MetaDistil (Zhou et al., 2021b) | Feature | $BERT_{base}$ | 2.00 | GLUE | 80.40 / 80.70 | 99.0% | Link |
| TED (Liang et al., 2023a) | Feature | $DeBERTaV3_{base}$ | 2.60 | GLUE | 87.50 / 88.90 | 98.0% | Link |
| MiniLLM (Gu et al., 2024) | Feature | GPT-2 | 2.00 | DollyEval | 57.40 / 58.40 | 94.0% | Link |
| PC-LoRA (Hwang et al., 2024) | Feature | $BERT_{base}$ | 3.73 | IMDB | 92.78 / 94.00 | 98.0% | – |
| GKD (Agarwal et al., 2024) | Logit | $T5_{XL}$ | 3.75 | GSM8K | 29.10 / 29.00 | 100.3% | – |
| **Black-box Distillation** | | | | | | | |
| ILD (Huang et al., 2022e) | ICL | $GPT2_{large}$ | 6.00 | CrossFit | 61.20 / 66.20 | 92.0% | – |
| LLM-R (Wang et al., 2023h) | ICL | $LLaMA_{13B}$ | 2.00 | CommonsenseQA | 48.80 / 49.60 | 98.0% | – |
| AICD (Liu, 2024) | ICL | GPT-3.5-Turbo | – | SVAMP | 51.70 / 20.70 | 250.0%* | – |
| UniversalNER (Zhou et al., 2023c) | IF | GPT-3.5-Turbo | – | UNIVERSAL NER | 41.70 / 34.90 | 119.0% | Link |
| LaMini-LM (Wu et al., 2023c) | IF | $Alpaca_{7B}$ | 9.00 | OpenBookQA | 34.00 / 43.20 | 79.0% | Link |
| Lion (Jiang et al., 2023b) | IF | $GPT_{175B}$ | 25.00 | BBH | 32.00 / 48.90 | 65.0% | Link |
| Fine-tune-CoT (Ho et al., 2022) | CoT | $GPT_{175B}$ | 26.00 | SVAMP | 30.33 / 64.67 | 47.0% | Link |
| Li et al. (Li et al., 2022e) | CoT | $GPT_{175B}$ | 58.00 | CommonsenseQA | 82.47 / 73.00 | 113.0% | – |
| Magister et al. (Magister et al., 2022) | CoT | $GPT_{175B}$ | 16.00 | StrategyQA | 63.77 / 65.40 | 97.0% | – |
| MCC-KD (Chen et al., 2023a) | CoT | GPT-3.5-Turbo | – | SVAMP | 68.66 / 75.14 | 91.0% | Link |
| CSoTD (Li et al., 2023d) | CoT | $GPT_{175B}$ | 135.00 | CommonsenseQA | 67.00 / 82.10 | 82.0% | – |
| SCOTT (Wang et al., 2023j) | CoT | $GPTneox_{20B}$ | 7.00 | CommonsenseQA | 74.70 / 60.40 | 124.0% | Link |
| Distilling step-by-step (Hsieh et al., 2023) | CoT | $Palm_{540B}$ | 700.00 | SVAMP | 58.40 / 72.30 | 81.0% | Link |
| PaD (Zhu et al., 2023b) | CoT | GPT-3.5-Turbo | – | SVAMP | 36.70 / 57.80 | 64.0% | – |
| TDG (Li et al., 2024g) | CoT | GPT-3.5-Turbo | – | MATH | 6.81 / 3.88 | 175.0%* | Link |
| RevThink (Chen et al., 2024c) | CoT | Gemini-1.5-Pro-001 | 30.00 | CommonsenseQA | 75.76 / 76.72 | 99.0% | – |

# 8 Discussion

## 8.1 Principles of Method Classification

To structure this survey, the reviewed methods are categorized according to their primary contributions across five dimensions: distillation sources, schemes, algorithms, modalities, and applications. While many approaches span multiple categories, each method is assigned based on its central innovation, that is, the main idea or technical novelty the paper introduces, with relevant cross-references provided where appropriate. This principled organization aims to balance clarity with the multi-dimensional nature of knowledge distillation research.

## 8.2 Guidelines for Selecting Knowledge Distillation Strategies

To complement the presented classification, the survey also offers practical guidance for selecting appropriate KD strategies based on common task-specific constraints. While modality and application are typically determined by the nature of the target domain, the selection of distillation source, scheme, and algorithm remains a critical aspect of model design. Table 17 provides a decision-oriented summary to map task scenarios to suitable KD components. This perspective bridges the gap between methodological classification

and practical usage, offering a starting point for design decisions in KD, while recognizing that the optimal choice ultimately depends on the specific goals, constraints, and context of each use case.

Table 17: Summary of key task constraints in knowledge distillation, with corresponding guidance for selecting suitable strategies, including KD sources, schemes, and suitable algorithms.

| **Task Requirement** | **KD Source(s)** | **KD Scheme(s)** | **KD Algorithm(s)** |
|---|---|---|---|
| Real-time Inference | Logit-based | Online | Attention-based<br>Adaptive |
| Data Scarcity / Few Shot | Feature-based | Self-distillation | Graph-based<br>Contrastive |
| Model Interpretability | Feature-based | Offline | Attention-based<br>Graph-based |
| Fine-grained Recognition Tasks | Feature-based<br>Similarity-based | Offline | Attention-based<br>Contrastive |
| Limited Compute Budget | Logit-based | Offline | Adaptive |
| Adversarial Robustness | Logit-based | Offline | Adversarial |
| Multimodal Fusion | Feature-based | Offline | Cross-modal |
| Continual/Incremental Learning | Logit-based<br>Similarity-based | Online | Adaptive |
| Multi-task Learning | Logit-based<br>Feature-based | Online | Adaptive<br>Multi-teacher |
| Learning from Noisy Labels | Similarity-based | Self-distillation | Contrastive |

### 8.3 Quantitative Comparison

Quantitative comparison of existing distillation methods is very challenging. The performance of each method is influenced by numerous hyperparameters and is highly dependent on the weights of the pre-trained teacher model. As a result, providing a clear and fair comparison is nearly impossible, and reported performance metrics often differ significantly from those in the original papers. Therefore, we have intentionally avoided making direct claims about which method performs better than others.

However, in each section of the survey, we have highlighted the strengths and advantages of each category over others. In addition, Tables 12, 13, 14, 15, and 16 report the performance of various distillation methods across different domains, tasks, and datasets. Where available, we also include results for cases where the teacher and student models have either similar or different architectures. This information enables meaningful comparison, as it demonstrates the impact of varying dataset sizes and the robustness of each method to architectural differences.

### 8.4 Practical and Ethical Aspects

KD can also be studied from various other perspectives, including its use cases beyond model compression, such as enhancing privacy, improving security, and addressing intellectual property concerns. While this is beyond the scope of this paper, it is important to highlight some of the key challenges in these areas.

Recent studies have explored the use of KD to enhance privacy and prevent data leakage. For instance, Fernandez et al. (2023) employed KD to mitigate data memorization in text-to-image generative models, Flemings & Annavaram (2024) applied it to reduce the risk of LLMs generating sensitive information, and Shao et al. (2024a) utilized KD to train a large teacher model without exposing the data of smaller edge devices. On the other hand, Zhao & Zhang (2025) demonstrated that techniques like Data-Free Knowledge

Distillation (DFKD) do not inherently guarantee privacy, and Cui et al. (2025) showed that such approaches may even introduce new vulnerabilities in LLMs. KD can also be used to guide the student model—intended as the final model—toward more fair and equitable behavior. (Chai et al., 2022)

Another important real-world consideration of KD is its potential to reduce computational demands, particularly relevant in the context of LLMs, which typically require substantial processing power. By enabling student models to achieve comparable performance with significantly lower resource consumption, KD can contribute to more energy-efficient and environmentally sustainable AI systems (Yuan et al., 2024c). However, the use of commercial models as teachers in the distillation process may raise legal concerns, as it could violate copyright protections. This limitation must be carefully considered in both research and deployment contexts.

### 8.5 Challenges

While KD is an effective method for improving the performance of smaller networks, it comes with several challenges. Below, some of the most important challenges are discussed.

**Knowledge Extraction**: Finding an appropriate source of knowledge for distillation is one of the main challenges in the distillation process (Heo et al., 2019a; Hsu et al., 2022; Ojha et al., 2023). Although using logits is the most straightforward approach to distillation, its effectiveness varies across different scenarios. For example, in tasks where labels are not available, logit-based distillation is not feasible. Furthermore, with the emergence of foundation models like CLIP, distilling rich features and embeddings has become crucial, which is not achievable through logit-based distillation alone, requiring feature-based and similarity-based distillation methods. In LLMs, this issue is even more pronounced, as some recent LLMs are not open-source (Achiam et al., 2023; Team et al., 2023), restricting access to intermediate features or logits, and making black-box distillation particularly challenging. Existing methods often leverage a combination of different distillation sources to effectively transfer the teacher model's knowledge to the student model.

**Choosing Proper Distillation Scheme**: Selecting an appropriate teacher-student architecture is another challenge, as it depends on the specific context. In scenarios where a pre-trained large model is available, offline distillation tends to be more effective, whereas, in other cases, online distillation may prove to be a better alternative. In self-supervised learning, self-distillation schemes have shown to be highly effective. Additionally, using multiple teachers to train a smaller network is a promising approach, often referred to as a mixture of experts. However, this comes with its own challenges, including the need for more powerful memory and GPUs, as well as the complexity of loading different teachers onto separate resources. In most existing methods, offline distillation is more commonly used. Specifically, when distilling from foundation models or LLMs, training the teacher alongside the student is not feasible due to the immense number of parameters in the teacher model. Table 17 presents recommended distillation strategies tailored to specific conditions.

**Capacity Gap**: Another important challenge is the capacity gap between the teacher and student models. It may be assumed that a larger teacher always leads to higher performance improvement in the student network, but when the teacher and student networks have a considerable size difference, it can cause capacity gap issues (Cho & Hariharan, 2019; Guo et al., 2020a; Huang et al., 2022a). This is because the receptive fields of the teacher and student models differ significantly. This challenge is even more pronounced in foundation models, as they contain a tremendous number of parameters, making the distillation of their knowledge into a smaller model challenging. Furthermore, in black-box distillation of LLMs, where intermediate layers are not accessible, directly transferring knowledge from the teacher to the student is a serious challenge. Existing methods employ various algorithms to reduce this gap and make the teacher's knowledge more comprehensible for the student.

**Architectural Difference**: Differences in the architecture of the teacher and student models present another significant challenge, which can impact the performance of the distillation method. In scenarios where the teacher and student have different architectures, which is very common, transferring proper knowledge becomes a new challenge. This issue becomes apparent when distilling knowledge from a teacher to a student with a different architecture, often leading to performance degradation (Cho & Hariharan, 2019; Guo et al., 2020a). In foundation models, the student's architecture can differ significantly from that of the teacher's.

Although logit-based distillation remains effective in such scenarios, feature-based and similarity-based distillation methods face additional challenges. More specifically, selecting appropriate layers with similar semantic information for feature distillation plays an important role in the performance of the method, as the teacher and student may even have different inductive biases. In black-box distillation for LLMs, CoT distillation can help address this problem, as it does not force the student to mimic the exact intermediate layers of the teacher. Instead, it provides the student with intermediate reasoning steps derived from the teacher. However, generating proper reasoning and addressing the capacity gap between the teacher and student remain significant challenges. Tables 12, 13, 14, and 15 present the effect of teacher-student architectural differences on the performance of knowledge distillation methods.

### 8.6 Future Directions

Although KD is not a new concept and has been around since its initial proposal, the number of new methods being proposed is rapidly increasing, along with its expanding applications in various fields. This presents a significant opportunity for further research. In the following, potential future directions are discussed.

**Feature-based Distillation**: While logit-based and similarity-based distillation are effective, recent trends and results show that feature-based distillation can lead to even greater improvements in student performance. However, it can also be combined with other methods. Recent feature-based distillation approaches have demonstrated that instead of defining complex relationships, a simple transformation on the features of either the teacher or student is sufficient to achieve SOTA results in enhancing the student's performance (Liu et al., 2023c; 2024g). Furthermore, with the rapid growth of foundation models like CLIP, distilling embeddings from these large-scale models is gaining increasing attention.

**Adaptive Distillation**: Despite the existence of numerous distillation approaches, effectively distilling informative knowledge from the teacher to the student remains largely unexplored (Hsu et al., 2022; Ojha et al., 2023). Several adaptive distillation methods have been proposed to effectively transfer the dark knowledge of the teacher while ignoring unused knowledge. This line of research is particularly interesting, as it prevents the distillation of incorrect knowledge to the student and helps reduce the gap between the teacher and student networks. This process closely resembles human learning: a teacher transfers knowledge to students, but directly transferring all knowledge is not always effective or practical (Cho & Hariharan, 2019; Mirzadeh et al., 2020). Therefore, various forms of adaptive distillation can be employed, such as using a teacher assistant, decreasing the impact of distillation toward the end of training, or applying adaptive loss functions based on the importance of samples. All of these approaches align with real-world learning scenarios in human training processes.

**Distillation from Foundation Models**: With the emergence of a variety of foundation and multi-modal models, they have become a rich source of knowledge for distillation (Zhang et al., 2024d). For example, CLIP, with its image and text encoder, can provide student models with rich embeddings, where text can be used to improve vision tasks such as weakly-supervised semantic segmentation, making it a hot topic in research (Xie et al., 2022). Additionally, due to the general-purpose features of these large-scale models, they can be leveraged for open-vocabulary tasks (Liu et al., 2024f), where the task is not limited to predefined classes. These interesting applications of distillation present great potential for future research.

**Distillation in LLMs**: While distillation is a highly effective approach in vision, text, and speech tasks, its application in LLMs is even more important. As stated earlier, LLMs have been shown to be overparameterized (Kaplan et al., 2020; Fischer et al., 2024) with billions and even trillions of parameters, far exceeding the size of models in other fields. These weight matrices are often sparse (Hu et al., 2021), making distillation crucial in reducing model size while maintaining performance. Additionally, most powerful models are not open-source (Achiam et al., 2023; Team et al., 2023), increasing the importance of black-box distillation. CoT distillation is a particularly interesting technique for transferring reasoning capabilities from an LLM to an SLM. Unlike conventional distillation methods, CoT aims to enhance the generalization ability of the student model by teaching it how to reason. Results show that by distilling the knowledge of an LLM, the student model can achieve performance comparable to the teacher while significantly reducing model size. This is especially important because the teacher model may be too large to run on common GPUs, whereas

the distilled student model can be efficiently deployed on common hardware, making advanced AI more accessible to to a wider audience.

Besides its role in training SLMs, a promising new direction is the use of KD during the pre-training stage of LLMs in industry (Yang et al., 2025; Team et al., 2025). Training LLMs of different sizes is highly cost-intensive. KD plays an important role in creating smaller versions of a pre-trained LLM, where techniques such as quantization (Liu et al., 2023b) or pruning (Sreenivas et al., 2024; Muralidharan et al., 2024) are applied to the large model to extract a compact student model. This process can be used in a post-training setting, requiring only a small fraction of the dataset originally used to train the teacher model.

**Applications of Distillation**: In addition to proposing novel distillation methods, distillation has been applied across various tasks, resulting in significant performance improvements. For example, distillation has been successfully applied in adversarial attacks (Papernot et al., 2016b; Goldblum et al., 2020), mitigating backdoor attacks (Li et al., 2021a; Han et al., 2024b), and anomaly detection (Salehi et al., 2021). Furthermore, the concept of distillation is utilized in dataset distillation (Wang et al., 2018a), making it an interesting research direction. Another application of distillation is its integration with other compression methods, such as quantization or low-rank factorization.

## 9 Conclusion

Knowledge distillation has revolutionized the field of model compression and has played a significant role in various research topics in recent years. In this work, a comprehensive survey on knowledge distillation was proposed, reviewing its methods from various perspectives, including: sources, schemes, algorithms, modalities, applications, and performance comparisons. In contrast to existing surveys, this survey presented recent advancements in the field and focuses on the most important topics in KD. Adaptive and contrastive distillation were presented as two important algorithms that have gained increased attention in recent years. Additionally, KD was reviewed across different modalities, particularly for 3D inputs such as point clouds, which have established important connections with KD in recent research. Furthermore, the applications of distillation in self-supervised learning, diffusion models, foundation models, and LLMs was explored. Self-supervised learning methods primarily rely on self-distillation techniques, diffusion models utilize distillation to reduce the number of steps in the generation phase, and LLMs are distilled into smaller versions KD was reviewed across different modalities, particularly for 3D inputs such as point clouds, which have established important connections with KD in recent research. Furthermore, with recent advancements in foundation models like CLIP, distilling their rich knowledge into other models has become increasingly widespread.

## References

Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Compress: Self-supervised learning by compressing representations. *Advances in Neural Information Processing Systems*, 33:12980–12992, 2020.

Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–9, 2022.

Cihan Acar, Kuluhan Binici, Alp Tekirdağ, and Yan Wu. Visual-policy learning through multi-camera view to single-camera view knowledge distillation for robot manipulation tasks. *IEEE Robotics and Automation Letters*, 9(1):691–698, 2023.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alen Adamyan and Erik Harutyunyan. Smaller3d: Smaller models for 3d semantic segmentation using minkowski engine and knowledge distillation methods. *arXiv preprint arXiv:2305.03188*, 2023.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2143–2147. IEEE, 2020.

Rishabh Agarwal, Nino Vieillard, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. Gkd: Generalized knowledge distillation for auto-regressive sequence models. *arXiv preprint arXiv:2306.13649*, 2023.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024.

Nima Aghli and Eraldo Ribeiro. Combining weight pruning and knowledge distillation for cnn compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3191–3198. Computer Vision Foundation / IEEE, 2021.

Angeline Aguinaldo, Ping-Yeh Chiang, Alex Gain, Ameya Patil, Kolten Pearson, and Soheil Feizi. Compressing gans using knowledge distillation. *arXiv preprint arXiv:1902.00159*, 2019.

Rozhan Ahmadi and Shohreh Kasaei. Leveraging swin transformer for local-to-global weakly supervised semantic segmentation. In *2024 13th Iranian/3rd International Machine Vision and Image Processing Conference (MVIP)*, pp. 1–7. IEEE, 2024.

Sidra Aleem, Fangyijie Wang, Mayug Maniparambil, Eric Arazo, Julia Dietlmeier, Kathleen Curran, Noel EO' Connor, and Suzanne Little. Test-time adaptation with salip: A cascade of sam and clip for zero-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5184–5193, 2024.

Abdolmaged Alkhulaifi, Fahad Alsahli, and Irfan Ahmad. Knowledge distillation in deep learning and its applications. *PeerJ Computer Science*, 7:e474, 2021.

