Simplifying Knowledge Transfer in Pretrained Models

Anonymous authors
Paper under double-blind review

Abstract

Pretrained models are ubiquitous in the current deep learning landscape, offering strong results on a broad range of tasks. Recent works have shown that models differing in various design choices exhibit categorically diverse generalization behavior, resulting in one model grasping distinct data-specific insights unavailable to the other. In this paper, we propose to leverage large publicly available model repositories as an auxiliary source of model improvements. We introduce a data partitioning strategy where pretrained models autonomously adopt either the role of a student, seeking knowledge, or that of a teacher, imparting knowledge, fostering a collaborative learning environment. Experiments across various tasks demonstrate the effectiveness of our proposed approach. In image classification, we improved the performance of ViT-B by approximately 1.4% through bidirectional knowledge transfer with ViT-T. For semantic segmentation, our method boosted all evaluation metrics by enabling knowledge transfer both within and across backbone architectures. In video saliency prediction, our approach achieved a new state-of-the-art. We further extend our approach to knowledge transfer between multiple models, leading to considerable performance improvements for all model participants.

1 Introduction

Knowledge Distillation (KD) (Buciluǎ et al., 2006; Hinton et al., 2015; Beyer et al., 2022) intends to transfer knowledge from a large 'teacher' model to a smaller 'student' model. Traditional KD methods utilize the predictions from the pretrained teacher model to supervise the training of the student model, encouraging it to generalize better than if it were trained from scratch alone. However, vanilla KD is a two-stage process that begins with training a teacher model and then freezing it to distill knowledge into the student model, meaning that the knowledge can only be transferred from the teacher to the student. Online Knowledge Distillation methods (Zhang et al., 2018; Guo et al., 2020) overcome this limitation by adopting a one-stage training process, jointly training a set of student models that learn from each other in a peer-teaching manner.

Although online KD methods employ a single stage training process, they distill knowledge into untrained student models. Transferring knowledge in this way neglects the existence of complementary knowledge between pretrained models, a factor that, if considered, can enhance generalization (Gontijo-Lopes et al., 2022; Roth et al., 2024). Various design choices such as hyperparameters, model architecture, optimization strategies, and pretraining dataset shape the semantic knowledge a model acquires (Bouthillier et al., 2021; Wagner et al., 2022). Gontijo-Lopes et al. (2022) show that model pairs with diverging training methodologies produce increasingly uncorrelated errors. Even low-accuracy models may capture some data-specific insights that high-accuracy models might overlook. Roth et al. (2024) capitalize on this finding by transferring complementary knowledge between any pretrained teacher and student model pair. They propose a data partitioning strategy that divides the dataset into instances where knowledge transfer from a teacher is beneficial and those where retaining the student's behavior is preferred. However, their approach operates in a unidirectional manner, wherein the teacher remains fixed after pretraining and the student is trained to minimize the teacher-student gap. This contrasts with the real-world teacher-student dynamic, in which a teacher continuously improves their knowledge and teaching skills through ongoing interactions with the student (Cornelius-White, 2007; Wright, 2011).

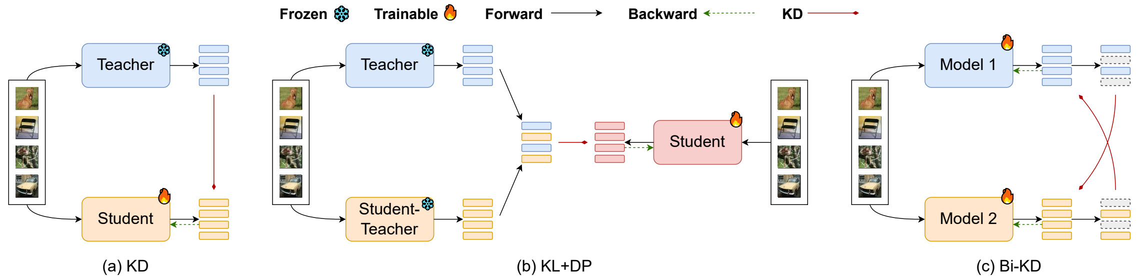

Figure 1: The figure illustrates the training process for a given batch of samples. (a) KD (Hinton et al., 2015) utilizes a pretrained teacher model and trains a student model to mimic the teachers predictions on each sample. (b) KL+DP (Roth et al., 2024) employs a frozen teacher and the original frozen student (called student-teacher) to jointly guide the training of the student model. They partition the dataset into samples where learning from the teacher is desired and ones where knowledge from the original frozen student should be retained. (c) In Bi-KD (ours), both models are trainable and learn from each other on every sample, resulting in bidirectional knowledge transfer. Different from KD and KL+DP, we are able to improve both models simultaneously.

Transferring knowledge among a group of pretrained models lets each one benefit from the unique strengths and insights that the others have already developed. They can complement each other's weaknesses and build more robust, generalized representations. When models are trained together, they iteratively update their predictions, effectively "teaching" each other. This simultaneous updating can lead to a performance boost that none of the models might achieve if trained independently (Zhang et al., 2018; Zhu et al., 2018; Guo et al., 2020).

