DETRDISTILL: A SIMPLE KNOWLEDGE DISTILLATION FRAMEWORK FOR DETR-FAMILIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformer-based detectors (DETRs) have attracted great attention due to their sparse training paradigm and the removal of post-processing operations, but the huge model can be computationally time-consuming and difficult to be deployed in real-world applications. To tackle this problem, knowledge distillation (KD) can be employed to compress the huge model by constructing a simple teacher-student learning framework. Different from the traditional CNN detectors, where the distillation targets can be naturally aligned through the feature map, DETR regards object detection as a set prediction problem, leading to an unclear relationship between teacher and student during distillation. In this paper, we propose **DETRDistill**, a novel knowledge distillation dedicated to DETR-families. We first explore a sparse matching paradigm with progressive stage-by-stage instance distillation. Considering the diverse attention mechanisms adopted in different DETRs, we propose attention-agnostic feature distillation module to overcome the ineffectiveness of conventional feature imitation. Finally, to fully leverage the intermediate products from the teacher, we introduce teacher-assisted assignment distillation, which greatly alleviates the instability of label assignment caused by bipartite graph matching. Extensive experiments demonstrate that our distillation method achieves significant improvement on various competitive DETR approaches, without introducing extra consumption in the inference phase. To the best of our knowledge, this is the first systematic study to explore a general distillation method for DETR-style detectors.

1 INTRODUCTION

Object detection aims to locate and classify objects of predefined categories from images. Early works usually employ convolutional neural networks (CNNs) and require post-processing procedures such as non-maximum suppression (NMS). Recently, transformer-based works propose to regard object detection as a set prediction task and train it end-to-end. DETR (Carion et al., 2020) firstly applied transformer in the detection field and eliminates the need for such hand-crafted post-process components in the CNN-based detector.

Although creating a novel paradigm of objection detection, it still suffers from low inference-speed performance and slow training convergence. To alleviate the problems, Deformable DETR (Zhu et al., 2020) designs a sparse attention mechanism to speed up the inference procedure; Conditional DETR (Meng et al., 2021) reconstructed object query to accelerate convergence to reduce the difficulty of training. However, the inference speed of the neural network can still be limited mainly by the model size. Generally, larger models can achieve higher performance but require more storage and slow down the inference speed. How to balance the efficiency and performance of the model is receiving increasing attention.

Knowledge distillation (Hinton et al., 2015) was first proposed and applied to decrease the complexity of the model while retaining most of the performance in the image classification task. The soft labels predicted by the teacher model provide similarity across categories (denoted as *logits*). With this "dark knowledge" as extra supervision, the student model can greatly improve performance. FitNet (Romero et al., 2014) allows the student to imitate features from the image backbone, providing supervision at an earlier stage. Most of the current distillation methods can be divided into logits-level and feature-level, inspired by the two works mentioned above. Recent work has shown that

imitating the whole feature equally is sub-optimal because of noise in the background. For example, FGFI (Wang et al., 2019) mimicked features within anchors near the ground truth (GT) boxes. DeFeat (Guo et al., 2021) applied different weights to the foreground and background area when imitating features. FGD (Yang et al., 2022a) separated the foreground and background in Focal Distill and imitated the relationship between pixels in Global Distill. LD (Zheng et al., 2022)operated on the logits from predicting heads rather than deep features. Turning boxes regression into classification problem through GFL (Li et al., 2020), and then make the student imitate the prediction of teacher location distribution.

DETR-families is the most attractive detection paradigm at present, but there are many difficulties in distilling knowledge on it. It's worth noting that DETR regards object detection as a set prediction problem, which means the instance predictions are sparse and unordered, making it difficult to align the intermediate properties between the teacher and student models. In addition, CNN has a fixed and regular receptive field, providing natural spatial correspondence for GT boxes and activated features. Besides, various DETRs utilize different attention strategies on cross-attention, to get more flexibility in sampling and aggregation of features from the backbone. Therefore, imitating features using artificially designed partitioning strategies will even deteriorate the detection performance.