Niki Amini-Naieni, Tengda Han, and Andrew Zisserman. Countgd: Multi-modal open-world counting. *Advances in Neural Information Processing Systems*, 37:48810–48837, 2024.

Abdollah Amirkhani, Amir Khosravian, Masoud Masih-Tehrani, and Hossein Kashiani. Robust semantic segmentation with multi-teacher knowledge distillation. *IEEE Access*, 9:119049–119066, 2021.

Shumin An, Qingmin Liao, Zongqing Lu, and Jing-Hao Xue. Efficient semantic segmentation via self-attention and self-distillation. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):15256–15266, 2022.

Alex Andonian, Shixing Chen, and Raffay Hamid. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16430–16441, 2022.

Simone Angarano, Francesco Salvetti, Mauro Martini, and Marcello Chiaberge. Generative adversarial super-resolution at the edge with knowledge distillation. *Engineering Applications of Artificial Intelligence*, 123: 106407, 2022.

Anshumann Anshumann, Mohd Abbas Zaidi, Akhil Kedia, Jinwoo Ahn, Taehwak Kwon, Kangwook Lee, Haejun Lee, and Joohyung Lee. Sparse logit sampling: Accelerating knowledge distillation in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18085–18108, 2025.

Umar Asif, Jianbin Tang, and Stefan Harrer. Ensemble knowledge distillation for learning improved and efficient networks. In *ECAI 2020*, pp. 953–960. IOS Press, 2020.

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27, 2014.

Ziga Babnik, Fadi Boutros, Naser Damer, Peter Peer, and Vitomir Struc. Ai-kd: Towards alignment invariant face image quality assessment using knowledge distillation. In *IWBF*, pp. 1–6. IEEE, 2024.

Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan L Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24256–24265, 2023.

Jawid Ahmad Baktash and Mursal Dawodi. Gpt-4: A review on advancements and opportunities in natural language processing. *arXiv preprint arXiv:2305.03195*, 2023.

Geonho Bang, Kwangjin Choi, Jisong Kim, Dongsuk Kum, and Jun Won Choi. Radardistill: Boosting radar-based object detection performance via knowledge distillation from lidar features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15491–15500, 2024.

Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35:33781–33794, 2022.

Nader Karimi Bavandpour and Shohreh Kasaei. Class attention map distillation for efficient semantic segmentation. In *2020 International Conference on Machine Vision and Image Processing (MVIP)*, pp. 1–6. IEEE, 2020.

Vasileios Belagiannis, Azade Farshad, and Fabio Galasso. Adversarial network compression. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Tariq Berrada, Camille Couprie, Karteek Alahari, and Jakob Verbeek. Guided distillation for semi-supervised instance segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472. IEEE, 2024.

Shweta Bhardwaj, Mukundhan Srinivasan, and Mitesh M Khapra. Efficient video classification using fewer frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 354–363, 2019.

Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, and Yi-Zhe Song. Text is text, no matter what: Unifying text recognition using knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 983–992, 2021.

Cunling Bian, Wei Feng, Liang Wan, and Song Wang. Structural knowledge distillation for efficient skeleton-based action recognition. *IEEE Transactions on Image Processing*, 30:2963–2976, 2021.

Fadi Boutros, Vitomir Struc, and Naser Damer. Adadistill: Adaptive knowledge distillation for deep face recognition. In *ECCV*, volume 15113 of *Lecture Notes in Computer Science*, pp. 163–182. Springer, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Mateusz Budnik and Yannis Avrithis. Asymmetric metric learning for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. Computer Vision Foundation / IEEE, 2021.

Eduarda Caldeira, Jaime S. Cardoso, Ana Filipa Sequeira, and Pedro Neto. MST-KD: Multiple specialized teachers knowledge distillation for fair face recognition. In *Proceedings of the 7th Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW) at ECCV*, 2024.

Fernando Camarena, Miguel Gonzalez-Mendoza, and Leonardo Chang. Knowledge distillation in video-based human action recognition: An intuitive approach to efficient and flexible model training. *Journal of Imaging*, 10(4):85, 2024.

Zehranaz Canfes, M Furkan Atasoy, Alara Dirik, and Pinar Yanardag. Text and image guided 3d avatar generation and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4421–4431, 2023.

Shengcao Cao, Mengtian Li, James Hays, Deva Ramanan, Yu-Xiong Wang, and Liangyan Gui. Learning lightweight object detectors via multi-teacher progressive distillation. In *International Conference on Machine Learning*, pp. 3577–3598. PMLR, 2023.

Weiwei Cao, Jianpeng Zhang, Yingda Xia, Tony C. W. Mok, Zi Li, Xianghua Ye, Le Lu, Jian Zheng, Yuxing Tang, and Ling Zhang. Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11238–11247. IEEE, 2024.

Umberto Cappellazzo, Daniele Falavigna, and Alessio Brutti. An investigation of the combination of rehearsal and knowledge distillation in continual learning for spoken language understanding. *arXiv preprint arXiv:2211.08161*, 2022.

Umberto Cappellazzo, Muqiao Yang, Daniele Falavigna, and Alessio Brutti. Sequence-level knowledge distillation for class-incremental end-to-end spoken language understanding. *arXiv preprint arXiv:2305.13899*, 2023.

Adriano Cardace, Riccardo Spezialetti, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Self-distillation for unsupervised 3d domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4166–4177, 2023.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Jun Cen, Shiwei Zhang, Yixuan Pei, Kun Li, Hang Zheng, Maochun Luo, Yingya Zhang, and Qifeng Chen. Cmdfusion: Bidirectional fusion network with cross-modality knowledge distillation for lidar semantic segmentation. *IEEE Robotics and Automation Letters*, 9(1):771–778, 2023.

Hyungjoo Chae, Yongho Song, Kai Tzu-iunn Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. *arXiv preprint arXiv:2310.09343*, 2023.

Junyi Chai, Taeuk Jang, and Xiaoqian Wang. Fairness without demographics through knowledge distillation. *Advances in Neural Information Processing Systems*, 35:19152–19164, 2022.

Chun-Chieh Chang, Chieh-Chi Kao, Ming Sun, and Chao Wang. Intra-utterance similarity preserving knowledge distillation for audio tagging. *arXiv preprint arXiv:2009.01759*, 2020. URL `https://arxiv.org/abs/2009.01759`.

Jiahao Chang, Shuo Wang, Hai-Ming Xu, Zehui Chen, Chenhongyi Yang, and Feng Zhao. Detrdistill: A universal knowledge distillation framework for detr-families. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6898–6908, 2023.

Yevgen Chebotar and Austin Waters. Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech*, pp. 3439–3443, 2016.

Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 3430–3437, 2020a.

Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 7028–7036, 2021a.

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11933–11942, 2022a.

Dong Chen, Ning Liu, Yichen Zhu, Zhengping Che, Rui Ma, Fachao Zhang, Xiaofeng Mou, Yi Chang, and Jian Tang. Epsd: Early pruning with self-distillation for efficient model compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11258–11266, 2024a.

Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3514–3522, 2019a.

Hanting Chen, Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao. Learning student networks via feature embedding. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):25–35, 2020b.

Hongzhan Chen, Siyue Wu, Xiaojun Quan, Rui Wang, Ming Yan, and Ji Zhang. Mcc-kd: Multi-cot consistent knowledge distillation. *arXiv preprint arXiv:2310.14747*, 2023a.

Jiade Chen, Jin Wang, Yunhui Shi, Nam Ling, and Baocai Yin. Mvp-net: Multi-view depth image guided cross-modal distillation network for point cloud upsampling. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9759–9768, 2024b.

Jun Chen, Deyao Zhu, Guocheng Qian, Bernard Ghanem, Zhicheng Yan, Chenchen Zhu, Fanyi Xiao, Sean Chang Culatana, and Mohamed Elhoseiny. Exploring open-vocabulary semantic segmentation from clip vision encoder distillation only. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 699–710, 2023b.

Justin Chih-Yao Chen, Zifeng Wang, Hamid Palangi, Rujun Han, Sayna Ebrahimi, Long Le, Vincent Perot, Swaroop Mishra, Mohit Bansal, Chen-Yu Lee, et al. Reverse thinking makes llms stronger reasoners. *arXiv preprint arXiv:2411.19865*, 2024c.

Kecheng Chen, Tiexin Qin, Victor Ho Fun Lee, Hong Yan, and Haoliang Li. Learning robust shape regularization for generalizable medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(7): 2693–2706, 2024d.

Kuan-Yu Chen, Shih-Hung Liu, Berlin Chen, and Hsin-Min Wang. An information distillation framework for extractive summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26 (1):161–170, 2017.

Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16296–16305, 2021b.

Runze Chen, Haiyong Luo, Fang Zhao, Jingze Yu, Yupeng Jia, Juan Wang, and Xuepeng Ma. Structure-centric robust monocular depth estimation via knowledge distillation. In *ACCV*, volume 15480 of *Lecture Notes in Computer Science*, pp. 123–140. Springer, 2024e.

Sichen Chen, Yingyi Zhang, Siming Huang, Ran Yi, Ke Fan, Ruixin Zhang, Peixian Chen, Jun Wang, Shouhong Ding, and Lizhuang Ma. Sdpose: Tokenized pose estimation via circulation-guide self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1082–1090, 2024f.

Tianle Chen, Zheda Mai, Ruiwen Li, and Wei-lun Chao. Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. *arXiv preprint arXiv:2305.05803*, 2023c.

Tiansheng Chen and Lingfei Mo. Swin-fusion: swin-transformer with feature fusion for human action recognition. *Neural Processing Letters*, 55(8):11109–11130, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020c.

Xingjian Chen, Jianbo Su, and Jun Zhang. A two-teacher framework for knowledge distillation. In *Advances in Neural Networks–ISNN 2019: 16th International Symposium on Neural Networks, ISNN 2019, Moscow, Russia, July 10–12, 2019, Proceedings, Part I 16*, pp. 58–66. Springer, 2019b.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020d.

Xiuyi Chen, Guangcan Liu, Jing Shi, Jiaming Xu, and Bo Xu. Distilled binary neural network for monaural speech separation. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2018.

Xuanang Chen, Ben He, Kai Hui, Le Sun, and Yingfei Sun. Simplified tinybert: Knowledge distillation for document retrieval. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pp. 241–248. Springer, 2021c.

Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7016–7025, 2021d.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. Distilling knowledge learned in bert for text generation. *arXiv preprint arXiv:1911.03829*, 2019c.

Yifan Chen, Xiaozhen Qiao, Zhe Sun, and Xuelong Li. Comkd-clip: Comprehensive knowledge distillation for contrastive language-image pre-traning model. *arXiv preprint arXiv:2408.04145*, 2024g.

Yifei Chen, Binfeng Zou, Zhaoxin Guo, Yiyu Huang, Yifan Huang, Feiwei Qin, Qinhai Li, and Changmiao Wang. Scunet++: Swin-unet and cnn bottleneck hybrid architecture with multi-fusion dense skip connection for pulmonary embolism ct image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7759–7767, 2024h.

Yixin Chen, Pengguang Chen, Shu Liu, Liwei Wang, and Jiaya Jia. Deep structured instance graph for distilling object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4359–4368, 2021e.

Yuzhao Chen, Yatao Bian, Xi Xiao, Yu Rong, Tingyang Xu, and Junzhou Huang. On self-distilling graph neural network. *arXiv preprint arXiv:2011.02255*, 2020e.

Zailiang Chen, Xianxian Zheng, Hailan Shen, Ziyang Zeng, Yukun Zhou, and Rongchang Zhao. Improving knowledge distillation via category structure. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pp. 205–219. Springer, 2020f.

Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv preprint arXiv:2211.09386*, 2022b.

Zhaozheng Chen and Qianru Sun. Weakly-supervised semantic segmentation with image-level labels: from traditional models to foundation models. *ACM Computing Surveys*, 2023.

Zining Chen, Weiqiu Wang, Zhicheng Zhao, Fei Su, Aidong Men, and Hongying Meng. Practicaldg: Perturbation distillation on vision-language models for hybrid domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23501–23511, 2024i.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

Cheng Cheng, Wei Gao, and Xiaohang Bian. Knowledge distillation via inter-and intra-samples relation transferring. In *2024 20th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pp. 1–7. IEEE, 2024a.

Xin Cheng, Zhiqiang Zhang, Wei Weng, Wenxin Yu, and Jinjia Zhou. De-mkd: Decoupled multi-teacher knowledge distillation based on entropy. *Mathematics*, 12(11):1672, 2024b.

Zhihao Chi, Tu Zheng, Hengjia Li, Zheng Yang, Boxi Wu, Binbin Lin, and Deng Cai. Normkd: Normalized logits for knowledge distillation. *arXiv preprint arXiv:2308.00520*, 2023.

Hyeon Cho, Junyong Choi, Geonwoo Baek, and Wonjun Hwang. itkd: Interchange transfer-based knowledge distillation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13540–13549, 2023.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4794–4802, 2019.

Yubin Cho and Sukju Kang. Class attention transfer for semantic segmentation. In *2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pp. 41–45. IEEE, 2022.

Jiho Choi, Seonho Lee, Seungho Lee, Minhyun Lee, and Hyunjung Shim. Understanding multi-granularity for open-vocabulary part segmentation. *Advances in Neural Information Processing Systems*, 37:137161–137189, 2024.

Yoojin Choi, Jihwan Choi, Mostafa El-Khamy, and Jungwon Lee. Data-free network quantization with adversarial knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 710–711, 2020.

Inseop Chung, SeongUk Park, Jangho Kim, and Nojun Kwak. Feature-map-level online adversarial knowledge distillation. In *International Conference on Machine Learning*, pp. 2006–2015. PMLR, 2020.

Florinel-Alin Croitoru, Nicolae-Cătălin Ristea, Dana Dăscălescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Lightning fast video anomaly detection via multi-scale adversarial distillation. *Computer Vision and Image Understanding*, 247:104074, 2024.

Jiequan Cui, Zhuotao Tian, Zhisheng Zhong, Xiaojuan Qi, Bei Yu, and Hanwang Zhang. Decoupled kullback-leibler divergence loss. *Advances in Neural Information Processing Systems*, 37:74461–74486, 2024.

Ziyao Cui, Minxing Zhang, and Jian Pei. On membership inference attacks in knowledge distillation. *arXiv preprint arXiv:2505.11837*, 2025.

Claudia Cuttano, Gabriele Rosi, Gabriele Trivigno, and Giuseppe Averta. What does clip know about peeling a banana? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2238–2247, 2024.

Amirhossein Dadashzadeh, Alan Whone, and Majid Mirmehdi. Auxiliary learning for self-supervised video representation via similarity-based knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4231–4240, 2022.

Rui Dai, Srijan Das, and François Brémond. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13033–13044. IEEE, 2021a.

Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7842–7851. Computer Vision Foundation / IEEE, 2021b.

Arnav M Das, Ritwick Chaudhry, Kaustav Kundu, and Davide Modolo. Prompting foundational models for omni-supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1583–1592, 2024.

Sayantan Dasgupta and Trevor Cohn. Improving language model distillation through hidden state matching. In *The Thirteenth International Conference on Learning Representations*, 2025.

Dale Decatur, Itai Lang, Kfir Aberman, and Rana Hanocka. 3d paintbrush: Local stylization of 3d shapes with cascaded score distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4473–4483, 2024.

Songhe Deng, Wei Zhuo, Jinheng Xie, and Linlin Shen. Qa-clims: Question-answer cross language image matching for weakly supervised semantic segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5572–5583, 2023.

Xiang Deng and Zhongfei Zhang. Graph-free knowledge distillation for graph neural networks. *arXiv preprint arXiv:2105.07519*, 2021.

Xing Di, Yiyu Zheng, Xiaoming Liu, and Yu Cheng. Pros: Facial omni-representation learning via prototype-based self-distillation. In *WACV*, pp. 6075–6086. IEEE, 2024.

Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11583–11592, 2022a.

Qianggang Ding, Sifan Wu, Hao Sun, Jiadong Guo, and Shu-Tao Xia. Adaptive regularization of labels. *arXiv preprint arXiv:1908.05474*, 2019.

Yifeng Ding, Gaoming Yang, Shuting Yin, Ji Zhang, Xianjin Fang, and Wencheng Yang. Generous teacher: Good at distilling knowledge for student learning. *Image and Vision Computing*, 150:105199, 2024.

Yikang Ding, Qingtian Zhu, Xiangyue Liu, Wentao Yuan, Haotian Zhang, and Chi Zhang. Kd-mvs: Knowledge distillation based self-supervised learning for multi-view stereo. In *European conference on computer vision*, pp. 630–646. Springer, 2022b.

Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang. Ced: Consistent ensemble distillation for audio tagging. *arXiv preprint arXiv:2308.11957*, 2023. URL `https://arxiv.org/abs/2308.11957`.

Kien Do, Thai Hung Le, Dung Nguyen, Dang Nguyen, Haripriya Harikumar, Truyen Tran, Santu Rana, and Svetha Venkatesh. Momentum adversarial distillation: Handling large distribution shifts in data-free knowledge distillation. *Advances in Neural Information Processing Systems*, 35:10055–10067, 2022.

Jianfeng Dong, Minsong Zhang, Zheng Zhang, Xianke Chen, Daizong Liu, Xiaoye Qu, Xun Wang, and Baolong Liu. Dual learning with dynamic knowledge distillation for partially relevant video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11302–11312, 2023a.

Junhao Dong, Piotr Koniusz, Junxi Chen, and Yew-Soon Ong. Adversarially robust distillation by reducing the student-teacher variance gap. In *European Conference on Computer Vision*, pp. 92–111. Springer, 2024a.

Junhao Dong, Piotr Koniusz, Junxi Chen, Z Jane Wang, and Yew-Soon Ong. Robust distillation via untargeted and targeted intermediate adversarial samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28432–28442, 2024b.

Wenhui Dong, Bo Du, and Yongchao Xu. Shape-intensity knowledge distillation for robust medical image segmentation. *Frontiers of Computer Science*, 19(9):199705, 2025.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10995–11005, 2023b.

Yue-Jiang Dong, Fang-Lue Zhang, and Song-Hai Zhang. Mal: Motion-aware loss with temporal and distillation hints for self-supervised depth estimation. In *ICRA*, pp. 7318–7324. IEEE, 2024c.

Yushun Dong, Binchi Zhang, Yiling Yuan, Na Zou, Qi Wang, and Jundong Li. Reliant: Fair knowledge distillation for graph neural networks. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pp. 154–162. SIAM, 2023c.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14084–14093, 2022.

Zhixing Du, Rui Zhang, Ming Chang, Xishan Zhang, Shaoli Liu, Tianshi Chen, and Yunji Chen. Distilling object detectors with feature richness. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 5213–5224, 2021.