To this end, we propose a simple method for parallel multidirectional knowledge transfer between pretrained models, a challenging task due to their prior training. Building on the data partition strategy introduced by Roth et al. (2024), we dynamically assign the teacher role to the model with the highest prediction confidence for the ground-truth class, allowing it to transfer knowledge to the other model. Unlike their fixed partitioning approach, our confidence-based data partitioning evolves as the models improve, adapting throughout the training process (as illustrated in Figure 1). Despite the apparent simplicity of our method, it leads to consistent performance improvements for all model participants within a single training stage.

We validated our approach through extensive experiments on models with diverse architectures, performance levels, sizes, and training objectives. Our method was tested across multiple tasks, including image classification, semantic segmentation, and video saliency prediction, and was further extended to enable parallel knowledge transfer among multiple models. In all cases, we observed consistent performance improvements as more models were added to the collaborative training environment.

Overall, we make the following contributions:

- We demonstrate the ability for bidirectional knowledge transfer in pretrained models. Specifically, we show that knowledge can be transferred across both models simultaneously.
- We provide experiments across ImageNet classification, semantic segmentation, and video saliency prediction, where we observe consistent improvements for all participating models. In particular, our method sets a new state-of-the-art in video saliency prediction.
- We establish that our framework seamlessly extends to concurrent learning across multiple models, thereby progressively enhancing the performance of each individual model as additional models are integrated.

2 Related Work

Knowledge Distillation (KD), pioneered by Buciluǎ et al. (2006), aimed to compress large teacher models into smaller student models by aligning their soft target distributions. Hinton et al. (2015) refined this approach by incorporating temperature scaling to minimize the difference between the softened class probabilities of the teacher and student models. Beyer et al. (2022) further highlighted the importance of consistent image augmentations and extended training schedules for effective KD.

Building on these ideas, recent works have explored transferring knowledge beyond just output probabilities. Romero et al. (2014) proposed aligning the intermediate feature representations between the teacher and the student models, while Zagoruyko & Komodakis (2017) train a student to imitate the attention maps of teacher networks. Park et al. (2019) introduced a method that preserves the structural relationship between the outputs using distance-wise and angle-wise losses. However, these methods often require careful layer selection and loss balancing (Yun et al., 2019), making them highly dependent on specific network architectures. To address this, Srinivas & Fleuret (2018) proposed matching the Jacobian of network outputs, which has a dimension independent of the model's architecture.

The work most similar to ours, proposed by Roth et al. (2024), introduced a data partitioning strategy for knowledge transfer among pretrained models. However, their approach uses a fixed data partitioning strategy during training, and the teacher model is not optimized specifically to guide the student. In contrast, our work introduces an evolving data partitioning strategy, which enables continuous improvements in the models through repeated interactions.

Online Knowledge Distillation treats every network as a student and trains them simultaneously from scratch. DML (Zhang et al., 2018) enables peer student models to learn from each other's predictions using a combination of cross-entropy and distillation losses. Anil et al. (2018) extend this concept to large-scale distributed neural networks, accelerating training by updating models concurrently.

Other approaches (Song & Chai, 2018; Zhu et al., 2018) involve designing multiple branch classifiers that are trained together. However, these methods are inflexible as they force networks to share lower layers, restricting knowledge transfer to only the upper layers within a single model. Chen et al. (2020) incorporate a self-attention mechanism to assess the similarity between network groups, enhancing peer diversity to create a more effective leader. Similarly, KDCL (Guo et al., 2020) introduces an ensemble of logits, where the optimal weight distribution is determined using a Lagrange multiplier to minimize generalization error.

More recently, Wu & Gong (2021) introduce an extra temporal mean network for each peer, assigning it the teacher role. Li & Jin (2022) propose a proxy teacher that updates its weights based on predictions from the original teacher model, enabling bidirectional distillation with the student model. While these methods optimize the teacher model for distillation, they transfer knowledge between untrained models, disregarding the presence of complementary knowledge between pretrained models. Livanos et al. (2024) address this by transferring knowledge between trained models. They dynamically assign the teacher role to a model that correctly classifies an instance while others fail. The teacher then generates a counterfactual instance for each correct prediction, adding it to the training set of incorrect models, which are subsequently retrained. However, this approach requires multiple training stages, making it computationally inefficient.

In contrast, our proposed method improves every pretrained model involved in KD within a single training stage, leading to a more efficient and effective learning process.

Multi-teacher Knowledge Distillation. KD can naturally be extended to learning from multiple pretrained teachers. Fukuda et al. (2017) combine the distillation framework with a data augmentation strategy by creating multiple copies of the data with the corresponding soft output targets from multiple teachers. You et al. (2017) further extend this approach by incorporating multiple teacher networks in the intermediate layers, considering the dissimilarity between intermediate representations of different examples. Luo et al. (2019) propose a common feature learning scheme, in which the features of all teachers are transformed into a common space, and the student is required to imitate them all to amalgamate the intact knowledge. Instead of treating all teacher models equally, some works (Liu et al., 2020; Yuan et al., 2021) dynamically

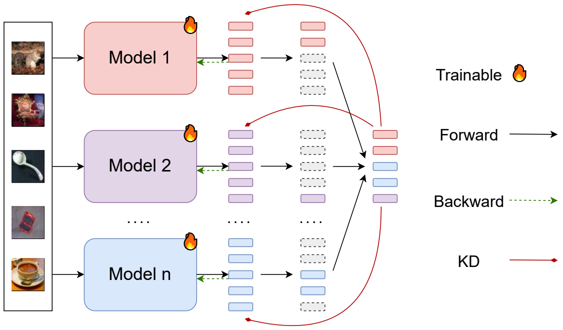

Figure 2: Generalization of our proposed method for transferring knowledge between multiple pretrained models. For each sample, we select a teacher model that guides the training of all other models. The teacher is selected based on the highest prediction probability, or the lowest loss, corresponding to the ground truth. This data partitioning strategy enables every model participating in the knowledge transfer to learn from each others strengths and address their weaknesses, resulting in all models improving within a single training stage.

assign weights to teacher models for different training instances and optimize the performance of the student model.