We address the above challenges by summing up the structural similarities of most DETRs while minimizing the intervention of distillation components in the model itself. To this end, we propose **DETRDistill**, to greatly boost the distillation performance in three aspects. First, we establish prediction matching between object queries with progressive stage-by-stage instance distillation to gradually transfer useful knowledge to the student. Second, instead of the conventional imitation of features, we regard the content query aggregated from each decoder layer for feature-level distillation, making it agnostic to a multifarious cross-attention mechanism. We shorten the distance between corresponding object features, and at the same time, monitor the relationship between queries. Finally, to fully leverage the query embedding trained by the teacher model, we view them as additional training queries for the student to provide more samples, and also transfer the characteristics of the teacher in an interactive process.

Extensive experiments on the MS COCO dataset prove the universality and effectiveness of our method. DETRDistill achieves considerable gains on transformer-based detectors compared with various state-of-the-art knowledge distillation methods. For instance, it obtains 2.4 mAP and 2.5 mAP improvements on two representative DETRs: AdaMixer (Gao et al., 2022) and Deformable DETR (Zhu et al., 2020), respectively. We also provide more experiments with lightweight backbone distillation, self distillation and other ablations to explore the effectiveness of our approach. To the best of our knowledge, this is the first systematic study to explore a general distillation method for DETR-style detectors.

2 Related Work

2.1 Vision-based Object Detectors

During the initial phase of object detection, two-stage detectors represented by Faster R-CNN (Ren et al., 2015) have shown great performance. They usually use the region proposal network (RPN) to select the foreground boxes, and then predict the specific categories and refine the box located in the next stage. Later, one-stage detectors, which output classification and regression results directly soon attracted the attention of researchers because of their simple structures and lower latency (Lin et al., 2017; Tian et al., 2019; Zhang et al., 2020; Chen et al., 2021b). Some detectors analyze and model the prediction results. For example, GFL (Li et al., 2020) transforms location regression into classification. IoU-Net (Jiang et al., 2018)adds a branch to predict the IoU between the detected box and matched GT boxes. The above diverse detectors have one thing in common: predicting results by connecting several convolutional layers behind the backbone.

With the excellent performance of Transformer (Vaswani et al., 2017) in natural language processing, researchers have also started to explore the application of Transformer structure to visual tasks. DETR was the first work to introduce the Transformer structure based on the attention mechanism into the field of object detection. However, the DETR training process is extremely inefficient so many follow-up works have attempted to accelerate convergence. One line of work tries to redesign the attention mechanism. For example, Dai et al. (Zhu et al., 2020) proposed Deformable DETR,