Niladri Shekhar Dutt, Sanjeev Muralikrishnan, and Niloy J Mitra. Diffusion 3d features (diff3f): Decorating untextured shapes with distilled semantic features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4494–4504, 2024.

Omar S. El-Assiouti, Ghada Hamed, Dina Reda Khattab, and Hala Mousher Ebeid. Hdkd: Hybrid data-efficient knowledge distillation network for medical image classification. *Engineering Applications of Artificial Intelligence*, 138:109430, 2024.

Brunó B Englert, Fabrizio J Piva, Tommie Kerssies, Daan De Geus, and Gijs Dubbelman. Exploring the benefits of vision foundation models for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1172–1180, 2024.

Jiawei Fan, Chao Li, Xiaolong Liu, Meina Song, and Anbang Yao. Augmentation-free dense contrastive knowledge distillation for efficient semantic segmentation. *Advances in Neural Information Processing Systems*, 36:51359–51370, 2023.

Zhaoxin Fan, Yulin He, Zhicheng Wang, Kejian Wu, Hongyan Liu, and Jun He. Reconstruction-aware prior distillation for semi-supervised point cloud completion. *arXiv preprint arXiv:2204.09186*, 2022.

Gongfan Fang, Jie Song, Chengchao Shen, Xinchao Wang, Da Chen, and Mingli Song. Data-free adversarial distillation. *arXiv preprint arXiv:1912.11006*, 2019.

Hangxiang Fang, Yongwen Long, Xinyi Hu, Yangtao Ou, Yuanjia Huang, and Haoji Hu. Dual cross knowledge distillation for image super-resolution. *Journal of Visual Communication and Image Representation*, 95: 103858, 2023a.

Haoshu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7157–7173, 2023b.

Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*, 2021.

Chen Feng, Duolikun Danier, Haoran Wang, Fan Zhang, Benoit Vallade, Alex Mackin, and David Bull. Rankdvqa-mini: knowledge distillation-driven deep video quality assessment. In *2024 Picture Coding Symposium (PCS)*, pp. 1–5. IEEE, 2024a.

Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *European Conference on Computer Vision*, pp. 701–717. Springer, 2022.

Weilun Feng, Chuanguang Yang, Zhulin An, Libo Huang, Boyu Diao, Fei Wang, and Yongjun Xu. Relational diffusion distillation for efficient image generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 205–213, 2024b.

Yingchao Feng, Xian Sun, Wenhui Diao, Jihao Li, and Xin Gao. Double similarity distillation for semantic image segmentation. *IEEE Transactions on Image Processing*, 30:5363–5376, 2021.

Yinqiu Feng, Bo Zhang, Lingxi Xiao, Yutian Yang, Tana Gegen, and Zexi Chen. Enhancing medical imaging with gans synthesizing realistic images from limited data. In *IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI)*, pp. 1192–1197, 2024c.

Virginia Fernandez, Pedro Sanchez, Walter Hugo Lopez Pinaya, Grzegorz Jacenków, Sotirios A Tsaftaris, and M Jorge Cardoso. Privacy distillation: reducing re-identification risk of diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 3–13. Springer, 2023.

Tim Fischer, Chris Biemann, et al. Large language models are overparameterized text encoders. *arXiv preprint arXiv:2410.14578*, 2024.

Jack FitzGerald, Shankar Ananthakrishnan, Konstantine Arkoudas, Davide Bernardi, Abhishek Bhagia, Claudio Delli Bovi, Jin Cao, Rakesh Chada, Amit Chauhan, Luoxin Chen, et al. Alexa teacher model: Pretraining and distilling multi-billion-parameter encoders for natural language understanding systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2893–2902, 2022.

James Flemings and Murali Annavaram. Differentially private knowledge distillation via synthetic text generation. *arXiv preprint arXiv:2403.00932*, 2024.

Kexue Fu, Peng Gao, Renrui Zhang, Hongsheng Li, Yu Qiao, and Manning Wang. Distillation with contrast is all you need for self-supervised point cloud representation learning. *arXiv preprint arXiv:2202.04241*, 2022.

Kui Fu, Peipei Shi, Yafei Song, Shiming Ge, Xiangju Lu, and Jia Li. Ultrafast video attention prediction with coupled knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10802–10809, 2020.

Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, pp. 3697–3701, 2017.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International conference on machine learning*, pp. 1607–1616. PMLR, 2018.

Marco Gaido, Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. End-to-end speech-translation with knowledge distillation: Fbk@ iwslt2020. *arXiv preprint arXiv:2006.02965*, 2020.

Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41 (4):1–13, 2022.

Alexander Gambashidze, Aleksandr Dadukin, Maxim Golyadkin, Maria Razzhivina, and Ilya Makarov. Weak-to-strong 3d object detection with x-ray distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15055–15064, 2024.

Chaochen Gao, Xing Wu, Peng Wang, Jue Wang, Liangjun Zang, Zhongyuan Wang, and Songlin Hu. Distilcse: Effective knowledge distillation for contrastive sentence embeddings. *arXiv preprint arXiv:2112.05638*, 2021.

Huan Gao, Jichang Guo, Guoli Wang, and Qian Zhang. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9903–9913. IEEE, 2022a.

Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *European Conference on Computer Vision*, pp. 266–282. Springer, 2022b.

Tianyi Gao, Wei Ao, Xing-Ao Wang, Yuanhao Zhao, Ping Ma, Mengjie Xie, Hang Fu, Jinchang Ren, and Zhi Gao. Enrich distill and fuse: Generalized few-shot semantic segmentation in remote sensing leveraging foundation model's assistance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2771–2780, 2024.

Yuting Gao, Jia-Xin Zhuang, Shaohui Lin, Hao Cheng, Xing Sun, Ke Li, and Chunhua Shen. Disco: Remedying self-supervised learning on lightweight models with distilled contrastive learning. In *European Conference on Computer Vision*, pp. 237–253. Springer, 2022c.

Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 103–118, 2018.

Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Learning with privileged information via adversarial discriminative modality distillation. *IEEE transactions on pattern analysis and machine intelligence*, 42 (10):2581–2593, 2019.

Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Ling Ge, Chunming Hu, Guanghui Ma, Jihong Liu, and Hong Zhang. Discrepancy and uncertainty aware denoising knowledge distillation for zero-shot cross-lingual named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18056–18064, 2024.

Zhengyang Geng, Ashwini Pokle, William Luo, Justin Lin, and J Zico Kolter. Consistency models made easy. *arXiv preprint arXiv:2406.14548*, 2024.

Kyle Genova, Xiaoqi Yin, Abhijit Kundu, Caroline Pantofaru, Forrester Cole, Avneesh Sud, Brian Brewington, Brian Shucker, and Thomas Funkhouser. Learning 3d semantic segmentation with only 2d image supervision. In *2021 International Conference on 3D Vision (3DV)*, pp. 361–372. IEEE, 2021.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pp. 540–557. Springer, 2022.

Mahdi Ghorbani, Fahimeh Fooladgar, and Shohreh Kasaei. Be your own best competitor! multi-branched adversarial knowledge transfer. *arXiv preprint arXiv:2010.04516*, 2020.

Mahsa Ghorbani, Mojtaba Bahrami, Anees Kazi, Mahdieh Soleymani Baghshah, Hamid R Rabiee, and Nassir Navab. Gkd: Semi-supervised graph knowledge distillation for graph-independent inference. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pp. 709–718. Springer, 2021.

Shantanu Ghosh, Ke Yu, and Kayhan Batmanghelich. Distilling blackbox to interpretable models for efficient transfer learning. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 14221 of *Lecture Notes in Computer Science*, pp. 628–638. Springer, 2023.

Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 3996–4003, 2020.

Guoqiang Gong, Jiaxing Wang, Jin Xu, Deping Xiang, Zicheng Zhang, Leqi Shen, Yifeng Zhang, JunhuaShu JunhuaShu, ZhaolongXing ZhaolongXing, Zhen Chen, et al. Beyond logits: Aligning feature dynamics for effective knowledge distillation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23067–23077, 2025.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Jianping Gou, Liyuan Sun, Baosheng Yu, Shaohua Wan, and Dacheng Tao. Hierarchical multi-attention transfer for knowledge distillation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(2):1–20, 2023.

Jianping Gou, Xiabin Zhou, Lan Du, Yibing Zhan, Wu Chen, and Zhang Yi. Difference-aware distillation for semantic segmentation. *IEEE Transactions on Multimedia*, 2024.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Jindong Gu, Wei Liu, and Yonglong Tian. Simple distillation baselines for improving small self-supervised models. *arXiv preprint arXiv:2106.11304*, 2021a.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021b.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Qi GUAN, Zihao SHENG, and Shibei XUE. Hrpose: Real-time high-resolution 6d pose estimation network using knowledge distillation. *Chinese Journal of Electronics*, 32(1):189–198, 2023.

Jia Guo, Minghao Chen, Yao Hu, Chen Zhu, Xiaofei He, and Deng Cai. Reducing the teacher-student gap via spherical knowledge disitllation. *arXiv preprint arXiv:2010.07485*, 2020a.

Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2154–2164. Computer Vision Foundation / IEEE, 2021a.

Jingwen Guo, Hong Liu, Shitong Sun, Tianyu Guo, Min Zhang, and Chenyang Si. Fsar: Federated skeleton-based action recognition with adaptive topology structure and knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10366–10376. IEEE, 2023a.

Jiongyu Guo, Defang Chen, and Can Wang. Alignahead: online cross-layer knowledge extraction on graph neural networks. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2022.

Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11020–11029, 2020b.

Shuxuan Guo, José M. Álvarez, and Mathieu Salzmann. Distilling image classifiers in object detectors. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1036–1047, 2021b.

Shuxuan Guo, Yinlin Hu, José M. Álvarez, and Mathieu Salzmann. Knowledge distillation for 6d pose estimation by aligning distributions of local predictions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18633–18642. IEEE, 2023b.

Yufeng Guo, Weiwei Zhang, Junhuang Wang, Ming Ji, Chenghui Zhen, and Zhengzheng Guo. Afmpm: adaptive feature map pruning method based on feature distillation. *International Journal of Machine Learning and Cybernetics*, 15(2):573–588, 2024a.

Zhen Guo, Pengzhou Zhang, and Peng Liang. Sakd: Sparse attention knowledge distillation. *Image and Vision Computing*, 146:105020, 2024b.

Zhichun Guo, Chunhui Zhang, Yujie Fan, Yijun Tian, Chuxu Zhang, and Nitesh V Chawla. Boosting graph neural networks via adaptive knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7793–7801, 2023c.

Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin. Class attention transfer based knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11868–11877, 2023d.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2827–2836, 2016.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Md Akmal Haidar and Mehdi Rezagholizadeh. Textkd-gan: Text generation using knowledge distillation and generative adversarial networks. In *Advances in Artificial Intelligence: 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019, Kingston, ON, Canada, May 28–31, 2019, Proceedings 32*, pp. 107–118. Springer, 2019.

Donghoon Han, Seunghyeon Seo, Eunhwan Park, Seong-Uk Nam, and Nojun Kwak. Unleash the potential of clip for video highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8275–8279, 2024a.

Tingxu Han, Shenghan Huang, Ziqi Ding, Weisong Sun, Yebo Feng, Chunrong Fang, Jun Li, Hanwei Qian, Cong Wu, Quanjun Zhang, et al. On the effectiveness of distillation in mitigating backdoors in pre-trained encoder. *arXiv preprint arXiv:2403.03846*, 2024b.

Zhiwei Hao, Jianyuan Guo, Kai Han, Han Hu, Chang Xu, and Yunhe Wang. Revisit the power of vanilla knowledge distillation: from small scale to large scale. *Advances in Neural Information Processing Systems*, 36:10170–10183, 2023.

Huarui He, Jie Wang, Zhanqiu Zhang, and Feng Wu. Compressing deep graph neural networks via adversarial knowledge distillation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 534–544, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 578–587, 2019.

Moein Heidari, Amirhossein Kazerouni, Milad Soltany, Reza Azad, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, and Dorit Merhof. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 6202–6212, 2023.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1921–1930, 2019a.

Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3779–3787, 2019b.

Congrui Hetang, Haoru Xue, Cindy Le, Tianwei Yue, Wenping Wang, and Yihui He. Segment anything model for road network graph extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2556–2566, 2024.

Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*, 2022.

Kiet Anh Hoang, Khanh Duong, Triet Nguyen Van Minh, Tung Le, and Huy Tien Nguyen. Integrating voice activity detection to enhance robustness of on-device speaker verification. In *Pacific Rim International Conference on Artificial Intelligence*, pp. 369–380. Springer, 2024.

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022a.

James Hong, Matthew Fisher, Michaël Gharbi, and Kayvon Fatahalian. Video pose distillation for few-shot, fine-grained sports action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9234–9243. IEEE, 2021.

Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *European Conference on Computer Vision*, pp. 87–104. Springer, 2022b.

Md Imtiaz Hossain, Sharmen Akhter, Choong Seon Hong, and Eui-Nam Huh. Single teacher, multiple perspectives: Teacher knowledge augmentation for enhanced knowledge distillation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xiaojun Hou, Jiazheng Xing, Yijie Qian, Yaowei Guo, Shuo Xin, Junhao Chen, Kai Tang, Mengmeng Wang, Zhengkai Jiang, Liang Liu, and Yong Liu. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26541–26551. IEEE, 2024.

Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8479–8488, 2022.

Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Yi-Ting Hsiao, Siavash Khodadadeh, Kevin Duarte, Wei-An Lin, Hui Qu, Mingi Kwon, and Ratheesh Kalarot. Plug-and-play diffusion distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13743–13752, 2024.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*, 2023.

Yen-Chang Hsu, James Smith, Yilin Shen, Zsolt Kira, and Hongxia Jin. A closer look at knowledge distillation with features, logits, and gradients. *arXiv preprint arXiv:2203.10163*, 2022.

Chengming Hu, Xuan Li, Dan Liu, Haolun Wu, Xi Chen, Ju Wang, and Xue Liu. Teacher-student architecture for knowledge distillation: A survey. *arXiv preprint arXiv:2308.04268*, 2023a.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Haifeng Hu, Yuyang Feng, Dapeng Li, Suofei Zhang, and Haitao Zhao. Monocular depth estimation via self-supervised self-distillation. *Sensors*, 24(13):4090, 2024a.

Haolin Hu, Huanqiang Zeng, Yi Xie, Yifan Shi, Jianqing Zhu, and Jing Chen. Global instance relation distillation for convolutional neural network compression. *Neural Computing and Applications*, 36(18): 10941–10953, 2024b.

Junjie Hu, Chenyou Fan, Hualie Jiang, Xiyue Guo, Yuan Gao, Xiangyong Lu, and Tin Lun Lam. Boosting lightweight depth estimation via knowledge distillation. In Zhi Jin, Yuncheng Jiang, Robert Andrei Buchmann, Yaxin Bi, Ana-Maria Ghiran, and Wenjun Ma (eds.), *Knowledge Science, Engineering and Management*, pp. 27–39. Springer Nature Switzerland, 2023b.

Minghao Hu, Yuxing Peng, Furu Wei, Zhen Huang, Dongsheng Li, Nan Yang, and Ming Zhou. Attention-guided answer distillation for machine reading comprehension. *arXiv preprint arXiv:1808.07644*, 2018.

Bo Huang, Mingyang Chen, Yi Wang, Junda Lu, Minhao Cheng, and Wei Wang. Boosting accuracy and robustness of student models via adaptive adversarial distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24668–24677, 2023a.

Junliang Huang, Haifeng Zheng, and Xinxin Feng. Multi-scale distillation for low scanline resolution depth completion. In *2024 9th International Conference on Computer and Communication Systems (ICCCS)*, pp. 854–859. IEEE, 2024a.

Lina Huang, Yixiong Liang, and Jianfeng Liu. Des-sam: Distillation-enhanced semantic sam for cervical nuclear segmentation with box annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 15009 of *Lecture Notes in Computer Science*, pp. 223–234. Springer, 2024b.

Rui Huang, Xuran Pan, Henry Zheng, Haojun Jiang, Zhifeng Xie, Cheng Wu, Shiji Song, and Gao Huang. Joint representation learning for text and 3d point cloud. *Pattern Recognition*, 147:110086, 2024c.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35:33716–33727, 2022a.

Tao Huang, Yuan Zhang, Shan You, Fei Wang, Chen Qian, Jian Cao, and Chang Xu. Masked distillation with receptive tokens. *arXiv preprint arXiv:2205.14589*, 2022b.

Tao Huang, Yuan Zhang, Mingkai Zheng, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge diffusion for distillation. *Advances in Neural Information Processing Systems*, 36:65299–65316, 2023b.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Dist+: Knowledge distillation from a stronger adaptive teacher. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Tianxin Huang, Jiangning Zhang, Jun Chen, Yuang Liu, and Yong Liu. Resolution-free point cloud sampling network with data distillation. In *European Conference on Computer Vision*, pp. 54–70. Springer, 2022c.

Xiao Huang, Wu Chen, and Wei Zhou. Class-wise adaptive logits distillation with meta-learning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025b.

Xiaohu Huang, Hao Zhou, Kun Yao, and Kai Han. Froster: Frozen clip is a strong teacher for open-vocabulary action recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024d.

Xun Huang, Hai Wu, Xin Li, Xiaoliang Fan, Chenglu Wen, and Cheng Wang. Sunshine to rainstorm: Cross-weather knowledge distillation for robust 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2409–2416, 2024e.

Yuanqing Huang, Yinggui Wang, Le Yang, and Lei Wang. Enhanced face recognition using intra-class incoherence constraint. In *ICLR*, 2024f.

Yuge Huang, Jiaxiang Wu, Xingkun Xu, and Shouhong Ding. Evaluation-oriented knowledge distillation for deep face recognition. In *CVPR*, pp. 18719–18728. IEEE, 2022d.

Yukun Huang, Yanda Chen, Zhou Yu, and Kathleen McKeown. In-context learning distillation: Transferring few-shot learning ability of pre-trained language models. *arXiv preprint arXiv:2212.10670*, 2022e.

Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

Zeyi Huang, Andy Zhou, Zijian Ling, Mu Cai, Haohan Wang, and Yong Jae Lee. A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11685–11695, 2023c.

Cuiying Huo, Di Jin, Yawen Li, Dongxiao He, Yu-Bin Yang, and Lingfei Wu. T2-gnn: Graph neural networks for graphs with incomplete features and structure via teacher-student distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4339–4346, 2023.

Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C2kd: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16006–16015, 2024.

Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7020–7031, 2022.

Injoon Hwang, Haewon Park, Youngwan Lee, Jooyoung Yang, and SunJae Maeng. Pc-lora: Low-rank adaptation for progressive model compression with knowledge distillation. *arXiv preprint arXiv:2406.09117*, 2024.