Although these works utilize predictions from multiple teacher models to reduce variance in network outputs, they fail to optimize teacher networks by considering the complementary knowledge between them. Our proposed method allows every model to benefit from each other's strengths and complement their weaknesses. This results in consistent performance improvements for all models in a single training stage, which ultimately leads to more robust predictions.

3 Bidirectional Knowledge Transfer

In this section, we first explain the traditional KD approach in Section 3.1, then introduce our proposed method for bidirectional knowledge transfer between two pretrained models in Section 3.2. Section 3.3 explains the extension of our method to dense tasks, namely semantic segmentation, and video saliency prediction, and finally, Section 3.4 highlights our approach for multidirectional knowledge transfer among multiple pretrained models.

3.1 Preliminaries

KD aims to improve the performance of the student network by leveraging the predictions of a teacher network as supervision. Hinton et al. (2015) propose minimizing the Kullback-Leibler (KL) divergence between the soft targets of the teacher and student models. The distillation loss is formulated as:

$$\mathcal{L}_{KL_{1,2}} = \frac{T^2}{N} \sum_{i=1}^{N} \mathrm{KL}\left[\sigma(\mathbf{z}_{1,i}/T), \sigma(\mathbf{z}_{2,i}/T)\right] \tag{1}$$

where T represents the temperature parameter, N denotes the batch size, and $\sigma(\mathbf{z}_1)$ and $\sigma(\mathbf{z}_2)$ correspond to class probabilities of student and teacher predictions, respectively. We use Equation 1 along with task-specific loss as our overall loss function.

3.2 Formulation

Given a training data batch $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$ of N samples, where $\mathcal{X} = \{x_i\}_{i=1}^{N}$ and $\mathcal{Y} = \{y_i\}_{i=1}^{N}$ represent the inputs and corresponding labels for C classes, we consider two models f_1 and f_2 parameterized by θ_1 and θ_2

```
Algorithm 1: Bi-KD
Input: Training set $\mathcal{X}$, label set $\mathcal{Y}$, learning rate $\eta$, epochs $T_{\max}$, iterations $N_{\max}$, models $f_1$ and $f_2$
       parameterized by $\theta_1$ and $\theta_2$ respectively
for $T = 1$ to $T_{\max}$ do
    Shuffle training set $\mathcal{X}$;
    for $N = 1$ to $N_{\max}$ do
        Fetch mini-batch $x$ from $\mathcal{X}$;
        Compute output logits $\mathbf{z}_1 = f_1(x; \theta_1)$ and $\mathbf{z}_2 = f_2(x; \theta_2)$;
        Obtain data mask $m_1 = \mathbb{I}\big[\sigma(\mathbf{z}_1)_{gt} > \sigma(\mathbf{z}_2)_{gt}\big]$ ;      // samples where $f_1$ acts as a teacher
        Obtain data mask $m_2 = \mathbb{I}\big[\sigma(\mathbf{z}_1)_{gt} \leq \sigma(\mathbf{z}_2)_{gt}\big]$ ;      // samples where $f_2$ acts as a teacher
        Get the distillation loss $\mathcal{L}_{\text{dist}} = m_1 \cdot \mathcal{L}_{KL_{2,1}} + m_2 \cdot \mathcal{L}_{KL_{1,2}}$;
        Get the cross-entropy losses $\mathcal{L}_{\text{ce}_1} = \text{CE}(f_1(x; \theta_1), \mathcal{Y})$ and $\mathcal{L}_{\text{ce}_2} = \text{CE}(f_2(x; \theta_2), \mathcal{Y})$;
        Compute overall loss $\mathcal{L} = \mathcal{L}_{\text{ce}_1} + \mathcal{L}_{\text{ce}_2} + \mathcal{L}_{\text{dist}}$;
        Update $\theta_1 \leftarrow \theta_1 - \eta \frac{\partial \mathcal{L}}{\partial \theta_1}$;
        Update $\theta_2 \leftarrow \theta_2 - \eta \frac{\partial \mathcal{L}}{\partial \theta_2}$;
Output: Trained models $f_1(\theta_1)$ and $f_2(\theta_2)$
```

respectively. Their output logits are computed as follows:

$$\mathbf{z}_1 = f_1(\mathcal{X}; \theta_1), \quad \mathbf{z}_2 = f_2(\mathcal{X}; \theta_2) \tag{2}$$

For each sample, we dynamically assign the teacher role to the model with the highest prediction probability for the corresponding ground-truth class, resulting in data masks m_1 and m_2 denoting models f_1 and f_2 as teachers respectively:

$$m_1 = \mathbb{I}\left[\sigma(\mathbf{z}_1)_{gt} > \sigma(\mathbf{z}_2)_{gt}\right], m_2 = \mathbb{I}\left[\sigma(\mathbf{z}_1)_{gt} \leq \sigma(\mathbf{z}_2)_{gt}\right] \tag{3}$$