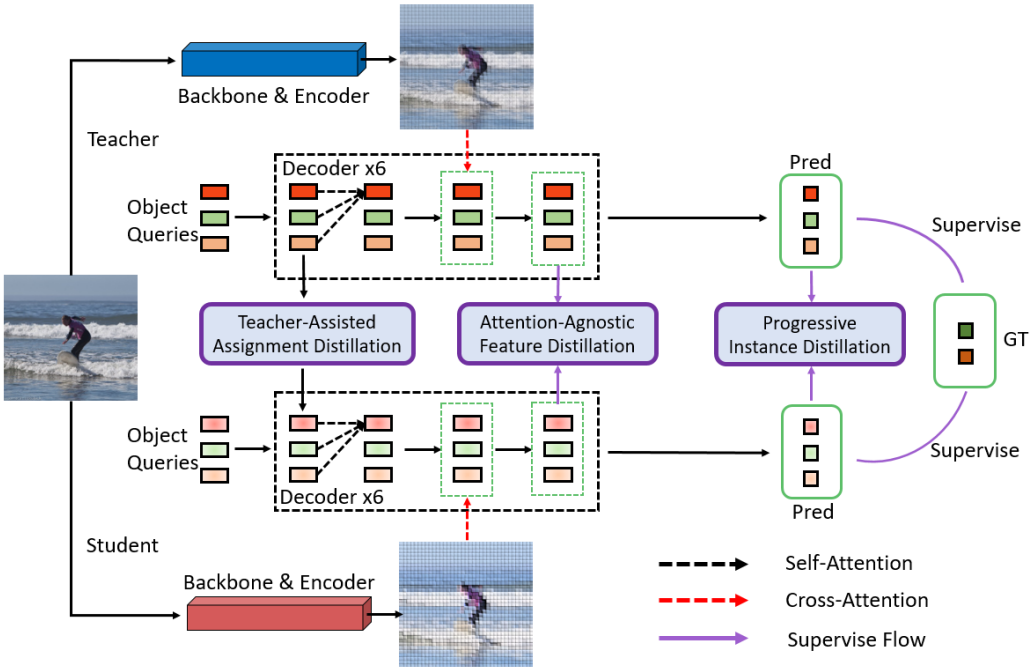

Figure 1: The overall architecture of our proposed DETRDistill, consists of a transformer-based teacher detector with a large backbone, a congener detector with a lightweight backbone, and the proposed distillation modules: (i) progressive instance distillation (ii) attention-agnostic feature distillation, and (iii) teacher-assisted assignment distillation.

which constructed a sparse attention mechanism by only interacting with the variable sampling point features around the reference points. SMCA (Gao et al., 2021) introduced Gaussian prior to limit cross-attention. AdaMixer (Gao et al., 2022) has designed a new adaptive adoption strategy without an encoder. Another line of work rethought the meaning of the query. Meng et al. (Meng et al., 2021) visualized that it is ineffective for DETR to rely on content embedding in the cross-attention to locate object extremity, and therefore proposed decoupling queries into content part and position part. Anchor-DETR (Wang et al., 2022) directly treated the query's 2D reference points as its position embedding to guide attention. DAB-DETR (Liu et al., 2022) introduced width and height information besides location to the attention mechanism to model different scale objects.

2.2 Knowledge Distillation in Object Detection

Knowledge distillation is a commonly used method for model compression. (Hinton et al., 2015) first proposed this concept and applied it in the field of image classification. They argue that soft labels output by the teacher contains "dark knowledge" of inter-category similarity compared to the one-hot encoding, which contributes to the generalization of the model. Attention transfer (Zagoruyko & Komodakis, 2016) focused the distillation on the feature map and transferred knowledge by narrowing the attention distribution of the teacher and student instead of distilling output logits. FitNet (Romero et al., 2014) proposed to mimic the intermediate-level hints of the teacher model by hidden layers. (Chen et al., 2017) first applied knowledge distillation to solve the multi-class object detection. (Li et al., 2017) thought that the background regions will introduce noise and proposed to distill the regions sampled by RPN. DeFeat (Guo et al., 2021) distilled the foreground and background separately. FGD (Yang et al., 2022a) imitated the teacher in terms of both focal regions and global relations of features, respectively. LD (Zheng et al., 2022) extended soft-label distillation to positional regression, causing the student to fit the teacher's border prediction distribution. MGD (Yang et al., 2022b) used masked image modeling (MIM) to transform the imitation task into a generation task.

3 METHOD

In this section, we first review the basic architecture of DETR and then introduce the concrete implementation of our proposed DETRDistill, which consists of three components: (i) progressive instance learning (ii) attention-agnostic feature distillation, and (iii) teacher-assisted label distillation. Figure 1 illustrates the overall architecture of DETRDistill.

3.1 A REVIEW OF DETR

DETR is an end-to-end object detector that includes a CNN backbone, learnable query embeddings, Transformer encoders and decoders. Given an image $I \in R^{H_0 \times W_0 \times 3}$, a CNN backbone extracts its spatial features and then Transformer encoders enhance the feature. With several updated features $F \in R^{HW \times d}$, query embeddings $Q \in R^{N \times d}$ are fed into Transformer decoders to produce the detection results, where d is the feature dimension, N is the number of queries, and H_0, W_0, and H, W denote the size of the image and the feature, respectively.