Sangwon Hwang, Junhyeop Lee, Woo Jin Kim, Sungmin Woo, Kyungjae Lee, and Sangyoun Lee. Lidar depth completion using color-embedded information via knowledge distillation. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):14482–14496, 2021.

Junha Hyung, Sungwon Hwang, Daejin Kim, Hyunji Lee, and Jaegul Choo. Local 3d editing via 3d distillation of clip knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12674–12684, 2023.

Hirofumi Inaguma, Tatsuya Kawahara, and Shinji Watanabe. Source and target bidirectional knowledge distillation for end-to-end speech translation. *arXiv preprint arXiv:2104.06457*, 2021.

Adrian Iordache, Bogdan Alexe, and Radu Tudor Ionescu. Multi-level feature distillation of joint teachers trained on distinct image datasets. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7133–7142. IEEE, 2025.

Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*, 2020.

Geethu Miriam Jacob, Vishal Agarwal, and Björn Stenger. Online knowledge distillation for multi-task learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2359–2368, 2023.

Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 867–876, 2022.

Soojin Jang, Jungmin Yun, Junehyoung Kwon, Eunju Lee, and Youngbin Kim. Dial: Dense image-text alignment for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, pp. 248–266. Springer, 2024.

Sujin Jang, Dae Ung Jo, Sung Ju Hwang, Dongwook Lee, and Daehyun Ji. Stxd: structural and temporal cross-modal distillation for multi-view 3d object detection. *Advances in Neural Information Processing Systems*, 36:29323–29342, 2023.

Younho Jang, Wheemyung Shin, Jinbeom Kim, Simon S. Woo, and Sung-Ho Bae. Glamd: Global and local attention mask distillation for object detectors. In *European Conference on Computer Vision (ECCV)*, volume 13670 of *Lecture Notes in Computer Science*, pp. 460–476. Springer, 2022.

Deyi Ji, Haoran Wang, Mingyuan Tao, Jianqiang Huang, Xian-Sheng Hua, and Hongtao Lu. Structural and statistical texture knowledge distillation for semantic segmentation. In *CVPR*, pp. 16855–16864. IEEE, 2022.

Zihao Jia, Shengkun Sun, Guangcan Liu, and Bo Liu. Mssd: Multi-scale self-distillation for object detection. *Visual Intelligence*, 2:1–11, 2024.

Feng Jiang, Heng Gao, Shoumeng Qiu, Haiqiang Zhang, Ru Wan, and Jian Pu. Knowledge distillation from 3d to bird's-eye-view for lidar semantic segmentation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 402–407. IEEE, 2023a.

Linjun Jiang, Yinghao Li, Yue Liu, Zhiyuan Dong, Mengyuan Yao, and Yusong Lin. Knowledge distillation-based point cloud registration method. In *International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2023)*, volume 13105, pp. 648–655. SPIE, 2024a.

Peng-Tao Jiang and Yuqi Yang. Segment anything is a good pseudo-label generator for weakly supervised semantic segmentation. *arXiv preprint arXiv:2305.01275*, 2023.

Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy. In *European Conference on Computer Vision*, pp. 38–57. Springer, 2024b.

Yuncheng Jiang, Zixun Zhang, Jun Wei, Chun-Mei Feng, Guanbin Li, Xiang Wan, Shuguang Cui, and Zhen Li. Let video teaches you more: Video-to-image knowledge distillation using detection transformer for medical video lesion detection. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 944–949. IEEE, 2024c.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. Lion: Adversarial distillation of proprietary large language models. *arXiv preprint arXiv:2305.12870*, 2023b.

Yuxuan Jiang, Chen Feng, Fan Zhang, and David Bull. Mtkd: Multi-teacher knowledge distillation for image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 15097 of *Lecture Notes in Computer Science*, pp. 364–382. Springer, 2024d.

Zheng Jiang, Jinqing Zhang, Yanan Zhang, Qingjie Liu, Zhenghui Hu, Baohui Wang, and Yunhong Wang. Fsd-bev: Foreground self-distillation for multi-view 3d object detection. In *European Conference on Computer Vision*, pp. 110–126. Springer, 2024e.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

Heegon Jin, Seonil Son, Jemin Park, Youngseok Kim, Hyungjong Noh, and Yeonsoo Lee. Align-to-distill: Trainable attention alignment for knowledge distillation in neural machine translation. *arXiv preprint arXiv:2403.01479*, 2024.

Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11165–11172. AAAI Press, 2020.

Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24276–24285. IEEE, 2023a.

Yufeng Jin, Guosheng Hu, Haonan Chen, Duoqian Miao, Liang Hu, and Cairong Zhao. Cross-modal distillation for speaker recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12977–12985, 2023b.

Jaewon Jung, Hongsun Jang, Jaeyong Song, and Jinho Lee. Peeraid: Improving adversarial distillation from a specialized peer tutor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24482–24491, 2024.

Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12115–12125, 2021.

Woonghyun Ka, Jae Young Lee, Jaehyun Choi, and Junmo Kim. Stereo-matching knowledge distilled monocular depth estimation filtered by multiple disparity consistency. In *ICASSP*, pp. 3840–3844. IEEE, 2024.

Dahyun Kang, Piotr Koniusz, Minsu Cho, and Naila Murray. Distilling self-supervised vision transformers for weakly-supervised few-shot classification 'i&' segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19627–19638. IEEE, 2023a.

Dahyun Kang, Piotr Koniusz, Minsu Cho, and Naila Murray. Distilling self-supervised vision transformers for weakly-supervised few-shot classification & segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19627–19638, 2023b.

Minchan Kang, Sanghyeok Son, and Daeshik Kim. Adaptive class token knowledge distillation for efficient vision transformer. *Knowledge-Based Systems*, 304:112531, 2024.

Zijian Kang, Peizhen Zhang, Xiangyu Zhang, Jian Sun, and Nanning Zheng. Instance-conditional knowledge distillation for object detection. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 16468–16480, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Sandra Kara, Hejer Ammar, Julien Denize, Florian Chabot, and Quoc-Cuong Pham. Diod: Self-distillation meets object discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3975–3985, 2024.

Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. *arXiv preprint arXiv:2306.09316*, 2023.

Ayoub Karine, Thibault Napoléon, and Maher Jridi. Channel-spatial knowledge distillation for efficient semantic segmentation. *Pattern Recognition Letters*, 180:48–54, 2024.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota, 2019.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19729–19739, 2023.

Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11144–11154, 2023a.

Hojin Kim, Seunghun Lee, Hyeon Kang, and Sunghoon Im. Offline-to-online knowledge distillation for video instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 159–168, 2024a.

Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018.

Janghyun Kim, Jeonghyun Noh, Mingyu Jeong, Wonju Lee, Yeonchool Park, and Jinsun Park. Adnet: Non-local affinity distillation network for lightweight depth completion with guidance from missing lidar points. *IEEE Robotics and Automation Letters*, 2024b.

Jeong-Uk Kim, Seong-Hu Kim, Hye-Young Kim, Hye-Jin Kim, and Hong-Goo Kang. Knowledge distillation-based training of speech enhancement for noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:149–160, 2023b. doi: 10.1109/TASLP.2023.3241700. URL `https://ieeexplore.ieee.org/document/10535505/`.

Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation with progressive refinement of targets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6567–6576, 2021a.

Seongbin Kim, Gyuwan Kim, Seongjin Shin, and Sangmin Lee. Two-stage textual knowledge distillation for end-to-end spoken language understanding. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7463–7467. IEEE, 2021b.

Seonghak Kim, Gyeongdo Ham, Suin Lee, Donggon Jang, and Daeshik Kim. Maximizing discrimination capability of knowledge distillation with energy function. *Knowledge-Based Systems*, 296:111911, 2024c.

Seung Wook Kim and Hyo-Eun Kim. Transferring knowledge to smaller network with class-distance loss. 2017.

Sunwoo Kim and Minje Kim. Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 176–180. IEEE, 2021.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*, 2016.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Marvin Klingner, Shubhankar Borse, Varun Ravi Kumar, Behnaz Rezaei, Venkatraman Narayanan, Senthil Yogamani, and Fatih Porikli. X3kd: Knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13343–13353, 2023.

Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 35:23311–23330, 2022.

Fei Kong, Jinhao Duan, Lichao Sun, Hao Cheng, Renjing Xu, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. Act-diffusion: Efficient adversarial consistency training for one-step diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8890–8899, 2024.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022.

Gakuto Kurata and George Saon. Knowledge distillation from offline to streaming rnn transducer for end-to-end speech recognition. In *Interspeech*, pp. 2117–2121, 2020.

Hyeokjun Kweon and Kuk-Jin Yoon. From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19499–19509, 2024.

Qizhen Lan and Qing Tian. Gradient-guided knowledge distillation for object detectors. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 423–432. IEEE, 2024.

Xu Lan, Xiatian Zhu, and Shaogang Gong. Self-referenced deep learning. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14*, pp. 284–300. Springer, 2019.

Shanshan Lao, Guanglu Song, Boxiao Liu, Yu Liu, and Yujiu Yang. Unikd: Universal knowledge distillation for mimicking homogeneous or heterogeneous object detectors. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6339–6349. IEEE, 2023.

Bokyeung Lee, Kyungdeuk Ko, Jonghwan Hong, and Hanseok Ko. Prune channel and distill: Discriminative knowledge distillation for semantic segmentation. In *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 339–345. IEEE, 2024.

Hongsin Lee, Seungju Cho, and Changick Kim. Indirect gradient matching for adversarial robust distillation. *arXiv preprint arXiv:2312.03286*, 2023a.

Pilhyeon Lee, Taeoh Kim, Minho Shim, Dongyoon Wee, and Hyeran Byun. Decomposed cross-modal distillation for rgb-based temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2373–2383, 2023b.

Seunghyun Lee and Byung Cheol Song. Graph-based knowledge distillation by multi-head attention network. *arXiv preprint arXiv:1907.02226*, 2019.

Seunghyun Lee and Byung Cheol Song. Interpretable embedding procedure knowledge transfer via stacked principal component analysis and graph neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8297–8305, 2021.

YuXing Lee and Wei Wu. Feature adversarial distillation for point cloud classification. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 970–974. IEEE, 2023.

Jan Lehečka, Jan Švec, Pavel Ircing, and Luboš Šmídl. Bert-based sentiment analysis using distillation. In *International conference on statistical language and speech processing*, pp. 58–70. Springer, 2020.

Yikun Lei, Zhengshan Xue, Xiaohu Zhao, Haoran Sun, Shaolin Zhu, Xiaodong Lin, and Deyi Xiong. Ckdst: Comprehensively and effectively distill knowledge from machine translation to end-to-end speech translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3123–3137, 2023.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022a.

Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1482–1491. IEEE, 2021.

Chuan Li, Xiao Teng, Yan Ding, and Long Lan. Ntce-kd: Non-target-class-enhanced knowledge distillation. *Sensors*, 24(11):3617, 2024a.

Cong Li, Gong Cheng, and Junwei Han. Boosting knowledge distillation via intra-class logit distribution smoothing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023a.

Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1306–1313. AAAI Press, 2022b.

Jianing Li, Ming Lu, Jiaming Liu, Yandong Guo, Li Du, and Shanghang Zhang. Bev-lgkd: A unified lidar-guided knowledge distillation framework for bev 3d object detection. *arXiv preprint arXiv:2212.00623*, 2022c.

Jingzhi Li, Zidong Guo, Hui Li, Seungju Han, Ji-Won Baek, Min Yang, Ran Yang, and Sungjoo Suh. Rethinking feature-based knowledge distillation for face recognition. In *CVPR*, pp. 20156–20165. IEEE, 2023b.

Jinpeng Li, Guangyong Chen, Hangyu Mao, Danruo Deng, Dong Li, Jianye Hao, Qi Dou, and Pheng-Ann Heng. Flat-aware cross-stage distilled framework for imbalanced medical image classification. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 13433 of *Lecture Notes in Computer Science*, pp. 217–226. Springer, 2022d.

Lebin Li, Ning Jiang, Jialiang Tang, and Xinlei Huang. Amalgamating knowledge for comprehensive classification with uncertainty suppression. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5. IEEE, 2024b.

Liangqi Li, Jiaxu Miao, Dahu Shi, Wenming Tan, Ye Ren, Yi Yang, and Shiliang Pu. Distilling detr with visual-linguistic knowledge for open-vocabulary object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6478–6487. IEEE, 2023c.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. Symbolic chain-of-thought distillation: Small models can also" think" step-by-step. *arXiv preprint arXiv:2306.14050*, 2023d.

Maohui Li, Michael Halstead, and Chris Mccool. Knowledge distillation for efficient instance semantic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5432–5439, 2024c.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 16189–16211, 2024d.

Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, and Song Han. Gan compression: Efficient architectures for interactive conditional gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5284–5294, 2020a.

Peng Li, Chang Shu, Yuan Xie, Yan Qu, and Hui Kong. Hierarchical knowledge squeezed adversarial network compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11370–11377, 2020b.

Qing Li, Xiaojiang Peng, Chuan Yan, Pan Gao, and Qi Hao. Self-ensembling for 3d point cloud domain adaptation. *Image and Vision Computing*, 154:105409, 2025a.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*, 2022e.

Te Li, Huajun Wang, Guangzhi Li, Songshan Liu, and Li Tang. Swinf: Swin transformer with feature fusion in target detection. In *Journal of Physics: Conference Series*, volume 2284, pp. 012027. IOP Publishing, 2022f.

Tong Li, Long Liu, Kang Liu, Xin Wang, Bo Zhou, Hongguang Yang, and Kai Lu. Adaptive dual guidance knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18457–18465, 2025b.

Wei-Hong Li and Hakan Bilen. Knowledge distillation for multi-task learning. In *European Conference on Computer Vision*, pp. 163–176. Springer, 2020.

Wei-Hong Li, Xialei Liu, and Hakan Bilen. Universal representations: A unified look at multiple task and domain learning. *International Journal of Computer Vision*, 132(5):1521–1545, 2024e.

Xiaojie Li, Shaowei He, Jianlong Wu, Yue Yu, Liqiang Nie, and Min Zhang. Mask again: Masked knowledge distillation for masked video modeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 2221–2232, 2023e.

Xuanlin Li, Yunhao Fang, Minghua Liu, Zhan Ling, Zhuowen Tu, and Hao Su. Distilling large vision-language model with out-of-distribution generalizability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2492–2503, 2023f.

Yanjing Li, Sheng Xu, Mingbao Lin, Jihao Yin, Baochang Zhang, and Xianbin Cao. Representation disparity-aware distillation for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6715–6724, 2023g.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *ICLR*, 2021a.

Yingahao Aaron Li, Rithesh Kumar, and Zeyu Jin. Dmdspeech: Distilled diffusion model surpassing the teacher in zero-shot speech synthesis via direct metric optimization. *arXiv preprint arXiv:2410.11097*, 2024f.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Bin Sun, Xinglin Wang, Heda Wang, and Kan Li. Turning dust into gold: Distilling complex reasoning capabilities from llms by leveraging negative data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18591–18599, 2024g.

Yongqi Li and Wenjie Li. Data distillation for text classification. *arXiv preprint arXiv:2104.08448*, 2021.

Zheng Li, Jingwen Ye, Mingli Song, Ying Huang, and Zhigeng Pan. Online knowledge distillation for efficient pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11740–11750, 2021b.

Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26617–26626, 2024h.

Zheng Li, Xiang Li, Lingfeng Yang, Renjie Song, Jian Yang, and Zhigeng Pan. Dual teachers for self-knowledge distillation. *Pattern Recognition*, 151:110422, 2024i.

Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Less is more: Task-aware layer-wise distillation for language model compression. In *International Conference on Machine Learning*, pp. 20852–20867. PMLR, 2023a.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7070, 2023b.

Peng Liang, Weiwei Zhang, Junhuang Wang, and Yufeng Guo. Neighbor self-knowledge distillation. *Information Sciences*, 654:119859, 2024.

Shining Liang, Ming Gong, Jian Pei, Linjun Shou, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. Reinforced iterative knowledge distillation for cross-lingual named entity recognition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3231–3239, 2021.

Xiaobo Liang, Lijun Wu, Juntao Li, Tao Qin, Min Zhang, and Tie-Yan Liu. Multi-teacher distillation with single model for neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:992–1002, 2022a.

Yunhao Liang, Yanhua Long, Yijie Li, Jiaen Liang, and Yuping Wang. Joint framework with deep feature distillation and adaptive focal loss for weakly supervised audio tagging and acoustic event detection. *Digital Signal Processing*, 123:103446, 2022b.

Dongping Liao, Xitong Gao, and Chengzhong Xu. Impartial adversarial distillation: Addressing biased data-free knowledge distillation via adaptive constrained optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3342–3350, 2024.

Yen-Lun Liao, Xuanjun Chen, Chung-Che Wang, and Jyh-Shing Roger Jang. Adversarial speaker distillation for countermeasure model on automatic speaker verification. *arXiv preprint arXiv:2203.17031*, 2022.

Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*, 2022a.

Fangzhou Lin, Haotian Liu, Haoying Zhou, Songlin Hou, Kazunori D Yamada, Gregory S Fischer, Yanhua Li, Haichong K Zhang, and Ziming Zhang. Loss distillation via gradient matching for point cloud completion with weighted chamfer distance. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 511–518. IEEE, 2024a.

Han Lin, Guangxing Han, Jiawei Ma, Shiyuan Huang, Xudong Lin, and Shih-Fu Chang. Supervised masked knowledge distillation for few-shot transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19649–19659, 2023a.

Jiaye Lin, Lin Li, Baosheng Yu, Weihua Ou, and Jianping Gou. Self-distillation via intra-class compactness. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 139–151. Springer, 2024b.

Liting Lin, Heng Fan, Zhipeng Zhang, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. *Advances in Neural Information Processing Systems*, 35:16743–16754, 2022b.

Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024c.

Ying-Chen Lin and Vincent S Tseng. Multi-view knowledge distillation transformer for human action recognition. *arXiv preprint arXiv:2303.14358*, 2023.

Yu-e Lin, Shuting Yin, Yifeng Ding, and Xingzhu Liang. Atmkd: adaptive temperature guided multi-teacher knowledge distillation. *Multimedia Systems*, 30(5):292, 2024d.

Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15305–15314, 2023b.

Bei Liu, Haoyu Wang, Zhengyang Chen, Shuai Wang, and Yanmin Qian. Self-knowledge distillation via feature enhancement for speaker verification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7542–7546. IEEE, 2022a.

Bochao Liu, Pengju Wang, and Shiming Ge. Learning differentially private diffusion models via stochastic adversarial distillation. In *European Conference on Computer Vision*, pp. 55–71. Springer, 2024a.