Here, $\sigma(\cdot)_{gt}$ denotes the softmax probability for the ground-truth class, and $\mathbb{I}$ represents the indicator function. This approach partitions the data into samples where the best performing model provides supervision to others while considering continuously improving models. The distillation loss is given as:

$$\mathcal{L}_{\text{dist}} = m_1 \cdot \mathcal{L}_{KL_{2,1}} + m_2 \cdot \mathcal{L}_{KL_{1,2}} \tag{4}$$

We also use a task-specific loss, cross-entropy for image classification, $\mathcal{L}_{\text{T}}$ for every model along with the distillation loss, resulting in the overall loss function:

$$\mathcal{L} = \mathcal{L}_{\text{T}_1} + \mathcal{L}_{\text{T}_2} + \mathcal{L}_{\text{dist}} \tag{5}$$

3.3 Knowledge Transfer for dense tasks

We further extend our approach for semantic segmentation and video saliency prediction. Instead of the aforementioned confidence-based data partition, we assign the teacher role to the model having the lowest loss for each sample. Let $\mathcal{L}_T$ be the task-specific loss, the data masks are then formulated as:

$$m_1 = \mathbb{I}\left[\mathcal{L}_{T_1} < \mathcal{L}_{T_2}\right], \quad m_2 = \mathbb{I}\left[\mathcal{L}_{T_1} \geq \mathcal{L}_{T_2}\right] \tag{6}$$

these masks are utilized in Equation 4 to obtain the distillation loss $\mathcal{L}_{\text{dist}}$, and the overall loss function is the same as in Equation 5.

Semantic Segmentation: For experiments on semantic segmentation, we use binary cross-entropy loss and dice loss (Milletari et al., 2016) for our mask loss:

$$\mathcal{L}_{\text{mask}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} \tag{7}$$

where $\lambda_{\text{ce}} = 5.0$ and $\lambda_{\text{dice}} = 5.0$. The final task-specific loss for semantic segmentation is a combination of mask loss and classification loss:

$$\mathcal{L}_{\text{semseg}} = \mathcal{L}_{\text{mask}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} \tag{8}$$

with $\lambda_{\text{cls}} = 2.0$ for correct predictions and 0.1 for incorrect ones.

Saliency Prediction: Video saliency prediction utilizes a combination of Equation 1 and Correlation Coefficient (CC), which calculates the Pearson correlation between the ground-truth and the predicted saliency maps, as the final task-specific loss:

$$\mathcal{L}_{\text{VSP}} = \mathcal{L}_{\text{KL}} - \mathcal{L}_{\text{CC}} \tag{9}$$

$$\mathcal{L}_{\text{CC}} = \frac{\sigma(P, Q)}{\sigma(P, P) \times \sigma(Q, Q)} \tag{10}$$

here P & Q are the predicted saliency map and ground-truth respectively, and $\sigma(P, Q)$ represents the covariance between P and Q.

3.4 Knowledge Transfer between multiple models

With the basic knowledge transfer between two models set up, extending it to parallel knowledge transfer between multiple models is the logical next step. As illustrated in Figure 2, the model with the highest prediction probability corresponding to the ground-truth class is selected as the teacher for a particular sample. Given K models $f_0, f_1, \ldots, f_{K-1}$ parameterized by $\theta_0, \theta_1, \ldots, \theta_{K-1}$ respectively. The data mask for a model k is computed as:

$$m_k = \mathbb{I}\left[\arg\max_k \left([\sigma(\mathbf{z}_0)_{gt}, \ldots, \sigma(\mathbf{z}_{K-1})_{gt}]\right) == k\right] \tag{11}$$

where m_k, with $k \in \{0, 1, \ldots, K-1\}$, represents whether the model f_k acts as a teacher for a particular sample, and $[\sigma(\mathbf{z}_0)_{gt}, \ldots, \sigma(\mathbf{z}_{K-1})_{gt}]$ denotes the concatenation of prediction probabilities corresponding to the ground-truth class for K models. The distillation loss is formulated as:

$$\mathcal{L}_{\text{dist}} = \sum_{i=0}^{K-1} \sum_{j=0, j\neq i}^{K-1} m_i \cdot \mathcal{L}_{KL_{j,i}} \tag{12}$$

with $\mathcal{L}_{KL_{j,i}}$ considering models f_j and f_i as student and teacher respectively. The overall loss is given as:

$$\mathcal{L} = \sum_{i=0}^{K-1} \mathcal{L}_{T_i} + \mathcal{L}_{\text{dist}} \tag{13}$$

Table 1: Selection of models used for experiments on the validation set of ImageNet.