3.2 DETRDISTILL

3.2.1 PROGRESSIVE INSTANCE DISTILLATION

One of the most common strategies for knowledge distillation is to directly pass the predicted soft labels of the teacher to the student model for learning. However, the sparsity of prediction results and the instability of query's predictions (Li et al., 2022) make it difficult for DETRs to orderly correspond the teacher's results to the student's predictions. To achieve this goal, we utilize the Hungarian algorithm to solve the matching problem of the sparse predictions for the DETRs. Formally, let y^T and y^S denote the predicted outcomes from the teacher and the student respectively, conforming to $y^T = \{y_i^T\}_{i=1}^M$ and $y^S = \{y_i^S\}_{i=1}^N$. y_i^T is composed of c_i^T and b_i^T, representing category and location projections, respectively. Similarly, y_i^S is composed of c_i^S and b_i^S. Due to the nature of knowledge distillation, M is usually greater than or equal to N. Then we can search for a permutation $\hat{\sigma}$ between the outputs of the teacher and the student with the lowest cost:

$$\hat{\sigma} = arg\,min \sum_{i=1}^{N} \mathcal{L}_{match}(y_{\sigma(i)}^T, y_i^S), \tag{1}$$

$\mathcal{L}_{match}$ is a pair-wise matching cost, which is defined as

$$\mathcal{L}_{match} = \mathcal{L}_{cls}(c_{\sigma(i)}^T, c_i^S) + \mathcal{L}_{bbox}(b_{\sigma(i)}^T, b_i^S), \tag{2}$$

where $\mathcal{L}_{cls}$ is the KL divergence loss and $\mathcal{L}_{bbox}$ is a mix combination of L_1 loss and $GIoU$ loss.

However, the above strategy of transferring the teacher's knowledge exclusively to the student may be sub-optimal. Motivated by the learning curves (Yelle, 1979) and knowledge review mechanism (Chen et al., 2021a), we argue that a person should learn different levels of knowledge at different ages, and the correct direction of learning at the current phase is a guarantee of successful learning at the next stage.

The process of knowledge distillation is analogous to the above situation, based on which we propose progressive instance distillation. We hope that each stage of the student model acquires different levels of knowledge, laying a solid foundation for smooth learning in the next stage. The decoder part of DETR usually contains K stages (K>1), and the prediction of the current stage is a refinement of the previous stage. Thus we can regard each stage as a learning phase. Instead of taking the teacher's final prediction as a distillation target, we carefully align the outputs of the teacher and the student at each stage, allowing the model to progressively learn different levels of knowledge, which greatly eases the learning difficulty and improves better performance. Formally, let c_i^{Tk} and b_i^{Tk} represent the category and position of the k-th stage of the teacher respectively and c_i^{Sk} and b_i^{Sk} denote the student ones. According to Equation 2, we can employ the outputs of the k-th stage of the teacher and student to obtain the corresponding permutation $\sigma_k(i)$. Therefore, the distillation loss for the k-th stage can be written as

$$\mathcal{L}_{stage\_k} = \sum_{i=1}^{N} \alpha \mathcal{L}_{bce}(c^{Tk}_{\sigma_k(i)}, c^{Sk}_i) + \beta \mathcal{L}_{giou}(b^{Tk}_{\sigma_k(i)}, b^{Sk}_i) + \gamma \mathcal{L}_{L1}(b^{Tk}_{\sigma_k(i)}, b^{Sk}_i), \tag{3}$$

where $\mathcal{L}_{bce}$ is the binary cross entropy loss, $\mathcal{L}_{giou}$ is $GIoU$ loss, $\mathcal{L}_{L1}$ is L_1 loss and α, β and γ are reweighted factors. Finally, the total loss can be derived as

$$\mathcal{L}_{pll} = \sum_{k=1}^{K} \mathcal{L}_{stage\_k}, \tag{4}$$

3.2.2 Attention-agnostic Feature Distillation

CNN-based detectors directly output prediction results by connecting multiple convolutional layers to the features from the backbone. Since the convolutional network has a fixed square receptive field, the features acquired by the detection head come from the uniform interpolation sampling of the target area, which can be easily divided based on the GT boxes. The main operations of the decoder in DETRs are self-attention and cross-attention. Self-attention is the interaction mode between queries, which is often understood as a structure to prevent repeated prediction. Cross-attention is the main manner to extract and aggregate the object information from features. Then the refined features are calculated by:

$$\text{MultiHeadAttn}\,(q, x) = \sum_{m=1}^{M} \boldsymbol{W}_m \left[\sum_{k \in \Omega_k} A_{mqk} \cdot \boldsymbol{W}'_m x_k \right], \tag{5}$$

where $q \in \mathbb{R}^d$ is the representation feature of the query. x means sampled feature set Ω_k from the backbone or encoder and x_k is the key element. m indexes the attention head, $\boldsymbol{W}'_m$ and $\boldsymbol{W}_m$ are learnable weights and A_{mqk} denotes attention weight between each query and key.

The difference between DETR variants mainly lies in the feature sampling strategy and the generation method of attention. Original DETR uses a redundant multi-head attention mechanism. By calculating the attention based on cosine similarity between queries and each location of the feature map, the object feature with richer semantic information can be obtained after weighted fusion. Deformable DETR proposed a learnable sampling method, which is similar to deformable convolutional networks (Dai et al., 2017). It can adaptively sample sparse features near the reference point without being limited by the fixed square area and generate attention weight through the neural network. AdaMixer's Adaptive Mixing method allows sampling across feature layers in the full image space. It is worth noting that this flexible sampling method is one of the greatest advantages of DETRs. However, this contradicts the previous way of decoupling distillation for the square receptive field, that is, GT boxes are not aligned with the sampled feature by DETRs. The foreground-background separation distillation of features through GT box mask will make performance worse. The schematic diagram is shown in Figure 2 and our ablation Table 5 further verified this conclusion.

Considering the diverse attention mechanisms, we do not design complex strategies to divide features, but adapt to the attention of different DETRs and choose to bring the student and the teacher closer in the query-based feature level after the cross-attention. Considering that there exists representative divergence among different queries, we adopt a softer imitation loss, InfoNCE (He et al., 2020) to regularize the similarity across query features:

$$\mathcal{L}_{q-feature} = -\log \frac{\exp\left(q^S_{\sigma(i)} \cdot q^T_i / \tau\right)}{\sum_{j=1}^{N} \exp\left(q^S_{\sigma(i)} \cdot q^T_j / \tau\right)}, \tag{6}$$

By pushing out the query features between unmatched pairs, we can also provide explicit supervision to prevent repeated prediction. Moreover, considering that the one-to-one feature distillation is incomplete and the instance relationship is indispensable, we adopt Euclidean distance to express

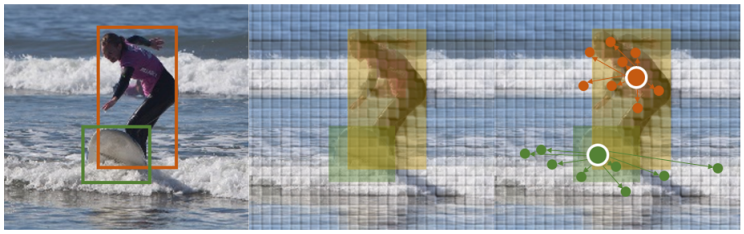

Figure 2: Effect of foreground-background decoupling distillation on different detectors. The middle column shows that the GT boxes mask region is aligned with the CNN receptive field. The flexible attention displayed on the right is more likely to use long-range features. For example, the characteristics of the surfboard itself are not obvious. It turns to rely on the ocean background to help speculate the foreground, but the GT box mask weight will affect attention in DETR.

the relationship information between q in teacher and transfer it to the student with L1 Loss, similar to GID (Dai et al., 2021):

$$\mathcal{L}_{q-relation} = \sum_{(i,j)\in\mathbb{N}^2} L1\left(\frac{1}{\phi(q^T)}\left\|q_i^T - q_j^T\right\|_2, \frac{1}{\phi(q^S)}\left\|q_{\sigma(i)}^S - q_{\sigma(j)}^S\right\|_2\right), \tag{7}$$

$$\phi(x) = \frac{1}{|\mathbb{N}^2|}\sum_{(i,j)\in\mathbb{N}^2}\left\|x_i - x_j\right\|_2, \tag{8}$$

where $\mathbb{N}^2 = \{(i,j) \mid 1 \leqslant i,j \leqslant N\}$ and the total loss is as follows:

$$\mathcal{L}_{feature} = \mathcal{L}_{q-feature} + \mathcal{L}_{q-relation}, \tag{9}$$