Boxiao Liu, Shenghan Zhang, Guanglu Song, Haihang You, and Yu Liu. Rectifying the data bias in knowledge distillation. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1477–1486, 2021a.

Cencen Liu, Dongyang Zhang, and Ke Qin. Knowledge distillation for single image super-resolution via contrastive learning. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 1079–1083, 2024b.

Feng Liu, Minchul Kim, Zhiyuan Ren, and Xiaoming Liu. Distilling clip with dual guidance for learning discriminative human body shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 256–266, 2024c.

Jian Liu, Yubo Chen, and Kang Liu. Exploiting the ground-truth: An adversarial imitation based knowledge distillation approach for event detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6754–6761, 2019a.

Jian Liu, Yufeng Chen, and Jinan Xu. Machine reading comprehension as data augmentation: A case study on implicit event argument extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2716–2725, 2021b.

Li Liu, Qingle Huang, Sihao Lin, Hongwei Xie, Bing Wang, Xiaojun Chang, and Xiaodan Liang. Exploring inter-channel correlation for diversity-preserved knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8271–8280, 2021c.

Liyang Liu, Zihan Wang, Minh Hieu Phan, Bowen Zhang, Jinchao Ge, and Yifan Liu. Bpkd: Boundary privileged knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1062–1072, 2024d.

Miao Liu, Xin Chen, Yun Zhang, Yin Li, and James M Rehg. Attention distillation for learning video representations. *arXiv preprint arXiv:1904.03249*, 2019b.

Mushui Liu, Yunlong Yu, Zhong Ji, Jungong Han, and Zhongfei Zhang. Tolerant self-distillation for image classification. *Neural Networks*, 174:106215, 2024e.

Peiye Liu, Wu Liu, Huadong Ma, Zhewei Jiang, and Mingoo Seok. Ktan: knowledge transfer adversarial network. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2020a.

Ruishan Liu, Nicolo Fusi, and Lester Mackey. Teacher-student compression with generative adversarial networks. *arXiv preprint arXiv:1812.02271*, 2018.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024f.

Tao Liu, Chenshu Chen, Xi Yang, and Wenming Tan. Rethinking knowledge distillation with raw features for semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1155–1164, 2024g.

Tian Yu Liu, Parth Agrawal, Allison Chen, Byung-Woo Hong, and Alex Wong. Monitored distillation for positive congruent depth completion. In *European Conference on Computer Vision*, pp. 35–53. Springer, 2022b.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482*, 2019c.

Xiaolong Liu, Lujun Li, Chao Li, and Anbang Yao. Norm: Knowledge distillation via n-to-one representation matching. *arXiv preprint arXiv:2305.13803*, 2023a.

Xiaoyu Liu, Bo Hu, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. Efficient biomedical instance segmentation via knowledge distillation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 13434 of *Lecture Notes in Computer Science*, pp. 14–24. Springer, 2022c.

Yang Liu, Sheng Shen, and Mirella Lapata. Noisy self-knowledge distillation for text summarization. *arXiv preprint arXiv:2009.07032*, 2020b.

Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2604–2613, 2019d.

Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. *Advances in Neural Information Processing Systems*, 36, 2024h.

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. End-to-end speech translation with knowledge distillation. *arXiv preprint arXiv:1904.08075*, 2019e.

Yueh-Cheng Liu, Yu-Kai Huang, Hung-Yueh Chiang, Hung-Ting Su, Zhe-Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H Hsu. Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining. *arXiv preprint arXiv:2104.04687*, 2021d.

Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7096–7104, 2019f.

Yuxuan Liu. Learning to reason with autoregressive in-context distillation. In *The Second Tiny Papers Track at ICLR 2024*, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021e.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. Llm-qat: Data-free quantization aware training for large language models. *arXiv preprint arXiv:2305.17888*, 2023b.

Zhengyi Liu, Yacheng Tan, Qian He, and Yun Xiao. Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4486–4497, 2021f.

Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. 3d-to-2d distillation for indoor scene parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4464–4474, 2021g.

Zhuoming Liu, Xuefeng Hu, and Ram Nevatia. Efficient feature distillation for zero-shot annotation object detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 882–891. IEEE, 2024i.

Ziwei Liu, Yongtao Wang, Xiaojie Chu, Nan Dong, Shengxiang Qi, and Haibin Ling. A simple and generic framework for feature distillation via channel-wise transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1129–1138, 2023c.

Yanxin Long, Jianhua Han, Runhui Huang, Hang Xu, Yi Zhu, Chunjing Xu, and Xiaodan Liang. Fine-grained visual–text prompt-driven self-training for open-vocabulary object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Christian Löwens, Thorben Funke, André Wagner, and Alexandru Paul Condurache. Unsupervised point cloud registration with self-distillation. *arXiv preprint arXiv:2409.07558*, 2024.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. *arXiv preprint arXiv:2410.11081*, 2024.

Guoming Lu, Heng Yin, Zhiyong Shu, Jielei Wang, and Guangchun Luo. Bdckd: Unlocking the power of brownian distance covariance in knowledge distillation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Jingze Lu, Yuxiang Zhang, Wenchao Wang, Zengqiang Shang, and Pengyuan Zhang. One-class knowledge distillation for spoofing speech detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11251–11255. IEEE, 2024.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7086–7096, 2022.

Sihui Luo, Xinchao Wang, Gongfan Fang, Yao Hu, Dapeng Tao, and Mingli Song. Knowledge amalgamation from heterogeneous networks by common feature learning. *arXiv preprint arXiv:1906.10546*, 2019.

Sihui Luo, Wenwen Pan, Xinchao Wang, Dazhou Wang, Haihong Tang, and Mingli Song. Collaboration by competition: Self-coordinated knowledge amalgamation for multi-talent student learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pp. 631–646. Springer, 2020.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023a.

Weijian Luo. A comprehensive survey on knowledge distillation of diffusion models. *arXiv preprint arXiv:2304.04262*, 2023.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36:76525–76546, 2023b.

Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 166–183, 2018.

Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós, and Juan C. San Miguel. Attention-based knowledge distillation in scene recognition: The impact of a dct-driven loss. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4769–4783, 2023.

Dongtong Ma, Kaibing Zhang, Qizhi Cao, Jie Li, and Xinbo Gao. Coordinate attention guided dual-teacher adaptive knowledge distillation for image classification. *Expert Systems with Applications*, 250:123892, 2024.

Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14074–14083, 2022.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*, 2022.

George Malki. Efficient sentiment analysis and topic modeling in nlp using knowledge distillation and transfer learning, 2023.

Amir M Mansourian, Arya Jalali, Rozhan Ahmadi, and Shohreh Kasaei. Attention-guided feature distillation for semantic segmentation. *arXiv preprint arXiv:2403.05451*, 2024.

Amir M Mansourian, Rozhan Ahamdi, and Shohreh Kasaei. Aicsd: Adaptive inter-class similarity distillation for semantic segmentation. *Multimedia Tools and Applications*, pp. 1–20, 2025.

Tianjun Mao and Chenghong Zhang. Diffslu: Knowledge distillation based diffusion model for cross-lingual spoken language understanding. In *Proc. of Interspeech*, 2023.

Konstantin Markov and Tomoko Matsui. Robust speech recognition using generalized distillation framework. In *Interspeech*, pp. 2364–2368, 2016.

Rémi Marsal, Florian Chabot, Angelique Loesch, William Grolleau, and Hichem Sahbi. Monoprob: Self-supervised monocular depth estimation with interpretable uncertainty. In *WACV*, pp. 3625–3634. IEEE, 2024.

Kidist Amde Mekonnen, Nicola Dall'Asen, and Paolo Rota. Adv-kd: Adversarial knowledge distillation for faster diffusion sampling. *arXiv preprint arXiv:2405.20675*, 2024.

Paul Micaelli and Amos J Storkey. Zero-shot knowledge transfer via adversarial belief matching. *Advances in Neural Information Processing Systems*, 32, 2019.

Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13492–13502, 2022.

Roy Miles and Krystian Mikolajczyk. Understanding the role of the projector in knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4233–4241, 2024.

Roy Miles, Mehmet Kerim Yucel, Bruno Manganelli, and Albert Saa-Garriga. Mobilevos: Real-time video object segmentation contrastive learning meets knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10480–10490, 2023.

Soma Minami, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Knowledge transfer graph for deep collaborative learning. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pp. 728–755. Springer, 2022.

Victoria Mingote, Antonio Miguel, Dayana Ribas, Alfonso Ortega, and Eduardo Lleida. Knowledge distillation and random erasing data augmentation for text-dependent speaker verification. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6824–6828. IEEE, 2020.

Victoria Mingote, Antonio Miguel, Alfonso Ortega, and Eduardo Lleida. Class token and knowledge distillation for multi-head self-attention speaker verification systems. *Digital Signal Processing*, 133:103859, 2023.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6707–6717, 2020.

Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems*, 33:3351–3361, 2020.

Alessio Monti, Angelo Porrello, Simone Calderara, Pasquale Coscia, Lamberto Ballan, and Rita Cucchiara. How many observations are enough? knowledge distillation for trajectory forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6553–6562, 2022.

Amir Moslemi, Anna Briskina, Zubeka Dang, and Jason Li. A survey on knowledge distillation: Recent advancements. *Machine Learning with Applications*, pp. 100605, 2024.

Subhabrata Mukherjee and Ahmed Awadallah. Xtremedistil: Multi-stage distillation for massive multilingual models. *arXiv preprint arXiv:2004.05686*, 2020.

Subhabrata Mukherjee, Ahmed Hassan Awadallah, and Jianfeng Gao. Xtremedistiltransformers: Task transfer for task-agnostic distillation. *arXiv preprint arXiv:2106.04563*, 2021.

Ravi Teja Mullapudi, Steven Chen, Keyi Zhang, Deva Ramanan, and Kayvon Fatahalian. Online model distillation for efficient video inference. In *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 3573–3582, 2019.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Raden Mu'az Mun'im, Nakamasa Inoue, and Koichi Shinoda. Sequence-level knowledge distillation for model compression of attention-based sequence-to-sequence speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6151–6155. IEEE, 2019.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation. *Advances in Neural Information Processing Systems*, 37:41076–41102, 2024.

Balamurali Murugesan, Rukhshanda Hussain, Rajarshi Bhattacharya, Ismail Ben Ayed, and Jose Dolz. Prompting classes: exploring the power of prompt class learning in weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 291–302, 2024.

Lakshmi Nair. Clip-embed-kd: Computationally efficient knowledge distillation using embeddings as teachers. *arXiv preprint arXiv:2404.06170*, 2024.

Sahar Almahfouz Nasser, Nihar Gupte, and Amit Sethi. Reverse knowledge distillation: Training a large model using a small one for retinal image matching on limited data. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7763–7772. IEEE, 2024.

KL Navaneet, Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Simreg: Regression as a simple yet effective tool for self-supervised knowledge distillation. *arXiv preprint arXiv:2201.05131*, 2022.

Pedro C. Neto, Ivona Colakovic, Sašo Karakatič, and Ana F. Sequeira. Synthetic gap mitigation using knowledge distillation in fair face recognition. In *Proceedings of the ECCV 2024 Workshop on Synthetic Data for Computer Vision*, 2024.

Chuong H. Nguyen, Thuy C. Nguyen, Tuan N. Tang, and Nam L. H. Phan. Improving object detection by label assignment distillation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1322–1331. IEEE, 2022.

Thong Thanh Nguyen and Anh Tuan Luu. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11103–11111, 2022.

Thanh Nguyen-Duc, Trung Le, He Zhao, Jianfei Cai, and Dinh Phung. Adversarial local distribution regularization for knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4681–4690, 2023.

Zhenliang Ni, Fukui Yang, Shengzhao Wen, and Gang Zhang. Dual relation knowledge distillation for object detection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1276–1284, 2023.

Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. Dynamic kernel distillation for efficient pose estimation in videos. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6941–6949. IEEE, 2019.

Mehdi Noroozi, Isma Hadji, Brais Martínez, Adrian Bulat, and Georgios Tzimiropoulos. You only need one step: Fast super-resolution with stable diffusion via scale distillation. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *ECCV (29)*, volume 15087 of *Lecture Notes in Computer Science*, pp. 145–161. Springer, 2024.

Utkarsh Ojha, Yuheng Li, Anirudh Sundara Rajan, Yingyu Liang, and Yong Jae Lee. What knowledge gets distilled in knowledge distillation? *Advances in Neural Information Processing Systems*, 36:11037–11048, 2023.

Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pp. 3918–3926. PMLR, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Hatef Otroshi-Shahreza, Anjith George, and Sébastien Marcel. Synthdistill: Face recognition with knowledge distillation from synthetic data. In *IJCB*, pp. 1–10. IEEE, 2023.

Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10870–10879, 2020.

Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016a.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pp. 582–597. IEEE, 2016b.

Hanhoon Park. Semantic super-resolution via self-distillation and adversarial learning. *IEEE Access*, 12: 2361–2370, 2024.

Hyejin Park and Dongbo Min. Dynamic guidance adversarial distillation with enhanced teacher knowledge. In *European Conference on Computer Vision*, pp. 204–219. Springer, 2024.

Hyun Joon Park, Wooseok Shin, Jin Sob Kim, and Sung Won Han. Leveraging non-causal knowledge via cross-network knowledge distillation for real-time speech enhancement. *IEEE Signal Processing Letters*, 2024.

Jeong-Sik Park, Jang-Hwan Cho, Han-Gyu Kim, and Jae-Ho Kim. End-to-end emotional speech recognition using acoustic model adaptation based on knowledge distillation. *Multimedia Tools and Applications*, 82: 14681–14697, 2023. doi: 10.1007/s11042-023-14680-y. URL `https://link.springer.com/article/10.1007/s11042-023-14680-y`.

Sangyong Park and Yong Seok Heo. Knowledge distillation for semantic segmentation using channel and spatial correlations and adaptive cross entropy. *Sensors*, 20(16):4616, 2020.

SeongUk Park and Nojun Kwak. Feature-level ensemble knowledge distillation for aggregating knowledge from multiple networks. In *ECAI 2020*, pp. 1411–1418. IOS Press, 2020.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3967–3976, 2019.

Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Heterogeneous knowledge distillation using information flow modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2339–2348, 2020.

Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, and Qun Liu. Alp-kd: Attention-based layer projection for knowledge distillation. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 35, pp. 13657–13665, 2021.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2085–2094, 2021.

Gaurav Patel, Konda Reddy Mopuri, and Qiang Qiu. Learning to retain while acquiring: Combating distribution-shift in adversarial data-free knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7786–7794, 2023.

Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18983–18992, 2023.

Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5007–5016, 2019.

Rui Peng, Ronggang Wang, Yawen Lai, Luyang Tang, and Yangang Cai. Excavating the potential capacity of self-supervised monocular depth estimation. In *ICCV*, pp. 15540–15549. IEEE, 2021.

Zhiyuan Peng, Xuanji He, Ke Ding, Tan Lee, and Guanglu Wan. Label-free knowledge distillation with contrastive loss for light-weight speaker recognition. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 324–328. IEEE, 2022.

Andres Perez, Valentina Sanguineti, Pietro Morerio, and Vittorio Murino. Audio-visual model distillation using acoustic images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2854–2863, 2020.

Minh-Hieu Phan, The-Anh Ta, Son Lam Phung, Long Tran-Thanh, and Abdesselam Bouzerdoum. Class similarity weighted knowledge distillation for continual semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16845–16854. IEEE, 2022.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Angelo Porrello, Luca Bergamini, and Simone Calderara. Robust re-identification by multiple views knowledge distillation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12355, pp. 93–110. Springer, 2020.

Paul Primus and Gerhard Widmer. A knowledge distillation approach to improving language-based audio retrieval models. Technical report, DCASE2024 Challenge, Tech. Rep, 2024.

Xingqun Qi, Zhuojie Wu, Wenxuan Zou, Min Ren, Yifan Gao, Muyi Sun, Shanghang Zhang, Caifeng Shan, and Zhenan Sun. Exploring generalizable distillation for efficient medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4170–4183, 2024.

Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning*, pp. 28223–28243. PMLR, 2023.

Yupeng Qiang, Xunde Dong, Xiuling Liu, and Yang Yang. Mt-mv-kdf: A novel multi-task multi-view knowledge distillation framework for myocardial infarction detection and localization. *Biomedical Signal Processing and Control*, 95:106382, 2024.

Dian Qin, Jia-Jun Bu, Zhe Liu, Xin Shen, Sheng Zhou, Jing-Jun Gu, Zhi-Hua Wang, Lei Wu, and Hui-Fen Dai. Efficient medical image segmentation based on knowledge distillation. *IEEE Transactions on Medical Imaging*, 40(12):3820–3831, 2021.

Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19446–19455, 2023.

Zhenkai Qin, Shuiping Ni, Mingfu Zhu, Yue Jia, Shangxin Liu, and Yawei Chen. A feature map fusion self-distillation scheme for image classification networks. *Electronics*, 14(1):182, 2025.

Shoumeng Qiu, Feng Jiang, Haiqiang Zhang, Xiangyang Xue, and Jian Pu. Multi-to-single knowledge distillation for point cloud semantic segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9303–9309. IEEE, 2023.

Gorjan Radevski, Dusan Grujicic, Matthew B. Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5190–5201. IEEE, 2023.

Alec Radford. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Md Mostafijur Rahman, Mustafa Munir, Debesh Jha, Ulas Bagci, and Radu Marculescu. Pp-sam: Perturbed prompts for robust adaption of segment anything model for polyp segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4989–4995, 2024.

Arushi Rai, Kyle Buettner, and Adriana Kovashka. Strategies to leverage foundational model knowledge in object affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1714–1723, 2024.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12490–12500, 2024.

Jun Rao, Xv Meng, Liang Ding, Shuhan Qi, Xuebo Liu, Min Zhang, and Dacheng Tao. Parameter-efficient and student-friendly knowledge distillation. *IEEE Transactions on Multimedia*, 2023.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18082–18091, 2022.

Amirhossein Rasoulian, Soorena Salari, and Yiming Xiao. Weakly supervised intracranial hemorrhage segmentation using head-wise gradient-infused self-attention maps from a swin transformer in categorical learning. *arXiv preprint arXiv:2304.04902*, 2023.

Jiaqian Ren, Hao Peng, Lei Jiang, Jia Wu, Yongxin Tong, Lihong Wang, Xu Bai, Bo Wang, and Qiang Yang. Transferring knowledge distillation for multilingual social event detection. *arXiv preprint arXiv:2108.03084*, 2021.

Sucheng Ren, Fangyun Wei, Zheng Zhang, and Han Hu. Tinymim: An empirical study of distilling mim pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3687–3697, 2023a.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Yeqing Ren, Haipeng Peng, Lixiang Li, Xiaopeng Xue, Yang Lan, and Yixian Yang. A voice spoofing detection framework for iot systems with feature pyramid and online knowledge distillation. *Journal of Systems Architecture*, 143:102981, 2023b.