Models	**Acc.**	**# Params. (M)**
SeNet154 (He et al., 2019)	81.378	115.09
SWSL-ResNext101 (Xie et al., 2017)	84.276	88.79
MAE (He et al., 2022)	83.446	86.57
ViT-B (Dosovitskiy et al., 2021)	79.152	86.57
PiT-B (Heo et al., 2021)	82.278	73.76
ResMLP-36 (Touvron et al., 2022)	79.576	44.69
MambaVision-T2 (Hatamizadeh & Kautz, 2024)	82.506	35.1
ResMLP-24-dist (Touvron et al., 2022)	80.548	30.02
DINOv2 (Oquab et al., 2023)	81.332	23.98
ViT-S (Dosovitskiy et al., 2021)	78.842	22.05
CoaT-lite-mini (Xu et al., 2021)	78.858	11.01
PiT-XS (Heo et al., 2021)	77.916	10.62
ViT-T (Dosovitskiy et al., 2021)	75.466	5.72

4 Experiments & Results

In this section, we perform a series of experiments to evaluate our proposed method on image classification, semantic segmentation, and video saliency prediction benchmarks. Section 4.1 introduces the datasets used for various experiments, and Section 4.2 explains the followed training choices. Finally, we discuss our results on various tasks in Section 4.3.

4.1 Datasets

We verify the effectiveness of our approach on multiple tasks using the following datasets:

ImageNet (Deng et al., 2009) consists of 1.2 million images for training and 50,000 images for validation. We report the results of our knowledge transfer between two or multiple models on the validation set.

ADE20K (Zhou et al., 2017) provides 150 object and stuff categories, with 20,210 images in the training set and 2,000 images in the validation set. We use the validation set to evaluate our approach for knowledge transfer on semantic segmentation.

DHF1K (Wang et al., 2018) is a benchmark dataset for video saliency prediction, comprising 600 videos in the training set and 100 videos in the validation set. We use the validation set for our evaluation.

Hollywood-2 (Mathe & Sminchisescu, 2014) is the largest dataset for video saliency prediction in terms of the number of videos, containing 1,707 clips sourced from 69 Hollywood movies. Following the standard evaluation protocol, we use the predefined split of 823 videos for training and the remaining 884 videos for testing.

4.2 Implementation details

We implement all the networks and training procedures in Pytorch (Paszke et al., 2019), and conduct all experiments on a single NVIDIA RTX A6000. We use the Adam optimizer for image classification and video saliency prediction, while experiments on semantic segmentation utilize the AdamW optimizer. For all experiments, the learning rate and weight decay are set to $1e$-6 and $1e$-5 respectively, with the temperature parameter set to 1 in Equation 1.

For experiments on ImageNet, we compare our approach with Roth et al. (2024) and follow their experimental setup. We use large open model libraries like *timm* (Wightman et al., 2019) and *huggingface* for our experiments. Furthermore, for knowledge transfer between semantic segmentation models, we utilize Mask2Former (Cheng et al., 2022) and follow its original experimental setup.

Finally, we perform knowledge transfer between state-of-the-art video saliency models proposed by Zhou et al. (2023) and Girmaji et al. (2025). We adopt their respective experimental setups to train the models from scratch, before transferring knowledge between them.

Table 2: Comparative results of change in Top-1 accuracy after transferring knowledge between models pretrained on ImageNet using different methods

Method	Model 1	$\Delta_{\text{top-1}}$	Model 2	$\Delta_{\text{top-1}}$
KL+DP Ours	DINOv2	0.582 **1.39**	MAE	0.306 **0.396**
KL+DP Ours	PiT-B	0.73 **0.822**	SWSL-ResNext101	**0.336** 0.23
KL+DP Ours	DINOv2	0.968 **1.472**	MambaVision-T2	-0.242 **0.036**
KL+DP Ours	DINOv2	0.818 **1.57**	SWSL-ResNext101	**0.538** 0.474
KL+DP Ours	CoaT-lite-mini	0.436 **0.482**	SeNet154	**0.48** 0.456
KL+DP Ours	CoaT-lite-mini	0.386 **0.584**	DINOv2	0.692 **1.364**
KL+DP Ours	PiT-XS	0.37 **0.472**	ResMLP-36	0.086 **0.274**
KL+DP Ours	CoaT-lite-mini	0.17 **0.38**	PiT-XS	0.222 **0.45**
KL+DP Ours	ViT-B	0.614 **1.174**	ViT-S	0.518 **0.684**
KL+DP Ours	ViT-B	0.492 **1.392**	ViT-T	0.828 **0.898**
KL+DP Ours	ViT-S	0.336 **0.838**	ViT-T	0.724 **0.946**

4.3 Results

Image Classification. For evaluating our approach on ImageNet, we utilize pretrained models listed in Table 1. The models were chosen to cover a wide range of architectures, performance levels, sizes, and training objectives. In Table 2, we report $\Delta_{\text{top-1}}$, which represents the change in Top-1 accuracy after transferring knowledge between models. Since Roth et al. (2024) have demonstrated that standard KD can negatively impact performance when learning from weaker or similarly performing teacher models, we compare our proposed method with their approach, referred to as KL+DP. It is important to note that, although our approach updates both models simultaneously in a single pass, applying KL+DP requires two separate runs, alternating the role of each model as the student, thereby incurring twice the computational cost and time during training.

From the results presented in Table 2, we observe that Bi-KD consistently improves the performance of both participating models. This finding substantiates our hypothesis that pretrained models serve as effective sources for transferring complementary knowledge between one another, thereby enabling the enhancement of each model independently. Furthermore, our method demonstrates superior performance compared to KL+DP, outperforming it in 19 out of 22 cases. The results provide supporting evidence for our hypothesis, demonstrating that simultaneous, bidirectional knowledge transfer, enabled through Bi-KD is more effective than the unidirectional knowledge transfer employed by KL+DP, where the frozen teacher model is not optimized for teaching.