3.2.3 Teacher-assisted Assignment Distillation

When the student model is initially trained, noise-filled query embedding can lead to unstable bipartite graph matching. As mentioned in (Li et al., 2022), a query is often matched with different objects in different epochs, which makes optimization ambiguous and inconstant. In the setting of knowledge distillation, we usually have all the parameters of the trained teacher model, including query embedding with sufficient information about the objects. It is intuitive to utilize the information to improve the stability of the optimization direction. Based on this motivation, we propose Teacher-assisted Label Assignment. Let Q^T and Q^S denote the queries of the teacher and student, respectively, and more specifically:

$$Q^T = \{q_1^T, \cdots, q_M^T\}, \;\; Q^S = \{q_1^S, \cdots, q_N^S\}, \tag{10}$$

where M and N are the numbers of teacher queries and student queries.

The queries of the teacher and student are fed into the model at the same time, sharing the parameters of the network without any information interaction between them. Formally, we can get

$$\begin{aligned} \hat{y}^T &= \mathcal{FFN}(Decoder((Q^T, \mathcal{F}), \phi), \psi), \\ y^S &= \mathcal{FFN}(Decoder((Q^S, \mathcal{F}), \phi), \psi), \end{aligned} \tag{11}$$

where $\mathcal{F}$ is the feature map obtained from the input image and ϕ and ψ are the parameters of the $Decoder$ and $\mathcal{FFN}$. The two sets of queries generate two separate sets of detection results, which are matched according to the ground truth to jointly optimize the model parameters. This indicates that no matter how the student query changes, there are always stable queries from the teacher to guide the optimization of $Decoder$ and $\mathcal{FFN}$.

4 EXPERIMENTS

4.1 SETUP AND IMPLEMENTATION DETAILS

We evaluate DETRDistill on the challenging large-scale MS COCO benchmark (Lin et al., 2014). The $train2017$ (118K images) is utilized for training and $val2017$ (5K images) is used for validation. Our codebase is built on MMdetection toolkit (Chen et al., 2019). All models are trained on 8 NVIDIA V100 GPUs. Unless otherwise specified, we train the teacher model for $1\times$ schedule (12 epochs) using ResNet-101 (He et al., 2016) as the backbone, and train the student model for $1\times$ schedule (12 epochs) using ResNet-50 as the backbone. We use the standard COCO-style measurement, $i.e.$, average precision (mAP) for evaluation.

4.2 MAIN RESULTS

We compare DETRDistill with other state-of-the-art knowledge distiallation methods on two representative DETRs: AdaMixer (Gao et al., 2022) and Deformable DETR (Zhu et al., 2020). The results are presented in Table 1. We observe that our method surpasses other methods with a large margin for the above two detectors. Classical methods for CNN networks like FitNet (Romero et al., 2014) can still bring some gain (0.6 mAP) for DETR-serious. For the latest distillation methods for feature levels like MGD (Yang et al., 2022b), there is almost no positive effect (0.0 mAP and -0.1 mAP). FGD (Yang et al., 2022a) and LD (Zheng et al., 2022) which are specific to object detection also do not work well with Transformer-based detectors, and even cause degradation of the results. FGD performs worse on the AdaMixer. We speculate that the sampling method utilized by AdaMixer is global and cross the FPN stages. While the DETRDistill can gain 2.4 mAP on the AdaMixer and 2.5 mAP on the Deformable DETR, which validates the effectiveness of our approach.

Table 1: A comparison between our DETRDistill with other state-of-the-art distillation methods on the COCO validation set.