Yeqing Ren, Haipeng Peng, Lixiang Li, and Yixian Yang. Lightweight voice spoofing detection using improved one-class learning and knowledge distillation. *IEEE Transactions on Multimedia*, 2023c.

Saed Rezayi, Handong Zhao, Sungchul Kim, Ryan A Rossi, Nedim Lipka, and Sheng Li. Edge: Enriching knowledge graph embeddings with external text. *arXiv preprint arXiv:2104.04909*, 2021.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14902–14912, 2021.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18603–18613, 2022.

V Sanh. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Pritam Sarkar and Ali Etemad. Xkd: Cross-modal knowledge distillation with domain alignment for video representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14875–14885, 2024.

Ioannis Sarridis, Christos Koutlis, Giorgos Kordopatis-Zilos, Ioannis Kompatsiaris, and Symeon Papadopoulos. Indistill: Information flow-preserving knowledge distillation for model compression. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 9033–9042. IEEE, 2025.

Bharat Bhusan Sau and Vineeth N Balasubramanian. Deep model compression: Distilling knowledge from noisy teachers. *arXiv preprint arXiv:1610.09650*, 2016.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9891–9901, 2022.

Florian Schmid, Khaled Koutini, and Gerhard Widmer. Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation. In *ICASSP 2023-2023 IEEE international Conference on acoustics, Speech and signal processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Siamak Shakeri, Abhinav Sethy, and Cheng Cheng. Knowledge distillation in document retrieval. *arXiv preprint arXiv:1911.11065*, 2019.

Chao Shang, Hongliang Li, Fanman Meng, Qingbo Wu, Heqian Qiu, and Lanxiao Wang. Incrementer: Transformer for class-incremental semantic segmentation with knowledge distillation focusing on old class. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7214–7224, 2023a.

Jinghuan Shang, Karl Schmeckpeper, Brandon B May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. *arXiv preprint arXiv:2407.20179*, 2024.

Ronghua Shang, Wenzheng Li, Songling Zhu, Licheng Jiao, and Yangyang Li. Multi-teacher knowledge distillation based on joint guidance of probe and adaptive corrector. *Neural Networks*, 164:345–356, 2023b.

Jiawei Shao, Fangzhao Wu, and Jun Zhang. Selective knowledge sharing for privacy-preserving federated distillation without a good teacher. *Nature Communications*, 15(1):349, 2024a.

Shuwei Shao, Zhongcai Pei, Weihai Chen, Ran Li, Zhong Liu, and Zhengguo Li. Urcdc-depth: Uncertainty rectified cross-distillation with cutflip for monocular depth estimation. *IEEE Trans. Multim.*, 26:3341–3353, 2024b.

Saurabh Sharma, Atul Kumar, and Joydeep Chandra. Confidence matters: Enhancing medical image classification through uncertainty-driven contrastive self-distillation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 15010 of *Lecture Notes in Computer Science*, pp. 133–142. Springer, 2024.

Chengchao Shen, Xinchao Wang, Jie Song, Li Sun, and Mingli Song. Amalgamating knowledge towards comprehensive classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3068–3075, 2019a.

Chengchao Shen, Mengqi Xue, Xinchao Wang, Jie Song, Li Sun, and Mingli Song. Customizing student networks from heterogeneous teachers via adaptive knowledge amalgamation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3504–3513, 2019b.

Fengyi Shen, Akhil Gurram, Ziyuan Liu, He Wang, and Alois Knoll. Diga: Distil to generalize and then adapt for domain adaptive semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15866–15877. IEEE, 2023.

Peng Shen, Xugang Lu, Sheng Li, and Hisashi Kawai. Feature representation of short utterances based on knowledge distillation for spoken language identification. In *Interspeech*, pp. 1813–1817, 2018.

Peng Shen, Xugang Lu, Sheng Li, and Hisashi Kawai. Interactive learning of teacher-student model for short utterance spoken language identification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5981–5985. IEEE, 2019c.

Peng Shen, Xugang Lu, Sheng Li, and Hisashi Kawai. Knowledge distillation-based representation learning for short-utterance spoken language identification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2674–2683, 2020.

Zhiqiang Shen, Zhankui He, and Xiangyang Xue. Meal: Multi-model ensemble via adversarial learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 4886–4893, 2019d.

Wentao Shi, Jing Xu, and Pan Gao. Ssformer: A lightweight transformer for semantic segmentation. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–5. IEEE, 2022.

Xuepeng Shi, Georgi Dikov, Gerhard Reitmayr, Tae-Kyun Kim, and Mohsen Ghafoorian. 3d distillation: Improving self-supervised monocular depth estimation on reflective surfaces. In *ICCV*, pp. 9099–9109. IEEE, 2023.

Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. *Advances in Neural Information Processing Systems*, 35:33754–33767, 2022a.

Sungho Shin, Joosoon Lee, Junseok Lee, Yeonguk Yu, and Kyoobin Lee. Teaching where to look: Attention similarity knowledge distillation for low resolution face recognition. In *ECCV*, volume 13672 of *Lecture Notes in Computer Science*, pp. 631–647. Springer, 2022b.

Wooseok Shin, Byung Hoon Lee, Jin Sob Kim, Hyun Joon Park, and Sung Won Han. Metricgan-okd: multi-metric optimization of metricgan via online knowledge distillation for speech enhancement. In *International Conference on Machine Learning*, pp. 31521–31538. PMLR, 2023.

Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5311–5320, 2021.

Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021.

Oriane Siméoni, Chloé Sekkat, Gilles Puy, Antonín Vobeckỳ, Éloi Zablocki, and Patrick Pérez. Unsupervised object localization: Observing the background to discover objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3176–3186, 2023.

Eunjin Son and Sang Jun Lee. Cabins: Clip-based adaptive bins for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4557–4567, 2024.

Kaiyou Song, Jin Xie, Shan Zhang, and Zimeng Luo. Multi-mode online knowledge distillation for self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11848–11857, 2023a.

Liangchen Song, Xuan Gong, Helong Zhou, Jiajie Chen, Qian Zhang, David Doermann, and Junsong Yuan. Exploring the knowledge transferred by response-based teacher-student distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 2704–2713, 2023b.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023c.

Efstathia Soufleri, Deepak Ravikumar, and Kaushik Roy. Advancing compressed video action recognition through progressive knowledge distillation. *arXiv preprint arXiv:2407.02713*, 2024.

Sharath Turuvekere Sreenivas, Saurav Muralidharan, Raviraj Joshi, Marcin Chochowski, Ameya Sunil Mahabaleshwarkar, Gerald Shen, Jiaqi Zeng, Zijia Chen, Yoshi Suhara, Shizhe Diao, et al. Llm pruning and distillation in practice: The minitron approach. *arXiv preprint arXiv:2408.11796*, 2024.

Suraj Srinivas and François Fleuret. Knowledge transfer with jacobian matching. In *International Conference on Machine Learning*, pp. 4723–4731. PMLR, 2018.

Pavel Suma and Giorgos Tolias. Large-to-small image resolution asymmetry in deep metric learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1451–1460. IEEE, 2023.

Huawei Sun, Nastassia Vysotskaya, Tobias Sukianto, Hao Feng, Julius Ott, Xiangyuan Peng, Lorenzo Servadei, and Robert Wille. Lircdepth: Lightweight radar-camera depth estimation via knowledge distillation and uncertainty guidance. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15731–15740, 2024a.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*, 2019.

Weixuan Sun, Zheyuan Liu, Yanhao Zhang, Yiran Zhong, and Nick Barnes. An alternative to wsss? an empirical study of the segment anything model (sam) on weakly-supervised semantic segmentation problems. *arXiv preprint arXiv:2305.01586*, 2023a.

Wujie Sun, Defang Chen, Can Wang, Deshi Ye, Yan Feng, and Chun Chen. Holistic weighted distillation for semantic segmentation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 396–401. IEEE, 2023b.

Ximeng Sun, Pengchuan Zhang, Peizhao Zhang, Hardik Shah, Kate Saenko, and Xide Xia. Dime-fm: Distilling multimodal and efficient foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15521–15533, 2023c.

Yue Sun, Lingfeng Huang, Qi Zhu, and Dong Liang. Cs-kd: Confused sample knowledge distillation for semantic segmentation of aerial imagery. In *International Conference on Intelligent Computing*, pp. 266–278. Springer, 2024b.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*, 2020.

Ryoichi Takashima, Sheng Li, and Hisashi Kawai. An investigation of a knowledge distillation method for ctc acoustic models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5809–5813, 2018. doi: 10.1109/ICASSP.2018.8461995. URL `https://ieeexplore.ieee.org/document/8461995/`.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. Multilingual neural machine translation with knowledge distillation. *arXiv preprint arXiv:1902.10461*, 2019.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*, 2019.

Ruining Tang, Zhenyu Liu, Yangguang Li, Yiguo Song, Hui Liu, Qide Wang, Jing Shao, Guifang Duan, and Jianrong Tan. Task-balanced distillation for object detection. *CoRR*, abs/2208.03006, 2022a.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20730–20740, 2022b.

Yuwu Tang, Ziang Ma, and Haitao Zhang. Enhanced feature learning with normalized knowledge distillation for audio tagging. In *Proc. Interspeech 2024*, pp. 1695–1699, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Ajinkya Tejankar, Soroush Abbasi Koohpayegani, Vipin Pillai, Paolo Favaro, and Hamed Pirsiavash. Isd: Self-supervised learning by iterative similarity distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9609–9618, 2021.

Fida Mohammad Thoker and Juergen Gall. Cross-modal knowledge distillation for action recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 6–10. IEEE, 2019.

Jialin Tian, Xing Xu, Zheng Wang, Fumin Shen, and Xin Liu. Relationship-preserving knowledge distillation for zero-shot sketch based image retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, pp. 5473–5481. ACM, 2021.

Yijun Tian, Chuxu Zhang, Zhichun Guo, Xiangliang Zhang, and Nitesh V Chawla. Nosmog: Learning noise-robust and structure-aware mlps on graphs. *arXiv preprint arXiv:2208.10010*, 2022a.

Yijun Tian, Shichao Pei, Xiangliang Zhang, Chuxu Zhang, and Nitesh Chawla. Knowledge distillation on graphs: A survey. *ACM Computing Surveys*, 2023a.

Yingjie Tian, Shiding Sun, and Jingjing Tang. Multi-view teacher–student network. *Neural Networks*, 146: 69–84, 2022b.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

Zhiqiang Tian, Weigang Li, Junwei Hu, and Chunhua Deng. Joint graph entropy knowledge distillation for point cloud classification and robustness against corruptions. *Information Sciences*, 648:119542, 2023b.

Zhuotao Tian, Pengguang Chen, Xin Lai, Li Jiang, Shu Liu, Hengshuang Zhao, Bei Yu, Ming-Chang Yang, and Jiaya Jia. Adaptive perspective distillation for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1372–1387, 2022c.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Duc-Tuan Truong, Ruijie Tao, Jia Qi Yip, Kong Aik Lee, and Eng Siong Chng. Emphasized non-target speaker knowledge in knowledge distillation for automatic speaker verification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10336–10340. IEEE, 2024.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1365–1374, 2019.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019.

Ardian Umam, Cheng-Kun Yang, Min-Hung Chen, Jen-Hui Chuang, and Yen-Yu Lin. Partdistill: 3d shape part segmentation by vision-language model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3470–3479, 2024.

Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11607–11616, 2021.

Tuan Van Vo, Minh Nhat Vu, Baoru Huang, Toan Nguyen, Ngan Le, Thieu Vo, and Anh Nguyen. Open-vocabulary affordance detection using knowledge distillation and text-point correlation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13968–13975. IEEE, 2024.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Jayakorn Vongkulbhisal, Phongtharin Vinayavekhin, and Marco Visentini-Scarzanella. Unifying heterogeneous classifiers with distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3175–3184, 2019.

Bo Wan and Tinne Tuytelaars. Exploiting clip for zero-shot hoi detection requires knowledge distillation at multiple levels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1805–1815, 2024.

Ziyu Wan, Despoina Paschalidou, Ian Huang, Hongyu Liu, Bokui Shen, Xiaoyu Xiang, Jing Liao, and Leonidas Guibas. Cad: Photorealistic 3d generation via adversarial distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10194–10207, 2024.

Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3835–3844, 2022a.

Can Wang, Zhe Wang, Defang Chen, Sheng Zhou, Yan Feng, and Chun Chen. Online adversarial knowledge distillation for graph neural networks. *Expert Systems with Applications*, 237:121671, 2024a.

Chao Wang and Zheng Tang. The staged knowledge distillation in video classification: Harmonizing student progress by a complementary weakly supervised framework. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Chen Wang, Jiang Zhong, Qizhu Dai, Rongzhen Li, Qien Yu, and Bin Fang. Local structure consistency and pixel-correlation distillation for compact semantic segmentation. *Applied Intelligence*, 53(6):6307–6323, 2023a.

Chen Wang, Jiang Zhong, Qizhu Dai, Yafei Qi, Rongzhen Li, Qin Lei, Bin Fang, and Xue Li. Prrd: Pixel-region relation distillation for efficient semantic segmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Chen Wang, Jiang Zhong, Qizhu Dai, Yafei Qi, Fengyuan Shi, Bin Fang, and Xue Li. Multi-view knowledge distillation for efficient semantic segmentation. *Journal of Real-Time Image Processing*, 20(2):39, 2023c.

Chen Wang, Jiang Zhong, Qizhu Dai, Yafei Qi, Qien Yu, Fengyuan Shi, Rongzhen Li, Xue Li, and Bin Fang. Channel correlation distillation for compact semantic segmentation. *International Journal of Pattern Recognition and Artificial Intelligence*, 37(03):2350004, 2023d.

Chen Wang, Jiang Zhong, Qizhu Dai, Qien Yu, Yafei Qi, Bin Fang, and Xue Li. Mted: multiple teachers ensemble distillation for compact semantic segmentation. *Neural Computing and Applications*, 35(16): 11789–11806, 2023e.

Chenyu Wang, Toshio Endo, Takahiro Hirofuchi, and Tsutomu Ikegami. Pyramid swin transformer: Different-size windows swin transformer for image classification and object detection. In *VISIGRAPP (5: VISAPP)*, pp. 583–590, 2023f.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *arXiv preprint arXiv:2411.03350*, 2024b.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. Selective knowledge distillation for neural machine translation. *arXiv preprint arXiv:2105.12967*, 2021a.

Guiqin Wang, Peng Zhao, Yanjiang Shi, Cong Zhao, and Shusen Yang. Generative model-based feature knowledge distillation for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15474–15482, 2024c.

Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3635–3647, 2024d.

Hong Wang, Yuefan Deng, Shinjae Yoo, Haibin Ling, and Yuewei Lin. Agkd-bml: Defense against adversarial attack by attention guided knowledge distillation and bi-directional metric learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7658–7667, 2021b.

Hongyuan Wang, Ziyan Wei, Qingting Tang, Shuli Cheng, Liejun Wang, and Yongming Li. Attention guidance distillation network for efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6287–6296. IEEE, 2024e.

Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 14223 of *Lecture Notes in Computer Science*, pp. 216–226. Springer, 2023g.

Jiabao Wang, Yuming Chen, Zhaohui Zheng, Xiang Li, Ming-Ming Cheng, and Qibin Hou. Crosskd: Cross-head knowledge distillation for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19723–19733. IEEE, 2024f.

Jiahang Wang, Sheng Jin, Wentao Liu, Weizhong Liu, Chen Qian, and Ping Luo. When human pose estimation meets robustness: Adversarial algorithms and benchmarks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11855–11864. Computer Vision Foundation / IEEE, 2021c.

Liang Wang, Nan Yang, and Furu Wei. Learning to retrieve in-context examples for large language models. *arXiv preprint arXiv:2307.07164*, 2023h.

Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3048–3068, 2021.

Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 608–619, 2021d.

Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14090–14100, 2021e.

Lu Wang, Liuchi Xu, Xiong Yang, Zhenhua Huang, and Jun Cheng. Debiased distillation for consistency regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7799–7807, 2025a.

Lubo Wang, Di Lin, Kairui Yang, Ruonan Liu, Qing Guo, Wuyuan Xie, Miaohui Wang, Lingyu Liang, Yi Wang, and Ping Li. Voxel proposal network via multi-frame knowledge distillation for semantic scene completion. *Advances in Neural Information Processing Systems*, 37:101096–101115, 2025b.

Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11186–11196, 2023i.

Ning-Hsu Wang and Yu-Lun Liu. Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. Scott: Self-consistent chain-of-thought distillation. *arXiv preprint arXiv:2305.01879*, 2023j.

Qi Wang, Lu Liu, Wenxin Yu, Shiyu Chen, Jun Gong, and Peng Chen. Bckd: Block-correlation knowledge distillation. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 3225–3229. IEEE, 2023k.

Qi Wang, Wenxin Yu, Lu Che, Chang Liu, Zhiqiang Zhang, Jun Gong, and Peng Chen. Similarity knowledge distillation with calibrated mask. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5770–5774. IEEE, 2024g.

Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6312–6322, 2023l.

Sijie Wang, Rui She, Qiyu Kang, Xingchao Jian, Kai Zhao, Yang Song, and Wee Peng Tay. Distilvpr: Cross-modal knowledge distillation for visual place recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 10377–10385, 2024h.

Song Wang, Wentong Li, Wenyu Liu, Xiaolu Liu, and Jianke Zhu. Lidar2map: In defense of lidar-based semantic map construction using online camera distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5186–5195, 2023m.

Song Wang, Jiawei Yu, Wentong Li, Wenyu Liu, Xiaolu Liu, Junbo Chen, and Jianke Zhu. Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14792–14801, 2024i.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018a.

Wanwei Wang, Wei Hong, Feng Wang, and Jinke Yu. Gan-knowledge distillation for one-stage object detection. *IEEE Access*, 8:60719–60727, 2020a.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020b.

Wenxuan Wang, Xingjian He, Yisi Zhang, Longteng Guo, Jiachen Shen, Jiangyun Li, and Jing Liu. Cm-masksd: Cross-modality masked self-distillation for referring image segmentation. *IEEE Transactions on Multimedia*, 2024j.

Xiao Wang, Shiao Wang, Chuanming Tang, Lin Zhu, Bo Jiang, Yonghong Tian, and Jin Tang. Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19248–19257. IEEE, 2024k.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Kdgan: Knowledge distillation with generative adversarial networks. *Advances in neural information processing systems*, 31, 2018b.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Adversarial distillation for learning with privileged provisions. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):786–797, 2019.