Table 3: Comparison of each model pair's performance before and after knowledge transfer using Bi-KD, along with the performance of their direct ensemble.

Model 1	Top-1		Model 2	Top-1		Ensemble	Recovered (%)	
	Before	After		Before	After		Model 1	Model 2
DINOv2	81.332	82.722	MAE	83.446	83.842	84.332	46.3	44.7
PiT-B	82.278	83.1	SWSL-ResNext101	84.276	84.506	85.206	28.1	24.7
DINOv2	81.332	82.804	Mamba Vision-T2	82.506	82.542	83.56	66	3.4
DINOv2	81.332	82.902	SWSL-ResNext101	84.276	84.75	85.162	40.1	53.5
CoaT-lite-mini	78.858	79.34	SeNet154	81.378	81.834	82.302	14	49.4
CoaT-lite-mini	78.858	79.442	DINOv2	81.332	82.694	82.732	15	97.3
PiT-XS	77.916	78.388	ResMLP-36	79.576	79.85	80.242	20.3	41.1
CoaT-lite-mini	78.858	79.238	PiT-XS	77.916	78.366	79.866	37.7	23.1
ViT-B	79.152	80.326	ViT-S	78.842	79.526	80.48	88.4	41.8
ViT-B	79.152	80.544	ViT-T	75.466	76.364	80.274	124	18.7
ViT-S	78.842	79.68	ViT-T	75.466	76.412	79.69	98.8	22.4

Our experiments further indicate that the most significant performance improvements occur when models with differing training methodologies are paired together. For example, pairing the self-supervised DINOv2 (Oquab et al., 2023) with any model trained using a supervised objective consistently results in performance improvements exceeding 1%. Interestingly, SWSL-ResNext101 (Xie et al., 2017), which has a Top-1 accuracy of 84.276%, benefits more when paired with the relatively weaker DINOv2 than with the stronger PiT-B (Heo et al., 2021). Similarly, CoaT-lite-mini (Xu et al., 2021) shows a greater improvement when paired with DINOv2 than with SeNet154 (He et al., 2019), despite both having comparable performance. These results align with the observation made by Gontijo-Lopes et al. (2022) that models trained through different methodologies tend to make uncorrelated errors, thereby making even lower performing models valuable contributors in knowledge transfer.

Another notable observation is that knowledge transfer between models with the same architecture also results in considerable performance improvements. All three models: ViT-B, ViT-S, and ViT-T, consistently yield performance improvements when paired with each other. Interestingly, both ViT-B and ViT-S experience greater performance gains when paired with the smaller ViT-T, rather than with each other. This further underscores that even smaller models can capture data-specific insights that may be absent in larger counterparts.

Finally, Table 3 presents a comparison between the performance of Bi-KD and a direct ensemble of its two constituent models. For the ensemble, the final classification is obtained by averaging the softmax scores of the individual models. To quantify this comparison, we calculate the percentage of the ensemble's performance retained by each model using the following formula:

$$Recovered = \frac{\text{Top-1}_{\text{after}} - \text{Top-1}_{\text{before}}}{\text{Ensemble} - \text{Top-1}_{\text{before}}} \tag{14}$$

Table 4: Performance comparison of Mask2Former models before and after knowledge transfer on ADE20K dataset.

Backbones	Before				After			
	mIoU	fwIoU	mACC	pACC	mIoU	fwIoU	mAcc	pACC
R50	47.23	70.95	60.11	81.72	47.69	71.12	60.43	81.83
Swin-T	47.7	72.37	61.44	82.7	48.32	72.66	61.76	82.99
Swin-S	51.33	73.26	65.11	83.48	51.67	73.35	65.3	83.56
Swin-T	47.7	72.37	61.44	82.7	48.48	72.52	62.07	82.83

As shown in Table 3, our method allows individual models to recover more than half of the ensemble's performance in most scenarios. Larger models, such as SeNet154, SWSL-ResNext101, MAE (He et al., 2022), ViT-B, and PiT-B, prove particularly effective, recovering approximately 60% of the ensemble performance on average. Among these, ViT-B stands out by recovering nearly 90% of the ensemble accuracy in one instance and surpassing the ensemble performance in another.

In contrast, smaller models, such as DINOv2, ViT-S, CoaT-lite-mini, PiT-XS (Heo et al., 2021), and ViT-T, recover roughly 40% of the ensemble's performance on average. In particular, DINOv2 and ViT-S perform consistently well, each recovering at least 40% of the ensemble performance in all cases and closely approaching it in some instances.

Overall, these findings underscore the effectiveness of our approach in enabling individual models, both large and small, to recover a noticeable portion of the ensemble-level performance.

Semantic Segmentation. Table 4 compares the performance of Mask2Former models with various backbones before and after applying our proposed knowledge transfer method. The models are evaluated using four standard semantic segmentation metrics: mean Intersection-over-union (mIoU), frequency weighted Intersection-over-union (fwIoU), mean Accuracy (mACC), and pixel Accuracy (pACC).