Xucong Wang, Pengkun Wang, Shurui Zhang, Miao Fang, and Yang Wang. Multi-label self knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21330–21338, 2025c.

Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3124–3134, 2023n.

Yanan Wang, Donghuo Zeng, Shinya Wada, and Satoshi Kurihara. Videoadviser: Video knowledge distillation for multimodal transfer learning. *IEEE Access*, 11:51229–51240, 2023o.

Yanbo Wang, Shaohui Lin, Yanyun Qu, Haiyan Wu, Zhizhong Zhang, Yuan Xie, and Angela Yao. Towards compact single image super-resolution via contrastive self-distillation. In Zhi-Hua Zhou (ed.), *IJCAI*, pp. 1122–1128. ijcai.org, 2021f.

Yang Wang and Tao Zhang. Osffnet: Omni-stage feature fusion network for lightweight image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 5660–5668. AAAI Press, 2024.

Yaqi Wang, Peng Cao, Qingshan Hou, Linqi Lan, Jinzhu Yang, Xiaoli Liu, and Osmar R. Zaïane. Progressively correcting soft labels via teacher team for knowledge distillation in medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 15009 of *Lecture Notes in Computer Science*, pp. 521–530. Springer, 2024l.

Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9332–9341, 2020c.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022b.

Yu Wang, Xin Li, Shengzhao Weng, Gang Zhang, Haixiao Yue, Haocheng Feng, Junyu Han, and Errui Ding. Kd-detr: Knowledge distillation for detection transformer with consistent distillation points sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16016–16025, 2024m.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. Sinsr: Diffusion-based image super-resolution in a single step. In *CVPR*, pp. 25796–25805. IEEE, 2024n.

Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pp. 346–362. Springer, 2020d.

Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao. Adversarial learning of portable student networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018c.

Yutian Wang and Yi Xu. Relation-based multi-teacher knowledge distillation. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, 2024.

Yuzheng Wang, Zhaoyu Chen, Dingkang Yang, Pinxue Guo, Kaixun Jiang, Wenqiang Zhang, and Lizhe Qi. Out of thin air: Exploring data-free adversarial robustness distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5776–5784, 2024o.

Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, and Xiaodong Yang. Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8637–8646, 2023p.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36:8406–8441, 2023q.

Zi-Rui Wang and Jun Du. Joint architecture and knowledge distillation in cnn for chinese text recognition. *Pattern Recognition*, 111:107722, 2021.

Zipeng Wang, Xuehui Yu, Xumeng Han, Wenwen Yu, Zhixun Huang, Jianbin Jiao, and Zhenjun Han. P2seg: Pointly-supervised segmentation via mutual distillation. In *International Conference on Learning Representations (ICLR)*, 2024p.

Taiba Majid Wani and Irene Amerini. Audio deepfake detection: A continual approach with feature distillation and dynamic class rebalancing. In *International Conference on Pattern Recognition*, pp. 211–227. Springer, 2025.

Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R Hershey. Student-teacher network learning with enhanced features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5275–5279. IEEE, 2017.

Min Wei, Jingkai Zhou, Junyao Sun, and Xuesong Zhang. Adversarial score distillation: When score distillation meets gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8131–8141, 2024a.

Shicai Wei, Chunbo Luo, and Yang Luo. Scaled decoupled distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15975–15983, 2024b.

Yi Wei, Zibu Wei, Yongming Rao, Jiaxin Li, Jie Zhou, and Jiwen Lu. Lidar distillation: Bridging the beam-induced domain gap for 3d object detection. In *European Conference on Computer Vision*, pp. 179–195. Springer, 2022.

Yongxian Wei, Zixuan Hu, Li Shen, Zhenyi Wang, Chun Yuan, and Dacheng Tao. Open-vocabulary customization from clip via data-free knowledge distillation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Di Wu, Pengfei Chen, Xuehui Yu, Guorong Li, Zhenjun Han, and Jianbin Jiao. Spatial self-distillation for object detection with inaccurate bounding boxes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6832–6842. IEEE, 2023a.

Guile Wu and Shaogang Gong. Peer collaborative learning for online knowledge distillation. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 35, pp. 10302–10310, 2021.

Hui Wu, Min Wang, Wengang Zhou, Houqiang Li, and Qi Tian. Contextual similarity distillation for asymmetric image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9479–9488. IEEE, 2022.

Jianfeng Wu, Yongzhu Hua, Shengying Yang, Hongshuai Qin, and Huibin Qin. Speech enhancement using generative adversarial network by distilling knowledge from statistical method. *Applied Sciences*, 9(16): 3396, 2019.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21970–21980, 2023b.

Lirong Wu, Haitao Lin, Zhangyang Gao, Guojiang Zhao, and Stan Z Li. A teacher-free graph knowledge distillation framework with dual self-distillation. *IEEE Transactions on Knowledge and Data Engineering*, 2024a.

Meng-Chieh Wu and Ching-Te Chiu. Multi-teacher knowledge distillation for compressed video action recognition based on deep learning. *Journal of systems architecture*, 103:101695, 2020.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. Lamini-lm: A diverse herd of distilled models from large-scale instructions. *arXiv preprint arXiv:2304.14402*, 2023c.

Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15254–15264, 2023d.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *International Conference on Learning Representations (ICLR)*, 2024b.

Yao Wu, Mingwei Xing, Yachao Zhang, Yuan Xie, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Cross-modal unsupervised domain adaptation for 3d semantic segmentation via bidirectional fusion-then-distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 490–498, 2023e.

Zizhang Wu, Yunzhe Wu, Jian Pu, Xianzhi Li, and Xiaoquan Wang. Attention-based depth distillation with 3d-aware positional encoding for monocular 3d object detection. In *AAAI*, pp. 2892–2900. AAAI Press, 2023f.

Monika Wysoczańska, Michaël Ramamonjisoa, Tomasz Trzciński, and Oriane Siméoni. Clip-diy: Clip dense inference yields open-vocabulary semantic segmentation for-free. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1403–1413, 2024a.

Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pp. 320–337. Springer, 2024b.

Zhaoyang Xia, Youquan Liu, Xin Li, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, and Yu Qiao. Scpnet: Semantic scene completion on point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17642–17651, 2023.

Jia-Wen Xiao, Chang-Bin Zhang, Jiekang Feng, Xialei Liu, Joost van de Weijer, and Ming-Ming Cheng. Endpoints weight fusion for class incremental semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7204–7213. IEEE, 2023a.

Jian Xiao, Keren Zhang, Xianyong Xu, Shuai Liu, Sheng Wu, Zhihong Huang, and Linfeng Li. Improving accuracy and efficiency of monocular depth estimation in power grid environments using point cloud optimization and knowledge distillation. *Energies*, 17(16):4068, 2024.

Zhiwen Xiao, Huanlai Xing, Bowen Zhao, Rong Qu, Shouxi Luo, Penglin Dai, Ke Li, and Zonghai Zhu. Deep contrastive representation learning with self-distillation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023b.

Jiafeng Xie, Bing Shuai, Jian-Fang Hu, Jingyang Lin, and Wei-Shi Zheng. Improving fast segmentation with teacher-student learning. *arXiv preprint arXiv:1810.08476*, 2018.

Jiao Xie, Linrui Gong, Shitong Shao, Shaohui Lin, and Linkai Luo. Hybrid knowledge distillation from intermediate layers for efficient single image super-resolution. *Neurocomputing*, 554:126592, 2023a.

Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: Cross language image matching for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4483–4492, 2022.

Johnathan Xie and Shuai Zheng. Zero-shot object detection through vision-language embedding alignment. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 1–15. IEEE, 2022.

Qingsong Xie, Zhenyi Liao, Zhijie Deng, Haonan Lu, et al. Tlcm: Training-efficient latent consistency model for image generation with 2-8 steps. *arXiv preprint arXiv:2406.05768*, 2024a.

Sirui Xie, Zhisheng Xiao, Diederik Kingma, Tingbo Hou, Ying Nian Wu, Kevin P Murphy, Tim Salimans, Ben Poole, and Ruiqi Gao. Em distillation for one-step diffusion models. *Advances in Neural Information Processing Systems*, 37:45073–45104, 2024b.

Yi Xie, Huaidong Zhang, Xuemiao Xu, Jianqing Zhu, and Shengfeng He. Towards a smaller student: Capacity dynamic distillation for efficient image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16006–16015. IEEE, 2023b.

Yi Xie, Yihong Lin, Wenjie Cai, Xuemiao Xu, Huaidong Zhang, Yong Du, and Shengfeng He. D3still: Decoupled differential distillation for asymmetric image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17181–17190. IEEE, 2024c.

Yi Xie, Hanxiao Wu, Yihong Lin, Jianqing Zhu, and Huanqiang Zeng. Pairwise difference relational distillation for object re-identification. *Pattern Recognition*, 152:110455, 2024d.

Yushan Xie, Yuejia Yin, Qingli Li, and Yan Wang. Deep mutual distillation for semi-supervised medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 14222 of *Lecture Notes in Computer Science*, pp. 540–550. Springer, 2023c.

Xiaomeng Xin, Heping Song, and Jianping Gou. A new similarity-based relational knowledge distillation method. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3535–3539. IEEE, 2024.

Xiaohan Xing, Yuenan Hou, Hang Li, Yixuan Yuan, Hongsheng Li, and Max Q-H Meng. Categorical relation-preserving contrastive knowledge distillation for medical image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pp. 163–173. Springer, 2021.

Haohang Xu, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Bag of instances aggregation boosts self-supervised distillation. *arXiv preprint arXiv:2107.01691*, 2021.

Jianqing Xu, Shen Li, Ailin Deng, Miao Xiong, Jiaying Wu, Jiaxiang Wu, Shouhong Ding, and Bryan Hooi. Probabilistic knowledge distillation of face ensembles. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3489–3498, 2023a.

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. Lrspeech: Extremely low-resource speech synthesis and recognition. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2802–2812, 2020a.

Kunlun Xu, Xu Zou, and Jiahuan Zhou. Lstkc: Long short-term knowledge consolidation for lifelong person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 16202–16210. AAAI Press, 2024a.

Kunran Xu, Lai Rui, Yishi Li, and Lin Gu. Feature normalized knowledge distillation for image classification. In *European conference on computer vision*, pp. 664–680. Springer, 2020b.

Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Learning multi-modal class-specific tokens for weakly supervised dense object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19596–19605, 2023b.

Lixiang Xu, Qingzhe Cui, Richang Hong, Wei Xu, Enhong Chen, Xin Yuan, Chenglong Li, and Yuanyan Tang. Group multi-view transformer for 3d shape analysis with spatial encoding. *IEEE Transactions on Multimedia*, 2024b.

Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pp. 736–753. Springer, 2022a.

Quanzheng Xu, Liyu Liu, and Bing Ji. Knowledge distillation guided by multiple homogeneous teachers. *Information Sciences*, 607:230–243, 2022b.

Ruochen Xu and Yiming Yang. Cross-lingual distillation for text classification. *arXiv preprint arXiv:1705.02073*, 2017.

Ting-Bing Xu and Cheng-Lin Liu. Data-distortion guided self-distillation for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 5565–5572, 2019.

Xuenan Xu, Haohe Liu, Mengyue Wu, Wenwu Wang, and Mark D Plumbley. Efficient audio captioning with encoder-level knowledge distillation. *arXiv preprint arXiv:2407.14329*, 2024c.

Yangyang Xu, Yibo Yang, and Lefei Zhang. Multi-task learning with knowledge distillation for dense prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21493–21502. IEEE, 2023c.

Yangyang Xu, Yibo Yang, and Lefei Zhang. Multi-task learning with knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21550–21559, 2023d.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2022c.

Zheng Xu, Yen-Chang Hsu, and Jiawei Huang. Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks. *arXiv preprint arXiv:1709.00513*, 2017.

Jun Xue, Cunhang Fan, Jiangyan Yi, Chenglong Wang, Zhengqi Wen, Dan Zhang, and Zhao Lv. Learning from yourself: A self-distillation method for fake speech detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Siming Yan, Chen Song, Youkang Kong, and Qixing Huang. Multi-view representation is what you need for point-cloud pre-training. *arXiv preprint arXiv:2306.02558*, 2023.

Xu Yan, Heshen Zhan, Chaoda Zheng, Jiantao Gao, Ruimao Zhang, Shuguang Cui, and Zhen Li. Let images give you more: Point cloud cross-modal training for shape analysis. *Advances in Neural Information Processing Systems*, 35:32398–32411, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Cheng Yang, Jiawei Liu, and Chuan Shi. Extract the knowledge of graph neural networks and go beyond it: An effective knowledge distillation framework. In *Proceedings of the web conference 2021*, pp. 1227–1237, 2021.

Chenhongyi Yang, Mateusz Ochal, Amos J. Storkey, and Elliot J. Crowley. Prediction-guided distillation for dense object detection. In *European Conference on Computer Vision (ECCV)*, volume 13669 of *Lecture Notes in Computer Science*, pp. 123–138. Springer, 2022a.

Chenxiao Yang, Qitian Wu, and Junchi Yan. Geometric knowledge distillation: Topology compression for graph neural networks. *Advances in Neural Information Processing Systems*, 35:29761–29775, 2022b.

Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12319–12328, 2022c.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, and Yongjun Xu. Clip-kd: An empirical study of distilling clip models. *arXiv preprint arXiv:2307.12732*, 2023a.

Chuanguang Yang, Zhulin An, Helong Zhou, Fuzhen Zhuang, Yongjun Xu, and Qian Zhang. Online knowledge distillation via mutual contrastive learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10212–10227, 2023b.

Chuanguang Yang, Xinqiang Yu, Zhulin An, and Yongjun Xu. Categories of response-based, feature-based, and relation-based knowledge distillation. In *Advancements in Knowledge Distillation: Towards New Horizons of Intelligent Systems*, pp. 1–32. Springer, 2023c.

Guoliang Yang, Shuaiying Yu, Yangyang Sheng, and Hao Yang. Attention and feature transfer based knowledge distillation. *Scientific Reports*, 13(1):18369, 2023d.

Jihan Yang, Shaoshuai Shi, Runyu Ding, Zhe Wang, and Xiaojuan Qi. Towards efficient 3d object detection with knowledge distillation. *Advances in Neural Information Processing Systems*, 35:21300–21313, 2022d.

Longrong Yang, Xianpan Zhou, Xuewei Li, Liang Qiao, Zheyang Li, Ziwei Yang, Gaoang Wang, and Xi Li. Bridging cross-task protocol inconsistency for distillation in dense object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17129–17138. IEEE, 2023e.

Xiaobo Yang and Xiaojin Gong. Foundation model assisted weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 523–532, 2024.

Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, and Daxin Jiang. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 690–698, 2020.

Ze Yang, Ruibo Li, Evan Ling, Chi Zhang, Yiming Wang, Dezhao Huang, Keng Teck Ma, Minhoe Hur, and Guosheng Lin. Label-guided knowledge distillation for continual semantic segmentation on 2d images and 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18601–18612, 2023f.

Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4633–4642. IEEE, 2022e.

Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *European Conference on Computer Vision (ECCV)*, pp. 53–69. Springer, 2022f.

Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, and Yu Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17185–17194, 2023g.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4212–4222. IEEE, 2023h.

Zhendong Yang, Zhe Li, Ailing Zeng, Zexian Li, Chun Yuan, and Yu Li. Vitkd: Feature-based knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1379–1388, 2024.

Lewei Yao, Renjie Pi, Hang Xu, Wei Zhang, Zhenguo Li, and Tong Zhang. G-detkd: Towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3571–3580. IEEE, 2021.

Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022.

Yuan Yao, Yuanhan Zhang, Zhenfei Yin, Jiebo Luo, Wanli Ouyang, and Xiaoshui Huang. 3d point cloud pre-training with knowledge distilled from 2d images. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Jianglong Ye, Naiyan Wang, and Xiaolong Wang. Featurenerf: Learning generalizable nerfs by distilling foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8962–8973, 2023a.

Jingwen Ye, Yixin Ji, Xinchao Wang, Xin Gao, and Mingli Song. Data-free knowledge amalgamation via group-stack dual-gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12516–12525, 2020.

Suhang Ye, Yingyi Zhang, Jie Hu, Liujuan Cao, Shengchuan Zhang, Lei Shen, Jun Wang, Shouhong Ding, and Rongrong Ji. Distilpose: Tokenized pose regression with heatmap distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2163–2172. IEEE, 2023b.

Yiwen Ye, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Desd: Self-supervised learning with deep self-distillation for 3d medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 13434 of *Lecture Notes in Computer Science*, pp. 545–555. Springer, 2022.

Jingjun Yi, Qi Bi, Hao Zheng, Haolan Zhan, Wei Ji, Yawen Huang, Shaoxin Li, Yuexiang Li, Yefeng Zheng, and Feiyue Huang. Hallucinated style distillation for single domain generalization in medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 15010 of *Lecture Notes in Computer Science*, pp. 438–448. Springer, 2024a.

Xuanyu Yi, Zike Wu, Qingshan Xu, Pan Zhou, Joo-Hwee Lim, and Hanwang Zhang. Diffusion time-step curriculum for one image to 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9948–9958, 2024b.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4133–4141, 2017.

Lirong Yin, Lei Wang, Zhuohang Cai, Siyu Lu, Ruiyang Wang, Ahmed AlSanad, Salman A AlQahtani, Xiaobing Chen, Zhengtong Yin, Xiaolu Li, et al. Dpal-bert: A faster and lighter question answering model. *CMES-Computer Modeling in Engineering & Sciences*, 141(1), 2024a.

Shenglin Yin, Zhen Xiao, Mingxuan Song, and Jieyi Long. Adversarial distillation based on slack matching and attribution region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24605–24614, 2024b.

Siqi Yin and Lifan Jiang. Distilling knowledge from multiple foundation models for zero-shot image classification. *PloS one*, 19(9):e0310730, 2024.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024c.

Yifang Yin, Harsh Shrivastava, Ying Zhang, Zhenguang Liu, Rajiv Ratn Shah, and Roger Zimmermann. Enhanced audio tagging via multi-to single-modal teacher-student mutual learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 10709–10717, 2021.

Ji Won Yoon, Hyeonseung Lee, Hyung Yong Kim, Won Ik Cho, and Nam Soo Kim. Tutornet: Towards flexible knowledge distillation for end-to-end speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1626–1638, 2021.

Ji Won Yoon, Beom Jun Woo, Sunghwan Ahn, Hyeonseung Lee, and Nam Soo Kim. Inter-kd: Intermediate knowledge distillation for ctc-based automatic speech recognition. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pp. 1–8, 2022. doi: 10.1109/SLT54892.2023.10022581. URL `https://arxiv.org/abs/2211.15075`.

Chenyu You, Nuo Chen, Fenglin Liu, Dongchao Yang, and Yuexian Zou. Towards data distillation for end-to-end spoken conversational question answering. *arXiv preprint arXiv:2010.08923*, 2020a.