Our approach leads to consistent improvements in each metric within a single stage of training, whether transferring knowledge between different architectures or similar ones. Notably, the Swin-T (Liu et al., 2021) backbone benefits more when paired with the stronger Swin-S (Liu et al., 2021) backbone than with the similarly performing R50 (He et al., 2016) backbone. For instance, Swin-T's mIoU increases from 47.7 to 48.32 when paired with R50, but increases further to 48.48 when paired with Swin-S. While the improvements are not dramatic, their consistency demonstrates the utility of our proposed approach in extracting additional performance from already well-trained models.

Video Saliency Prediction. Table 5 presents a comparative evaluation of video saliency prediction task on the DHF1K and Hollywood-2 datasets using two standard metrics: CC and Normalized Scanpath Saliency (NSS). We perform knowledge transfer on the TMFI-Net (Zhou et al., 2023) and ViNet-A (Girmaji et al., 2025) models. The performance of their Bi-KD variants is compared against their original versions as well as other state-of-the-art methods.

On the DHF1K dataset, TMFI-Net (Bi-KD) establishes a new state-of-the-art, improving its CC from 0.552 to 0.558 and its NSS from 3.188 to 3.216. Similarly, ViNet-A (Bi-KD) demonstrates consistent gains, with its CC increasing from 0.525 to 0.536 and NSS from 3.019 to 3.077.

On the Hollywood-2 dataset, the performance gains are even more substantial. TMFI-Net (Bi-KD) improves from a CC of 0.737 to 0.750 and from an NSS of 4.054 to 4.148. ViNet-A (Bi-KD) achieves a new state-of-the-art, raising its CC from 0.756 to 0.762 and its NSS from 4.119 to 4.198.

These results validate the efficacy of our approach, demonstrating that mutual knowledge transfer consistently enhances performance across diverse architectures and datasets, thereby advancing the state-of-the-art in video saliency prediction.

Table 5: We apply Bi-KD between TMFI-Net and ViNet-A, and compare the resulting models against a range of individually trained saliency prediction methods. The last two rows represent the Bi-KD variants.

Models	DHF1K		Hollywood-2	
	CC	NSS	CC	NSS
ViNet (Jain et al., 2021)	0.521	2.957	0.693	3.73
TSFP-Net (Chang & Zhu, 2021)	0.529	3.009	0.711	3.91
STSA-Net (Wang et al., 2021)	0.539	3.082	0.705	3.908
TMFI-Net (Zhou et al., 2023)	0.552	3.188	0.737	4.054
THTD-Net (Moradi et al., 2024)	0.553	3.188	0.726	3.965
ViNet-S (Girmaji et al., 2025)	0.529	3.008	0.728	3.941
ViNet-A (Girmaji et al., 2025)	0.525	3.019	0.756	4.119
TMFI-Net (Zhou et al., 2023) (Bi-KD)	**0.558**	**3.216**	0.75	4.148
ViNet-A (Girmaji et al., 2025) (Bi-KD)	0.536	3.077	**0.762**	**4.198**

Table 6: Parallel multidirectional knowledge transfer across multiple models. For image classification, we report results for two-way, three-way, and four-way knowledge transfer using our proposed approach. For video saliency prediction, we present results for two-way and three-way knowledge transfer.

Models	$\Delta_{\text{top-1}}$
CoaT-lite-mini	0.38
PiT-XS	0.45
CoaT-lite-mini	0.46
PiT-XS	0.476
ResMLP-24-dist	0.15
CoaT-lite-mini	0.572
PiT-XS	0.63
ResMLP-24-dist	0.286
DINOv2	1.28

(a) Image classification

Models	DHF1K	
	CC	NSS
TMFI-Net	0.558	3.216
ViNet-A	0.536	3.077
TMFI-Net	**0.561**	**3.224**
ViNet-A	0.54	3.087
ViNet-S	0.533	3.038

(b) Video saliency prediction

Multi-directional Transfer. Finally, we extend our knowledge transfer framework to support parallel, multidirectional transfer among multiple models within a single training stage. Table 6 demonstrates the effectiveness of our parallel multidirectional knowledge transfer strategy across both image classification and video saliency prediction tasks.

In image classification on ImageNet, all participating models consistently benefit as more models are incorporated into the collaborative learning setup. For example, the performance gain for CoaT-lite-mini increases from 0.38 to 0.46, and PiT-XS improves from 0.45 to 0.476 when ResMLP-24-dist (Touvron et al., 2022) is added. These gains are further amplified, reaching 0.572 for CoaT-lite-mini, 0.63 for PiT-XS, and 0.286 for ResMLP-24-dist, with the inclusion of DINOv2. These results highlight the scalability and effectiveness of our approach with respect to the number of participating models.

Importantly, our approach is also task-agnostic. In video saliency prediction on DHF1K, TMFI-Net benefits from the knowledge transferred from ViNet-A, achieving a CC of 0.558 and an NSS of 3.216. Incorporating ViNet-S (Girmaji et al., 2025) into the collaborative learning environment further boosts the performance of both TMFI-Net and ViNet-A, with TMFI-Net achieving a new state-of-the-art. These findings affirm the robustness and generality of our knowledge transfer strategy across both architectures and tasks.