Chenyu You, Nuo Chen, and Yuexian Zou. Contextualized attention-based knowledge transfer for spoken conversational question answering. *arXiv preprint arXiv:2010.11066*, 2020b.

Chenyu You, Nuo Chen, and Yuexian Zou. Knowledge distillation for improved accuracy in spoken question answering. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7793–7797. IEEE, 2021a.

Chenyu You, Nuo Chen, and Yuexian Zou. Mrd-net: Multi-modal residual knowledge distillation for spoken question answering. In *IJCAI*, pp. 3985–3991, 2021b.

Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S. Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(9):2228–2237, 2022.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1285–1294, 2017.

Nikolaos-Antonios Ypsilantis, Kaifeng Chen, André Araujo, and Ondrej Chum. Udon: Universal dynamic online distillation for generic image representations. *Advances in Neural Information Processing Systems*, 37:86836–86859, 2024.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 325–341, 2018.

Lei Yu, Xinpeng Li, Youwei Li, Ting Jiang, Qi Wu, Haoqiang Fan, and Shuaicheng Liu. Dipnet: Efficiency distillation and iterative pruning for image super-resolution. In *CVPR Workshops*, pp. 1692–1701. IEEE, 2023.

Lu Yu, Shichao Pei, Lizhong Ding, Jun Zhou, Longfei Li, Chuxu Zhang, and Xiangliang Zhang. Sail: Self-augmented graph contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8927–8935, 2022a.

Pengfei Yu, Heng Ji, and Prem Natarajan. Lifelong event detection with knowledge transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5278–5290, 2021.

Ping-Chung Yu, Cheng Sun, and Min Sun. Data efficient 3d learner via knowledge transferred from 2d model. In *European Conference on Computer Vision*, pp. 182–198. Springer, 2022b.

Jianlong Yuan, Minh Hieu Phan, Liyang Liu, and Yifan Liu. Fakd: Feature augmented knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 595–605, 2024a.

Mengyang Yuan, Bo Lang, and Fengnan Quan. Student-friendly knowledge distillation. *Knowledge-Based Systems*, 296:111915, 2024b.

Ye Yuan, Jiacheng Shi, Zongyao Zhang, Kaiwei Chen, Jingzhi Zhang, Vincenzo Stoico, and Ivano Malavolta. The impact of knowledge distillation on the energy consumption and runtime efficiency of nlp models. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pp. 129–133, 2024c.

Nir Zabari and Yedid Hoshen. Semantic segmentation in-the-wild without seeing any segmentation examples. *arXiv preprint arXiv:2112.03185*, 2021.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pp. 106–122. Springer, 2022.

Jia Zeng, Li Chen, Hanming Deng, Lewei Lu, Junchi Yan, Yu Qiao, and Hongyang Li. Distilling focal knowledge from imperfect expert for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 992–1001, 2023.

Bingfeng Zhang, Jimin Xiao, and Yao Zhao. Dynamic feature regularized loss for weakly supervised semantic segmentation. *arXiv preprint arXiv:2108.01296*, 2021a.

Bingfeng Zhang, Siyue Yu, Yunchao Wei, Yao Zhao, and Jimin Xiao. Frozen clip: A strong backbone for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3796–3806, 2024a.

Chenrui Zhang and Yuxin Peng. Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. *arXiv preprint arXiv:1804.10069*, 2018.

Dingkun Zhang, Sijia Li, Chen Chen, Qingsong Xie, and Haonan Lu. Laptop-diff: Layer pruning and normalized distillation for compressing diffusion models. *arXiv preprint arXiv:2404.11098*, 2024b.

Feng Zhang, Xiatian Zhu, and Mao Ye. Fast human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3517–3526. IEEE, 2019a.

Hailin Zhang, Defang Chen, and Can Wang. Adaptive multi-teacher knowledge distillation with meta-learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1943–1948. IEEE, 2023a.

Haiming Zhang, Xu Yan, Dongfeng Bai, Jiantao Gao, Pan Wang, Bingbing Liu, Shuguang Cui, and Zhen Li. Radocc: Learning cross-modality occupancy knowledge through rendering assisted distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7060–7068, 2024c.

Hanlin Zhang, Shuai Lin, Weiyang Liu, Pan Zhou, Jian Tang, Xiaodan Liang, and Eric P Xing. Iterative graph self-distillation. *IEEE Transactions on Knowledge and Data Engineering*, 2023b.

Haofei Zhang, Feng Mao, Mengqi Xue, Gongfan Fang, Zunlei Feng, Jie Song, and Mingli Song. Knowledge amalgamation for object detection with transformers. *IEEE Transactions on Image Processing*, 32:2093–2106, 2023c.

Jinbao Zhang and Jun Liu. Voxel-to-pillar: Knowledge distillation of 3d object detection in point cloud. In *Proceedings of the 4th European Symposium on Software Engineering*, pp. 99–104, 2023.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024d.

Jisi Zhang, Catalin Zorila, Rama Doddipatla, and Jon Barker. Teacher-student mixit for unsupervised and semi-supervised speech separation. *arXiv preprint arXiv:2106.07843*, 2021b.

Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations (ICLR)*, 2021.

Linfeng Zhang and Kaisheng Ma. Structured knowledge distillation for accurate and efficient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15706–15724, 2023.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3713–3722, 2019b.

Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2021c.

Linfeng Zhang, Xin Chen, Xiaobing Tu, Pengfei Wan, Ning Xu, and Kaisheng Ma. Wavelet knowledge distillation: Towards efficient image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12464–12474, 2022a.

Linfeng Zhang, Runpei Dong, Hung-Shuo Tai, and Kaisheng Ma. Pointdistiller: structured knowledge distillation towards efficient and compact 3d detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21791–21801, 2023d.

Linfeng Zhang, Yukang Shi, Ke Wang, Zhipeng Zhang, Hung-Shuo Tai, Yuan He, and Kaisheng Ma. Structured knowledge distillation towards efficient multi-view 3d object detection. In *BMVC*, pp. 339–344, 2023e.

Minjia Zhang, Niranjan Uma Naresh, and Yuxiong He. Adversarial data augmentation for task-specific knowledge distillation of pre-trained transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11685–11693, 2022b.

Peizhen Zhang, Zijian Kang, Tong Yang, Xiangyu Zhang, Nanning Zheng, and Jian Sun. Lgd: Label-guided self-distillation for object detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3309–3317. AAAI Press, 2022c.

Qijian Zhang, Junhui Hou, and Yue Qian. Pointmcd: Boosting deep point cloud encoders via multi-view cross-modal distillation for 3d shape recognition. *IEEE Transactions on Multimedia*, 2023f.

Quan Zhang, Xiaoyu Liu, Wei Li, Hanting Chen, Junchao Liu, Jie Hu, Zhiwei Xiong, Chun Yuan, and Yunhe Wang. Distilling semantic priors from sam to efficient image restoration models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25409–25419, 2024e.

Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21769–21780, 2023g.

Sha Zhang, Jiajun Deng, Lei Bai, Houqiang Li, Wanli Ouyang, and Yanyong Zhang. Hvdistill: Transferring knowledge from images to point clouds via unsupervised hybrid-view distillation. *International Journal of Computer Vision*, pp. 1–15, 2024f.

Shichang Zhang, Yozen Liu, Yizhou Sun, and Neil Shah. Graph-less neural networks: Teaching old mlps new tricks via distillation. *arXiv preprint arXiv:2110.08727*, 2021d.

Tianli Zhang, Mengqi Xue, Jiangtao Zhang, Haofei Zhang, Yu Wang, Lechao Cheng, Jie Song, and Mingli Song. Generalization matters: Loss minima flattening via parameter hybridization for efficient online knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20176–20185, 2023h.

Tianlu Zhang, Hongyuan Guo, Qiang Jiao, Qiang Zhang, and Jungong Han. Efficient rgb-t tracking via cross-modality distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5404–5413, 2023i.

Weijia Zhang, Dongnan Liu, Weidong Cai, and Chao Ma. Cross-view consistency regularisation for knowledge distillation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2011–2020, 2024g.

Wenhua Zhang, Wenjing Deng, Zhen Cui, Jia Liu, and Licheng Jiao. Object knowledge distillation for joint detection and tracking in satellite videos. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024h.

Wentao Zhang, Xupeng Miao, Yingxia Shao, Jiawei Jiang, Lei Chen, Olivier Ruas, and Bin Cui. Reliable data distillation on graph convolutional network. In *Proceedings of the 2020 ACM SIGMOD international conference on management of data*, pp. 1399–1414, 2020a.

Wentao Zhang, Yuezihan Jiang, Yang Li, Zeang Sheng, Yu Shen, Xupeng Miao, Liang Wang, Zhi Yang, and Bin Cui. Rod: reception-aware online distillation for sparse graphs. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2232–2242, 2021e.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, 2018.

Xiaobing Zhang, Shijian Lu, Haigang Gong, Zhipeng Luo, and Ming Liu. Amln: adversarial-based mutual learning network for online knowledge distillation. In *European Conference on Computer Vision*, pp. 158–173. Springer, 2020b.

Yachao Zhang, Yanyun Qu, Yuan Xie, Zonghao Li, Shanshan Zheng, and Cuihua Li. Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15520–15528, 2021f.

Yiman Zhang, Hanting Chen, Xinghao Chen, Yiping Deng, Chunjing Xu, and Yunhe Wang. Data-free knowledge distillation for image super-resolution. In *CVPR*, pp. 7852–7861. Computer Vision Foundation / IEEE, 2021g.

Yuan Zhang, Tao Huang, Jiaming Liu, Tao Jiang, Kuan Cheng, and Shanghang Zhang. Freekd: Knowledge distillation via semantic frequency prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15931–15940, 2024i.

Yuehan Zhang, Seungjun Lee, and Angela Yao. Pairwise distance distillation for unsupervised real-world image super-resolution. In *European Conference on Computer Vision (ECCV)*, volume 15087 of *Lecture Notes in Computer Science*, pp. 429–446. Springer, 2024j.

Yuxuan Zhang, Lei Liu, and Li Liu. Cuing without sharing: A federated cued speech recognition framework via mutual knowledge distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8781–8789, 2023j.

Zhengbo Zhang, Chunluan Zhou, and Zhigang Tu. Distilling inter-class distance for semantic segmentation. *arXiv preprint arXiv:2205.03650*, 2022d.

Zhilu Zhang and Mert Sabuncu. Self-distillation as instance-specific label smoothing. *Advances in Neural Information Processing Systems*, 33:2184–2195, 2020.

Zhuoyang Zhang, Yuhao Dong, Yunze Liu, and Li Yi. Complete-to-partial 4d distillation for self-supervised point cloud sequence representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17661–17670, 2023k.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11953–11962, 2022a.

Borui Zhao, Renjie Song, and Jiajun Liang. Cumulative spatial knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6146–6155, 2023a.

Chaoqiang Zhao, Matteo Poggi, Fabio Tosi, Lei Zhou, Qiyu Sun, Yang Tang, and Stefano Mattoccia. Gasmono: Geometry-aided self-supervised monocular depth estimation for indoor scenes. In *ICCV*, pp. 16163–16174. IEEE, 2023b.

Haimei Zhao, Qiming Zhang, Shanshan Zhao, Zhe Chen, Jing Zhang, and Dacheng Tao. Simdistill: Simulated multi-modal distillation for bev 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7460–7468, 2024a.

Haoran Zhao, Xin Sun, Junyu Dong, Milos Manic, Huiyu Zhou, and Hui Yu. Dual discriminator adversarial distillation for data-free model compression. *International Journal of Machine Learning and Cybernetics*, pp. 1–18, 2022b.

Kaiqi Zhao, Hieu Duy Nguyen, Animesh Jain, Nathan Susanj, Athanasios Mouchtaris, Lokesh Gupta, and Ming Zhao. Knowledge distillation via module replacing for automatic speech recognition with recurrent neural network transducer. In *23rd Interspeech Conference*, 2022c.

Lingjun Zhao, Jingyu Song, and Katherine A Skinner. Crkd: Enhanced camera-radar object detection with cross-modality knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15470–15480, 2024b.

Peisen Zhao, Lingxi Xie, Jiajie Wang, Ya Zhang, and Qi Tian. Progressive privileged knowledge distillation for online action detection. *Pattern Recognition*, 129:108741, 2022d.

Shiji Zhao, Jie Yu, Zhenlong Sun, Bo Zhang, and Xingxing Wei. Enhanced accuracy and robustness via multi-teacher adversarial distillation. In *European Conference on Computer Vision*, pp. 585–602. Springer, 2022e.

Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *European conference on computer vision*, pp. 159–175. Springer, 2022f.

Yunpeng Zhao and Jie Zhang. Does training with synthetic data truly protect privacy? *arXiv preprint arXiv:2502.12976*, 2025.

Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching. *arXiv preprint arXiv:2402.11148*, 2024.

Qiang Zheng, Chao Zhang, and Jian Sun. Efficient point cloud classification via offline distillation framework and negative-weight self-distillation technique. *arXiv preprint arXiv:2409.02020*, 2024a.

Wenqing Zheng, Edward W Huang, Nikhil Rao, Sumeet Katariya, Zhangyang Wang, and Karthik Subbian. Cold brew: Distilling graph node representations with incomplete or missing neighborhoods. *arXiv preprint arXiv:2111.04840*, 2021.

Xu Zheng, Yunhao Luo, Pengyuan Zhou, and Lin Wang. Distilling efficient vision transformers from cnns for semantic segmentation. *Pattern Recognition*, 158:111029, 2025.

Yujie Zheng, Chong Wang, Chenchen Tao, Sunqi Lin, Jiangbo Qian, and Jiafei Wu. Restructuring the teacher and student in self-distillation. *IEEE Transactions on Image Processing*, 2024b.

Zhaohui Zheng, Rongguang Ye, Qibin Hou, Dongwei Ren, Ping Wang, Wangmeng Zuo, and Ming-Ming Cheng. Localization distillation for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10070–10083, 2023.

Lanfeng Zhong, Xin Liao, Shaoting Zhang, and Guotai Wang. Semi-supervised pathological image segmentation via cross distillation of multiple attentions. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 14225 of *Lecture Notes in Computer Science*, pp. 570–579. Springer, 2023.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pp. 696–712. Springer, 2022a.

Huajun Zhou, Bo Qiao, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Texture-guided saliency distilling for unsupervised salient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7257–7267. IEEE, 2023a.

Jingtao Zhou, Hao Zheng, Wenkai Zhong, and Zhiqiang Bao. Improving knowledge distillation via cross-modal insights from clip. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022b.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022c.

Kaiyue Zhou, Ming Dong, Peiyuan Zhi, and Shengjin Wang. Cascaded network with hierarchical self-distillation for sparse point cloud classification. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024a.

Kaiyue Zhou, Zelong Tan, Shu Yang, and Shengjin Wang. Enhancing the encoding process in point cloud completion. In *Proceedings of the 2024 13th International Conference on Computing and Pattern Recognition*, pp. 59–65, 2024b.

Sheng Zhou, Yucheng Wang, Defang Chen, Jiawei Chen, Xin Wang, Can Wang, and Jiajun Bu. Distilling holistic knowledge with graph neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10387–10396, 2021a.

Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, and Chao Ma. Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5116–5125, 2023b.

Wangchunshu Zhou, Canwen Xu, and Julian McAuley. Bert learns to teach: Knowledge distillation with meta learning. *arXiv preprint arXiv:2106.04570*, 2021b.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*, 2023c.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pp. 350–368. Springer, 2022d.

Yuhang Zhou, Siyuan Du, Haolin Li, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Reprogramming distillation for medical foundation models. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 15011 of *Lecture Notes in Computer Science*, pp. 533–543. Springer, 2024c.

Yuhang Zhou, Yushu Zhang, Leo Yu Zhang, and Zhongyun Hua. Derd: data-free adversarial robustness distillation through self-adversarial teacher group. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10055–10064, 2024d.

Zaida Zhou, Chaoran Zhuge, Xinwei Guan, and Wen Liu. Channel distillation: Channel-wise attention for knowledge distillation. *arXiv preprint arXiv:2006.01683*, 2020.

Zhengming Zhou and Qiulei Dong. Self-distilled feature aggregation for self-supervised monocular depth estimation. In *ECCV*, volume 13661 of *Lecture Notes in Computer Science*, pp. 709–726. Springer, 2022.

Zhenyu Zhou, Defang Chen, Can Wang, Chun Chen, and Siwei Lyu. Simple and fast distillation of diffusion models. *Advances in Neural Information Processing Systems*, 37:40831–40860, 2024e.

Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11175–11185, 2023d.

Jianing Zhu, Jiangchao Yao, Bo Han, Jingfeng Zhang, Tongliang Liu, Gang Niu, Jingren Zhou, Jianliang Xu, and Hongxia Yang. Reliable adversarial distillation with unreliable teachers. *arXiv preprint arXiv:2106.04928*, 2021a.

Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Mingzhe Rong, Aijun Yang, and Xiaohua Wang. Complementary relation contrastive distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9260–9269, 2021b.

Jinjing Zhu, Yunhao Luo, Xu Zheng, Hao Wang, and Lin Wang. A good student is cooperative and reliable: Cnn-transformer collaborative learning for semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11686–11696. IEEE, 2023a.

Lianghui Zhu, Xinggang Wang, Jiapei Feng, Tianheng Cheng, Yingyue Li, Bo Jiang, Dingwen Zhang, and Junwei Han. Weakclip: Adapting clip for weakly-supervised semantic segmentation. *International Journal of Computer Vision*, pp. 1–21, 2024a.

Wencheng Zhu, Xin Zhou, Pengfei Zhu, Yu Wang, and Qinghua Hu. Ckd: Contrastive knowledge distillation from a sample-wise perspective. *IEEE Transactions on Image Processing*, 2025.

Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. *Advances in neural information processing systems*, 31, 2018.

Xuekai Zhu, Biqing Qi, Kaiyan Zhang, Xingwei Long, and Bowen Zhou. Pad: Program-aided distillation specializes large models in reasoning. *arXiv preprint arXiv:2305.13888*, 2023b.

Yichen Zhu, Qiqi Zhou, Ning Liu, Zhiyuan Xu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Scalekd: Distilling scale-aware knowledge in small object detector. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19723–19733. IEEE, 2023c.

Yuchang Zhu, Jintang Li, Liang Chen, and Zibin Zheng. The devil is in the data: Learning fair graph neural networks via partial knowledge distillation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 1012–1021, 2024b.

Xianwei Zhuang, Zhihong Zhu, Zhichang Wang, Xuxin Cheng, and Yuexian Zou. Unicott: A unified framework for structural chain-of-thought distillation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16443–16452, 2021.