5 Conclusion

In this work, we introduce a simple yet effective approach for simultaneous multidirectional knowledge transfer between pretrained models Our method employs a dynamic data partitioning scheme that selects the most suitable teacher model for each sample, resulting in consistent performance improvements across all participating models within a single training stage. By enabling each model to serve as both a learner and a teacher, our framework fosters mutual enhancement and contributes to the development of more robust model ensembles. We demonstrate the effectiveness of our approach across a range of model architectures and tasks, including image classification, semantic segmentation, and video saliency prediction. Notably, our method sets a new state-of-the-art in video saliency prediction, underscoring the potential of collaborative knowledge transfer in complex visual understanding tasks. Additionally, we extend our framework to support knowledge transfer among multiple models and find that performance continues to improve as more models are added to the collaborative environment. Our results provide compelling evidence for the viability of simultaneous multidirectional knowledge transfer between pretrained models. Future work could explore model merging as a pathway to consolidate the strengths of multiple models into a single, better performing model.

References

Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E. Dahl, and Geoffrey E. Hinton. Large scale distributed neural network training through online distillation. In *International Conference on Learning Representations (ICLR)*, 2018. URL `https://openreview.net/forum?id=rkr1UDeC-`.

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10925–10934, 2022.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3:747–769, 2021.

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.

Qinyao Chang and Shiping Zhu. Temporal-spatial feature pyramid for video saliency detection. *arXiv preprint arXiv:2105.04213*, 2021.

Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 34, pp. 3430–3437, 2020.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1299, 2022.

Jeffrey Cornelius-White. Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of educational research*, 77(1):113–143, 2007.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

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.

Rohit Girmaji, Siddharth Jain, Bhav Beri, Sarthak Bansal, and Vineet Gandhi. Minimalistic Video Saliency Prediction via Efficient Decoder & Spatio Temporal Action Cues. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2025.

Raphael Gontijo-Lopes, Yann Dauphin, and Ekin Dogus Cubuk. No one representation to rule them all: Overlapping features of training methods. In *International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=BK-4qbGgIE3`.

Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11020–11029, 2020.

Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. *arXiv preprint arXiv:2407.08083*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 558–567, 2019.

Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11936–11945, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyamgopal Karthik, Ramanathan Subramanian, and Vineet Gandhi. Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3520–3527. IEEE, 2021.

Lujun Li and Zhe Jin. Shadow knowledge distillation: Bridging offline and online knowledge transfer. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:635–649, 2022.

Yuang Liu, Wei Zhang, and Jun Wang. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113, 2020.

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 *International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

Michael Livanos, Ian Davidson, and Stephen Wong. Cooperative knowledge distillation: A learner agnostic approach. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 38, pp. 14124–14131, 2024.

Sihui Luo, Xinchao Wang, Gongfan Fang, Yao Hu, Dapeng Tao, and Mingli Song. Knowledge amalgamation from heterogeneous networks by common feature learning. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

Stefan Mathe and Cristian Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 37(7):1408–1424, 2014.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

Morteza Moradi, Simone Palazzo, and Concetto Spampinato. Transformer-based video saliency prediction with high temporal dimension decoding. *arXiv preprint arXiv:2401.07942*, 2024.

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. *Transactions on Machine Learning Research (TMLR)*, 2023.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3967–3976, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *International Conference on Learning Representations (ICLR)*, 2014.

Karsten Roth, Lukas Thede, A. Sophia Koepke, Oriol Vinyals, Olivier J Henaff, and Zeynep Akata. Fantastic gains and where to find them: On the existence and prospect of general knowledge transfer between any pretrained model. In *International Conference on Learning Representations (ICLR)*, 2024. URL `https://openreview.net/forum?id=m50eKHCttz`.

Guocong Song and Wei Chai. Collaborative learning for deep neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Suraj Srinivas and François Fleuret. Knowledge transfer with jacobian matching. In *International Conference on Machine Learning (ICML)*, pp. 4723–4731. PMLR, 2018.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 45(4):5314–5321, 2022.

Diane Wagner, Fabio Ferreira, Danny Stoll, Robin Tibor Schirrmeister, Samuel Müller, and Frank Hutter. On the importance of hyperparameters and data augmentation for self-supervised learning. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. URL `https://openreview.net/forum?id=oBmAN382UL`.

Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4894–4903, 2018.

Ziqiang Wang, Zhi Liu, Gongyang Li, Yang Wang, Tianhong Zhang, Lihua Xu, and Jijun Wang. Spatio-temporal self-attention network for video saliency prediction. *IEEE Transactions on Multimedia*, 25: 1161–1174, 2021.

Ross Wightman et al. Pytorch image models, 2019.

Gloria Brown Wright. Student-centered learning in higher education. *International journal of teaching and learning in higher education*, 23(1):92–97, 2011.

Guile Wu and Shaogang Gong. Peer collaborative learning for online knowledge distillation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 35, pp. 10302–10310, 2021.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1492–1500, 2017.

Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9981–9990, 2021.

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. Association for Computing Machinery, 2017. ISBN 9781450348874. doi: 10.1145/3097983.3098135. URL https://doi.org/10.1145/3097983.3098135.

Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. Reinforced multi-teacher selection for knowledge distillation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 35, pp. 14284–14291, 2021.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, pp. 6023–6032, 2019.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Sks9_ajex.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4320–4328, 2018.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641, 2017.

Xiaofei Zhou, Songhe Wu, Ran Shi, Bolun Zheng, Shuai Wang, Haibing Yin, Jiyong Zhang, and Chenggang Yan. Transformer-based multi-scale feature integration network for video saliency prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7696–7707, 2023.

Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.