Detector	Setting	Epoch	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
AdaMixer	Teacher	12	45.3	64.6	49.2	27.3	48.3	61.9
	Student	12	42.3	61.2	45.6	25.3	44.8	58.2
	FGD	12	40.7(-1.6)	59.3	43.4	23.4	43.3	55.8
	MGD	12	42.3(+0.0)	61.3	45.5	24.5	45.0	58.9
	FitNet	12	42.9(+0.6)	61.7	46.2	24.7	45.8	59.4
	LD	12	41.4(-0.7)	60.4	44.7	23.6	44.2	57.6
	Ours	12	**44.7**(+2.4)	**62.9**	**48.2**	**26.7**	**47.6**	**61.0**
Deformable DETR	Teacher	50	44.8	64.1	48.9	26.5	48.3	59.6
	Student	50	44.1	63.2	47.9	27.0	47.4	58.3
	FGD	50	44.1(+0.0)	63.1	48.0	25.9	47.7	58.8
	MGD	50	44.0(-0.1)	63.1	48.0	25.9	47.3	58.6
	FitNet	50	44.9(+0.8)	64.3	48.9	27.2	48.4	59.6
	LD	50	43.7(-0.4)	62.4	47.2	25.3	46.8	58.8
	Ours	50	**46.6**(+2.5)	**65.6**	**50.7**	**28.5**	**50.0**	**60.4**

4.3 ABLATION STUDIES

In this section, we explore the role of each module in DETRDistill through more detailed ablation studies. More specifically, we adopt Adamixer as the baseline model. If not specified, teachers use ResNet-101 with 300 queries and students use ResNet-50 with 100 queries.

4.3.1 MAIN ABLATIONS

To study the impact of each component in DETRDistill, we report the performance of each module in Table 2. Our baseline starts from 42.3 mAP. When progressive instance distillation, attention-agnostic feature distillation and teacher-assisted assignment distillation are applied separately, we

can obtain the gain of 1.4 mAP, 1.0 mAP and 1.1 mAP respectively. Concentrating on the more specific AP_S, AP_M and AP_L, the three components have different emphases when it comes to transferring teacher's knowledge. Finally, the AP performance achieves 44.7 when all three modules are applied together, gaining a 2.4 absolute improvement and validating our approach.

Table 2: Effectiveness of each component in DETRDistill. Results are reported on AdaMixer-R50 with standard 1 $\times$ schedule setting.

Instance Distill	Feature Distill	Assign Distill	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
			42.3	61.2	45.6	25.3	44.8	58.2
✓			43.7	61.7	47.2	25.3	46.5	60.7
	✓		43.3	62.4	46.6	26.1	46.2	59.5
		✓	43.4	62.2	46.4	25.3	46.0	59.9
✓	✓		44.3	62.4	48.0	25.8	47.0	61.0
✓	✓	✓	**44.7**	**62.9**	**48.2**	**26.7**	**47.6**	**61.0**

4.3.2 Ablation on Progressive Instance Distillation

Explore the impact of different imitation learning strategies for instance distillation.

To verify the effectiveness of progressive distillation, we additionally explore two ways of transferring a teacher's predicted outputs to the student. The first one is to regard the final predictions of the teacher as the goal of the student for all stages, represented by direct distillation. As we can see from Table 3, the progressive instance distillation can achieve 0.6 mAP higher than direct distillation, which verifies the effectiveness of our method. However, the different information levels in different stages lead to inconsistency when matching the same query of the student in different stages with different teacher's queries in bipartite graph matching. Based on this, we propose another learning approach progressive guided distillation, which only performs bipartite graph matching in the last stage, and then uses this matching result as the correspondence between student's queries and teacher's queries in the previous stages. Through the experiment, we can find that this strategy causes a sharp decline (1.0 mAP) in the final result, indicating that it is better to let student choose what they learn on demand at different stages rather than forcing.

Table 3: Comparison on different imitation learning approaches for instance distillation.

Strategies	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
-	42.3	61.2	45.6	25.3	44.8	58.2
Direct Distillation	43.1	61.7	46.3	**25.4**	45.9	59.7
Progressive Guided Distillation	42.7	60.9	46.2	24.9	45.4	59.3
Progressive Instance Distillation	**43.7**	**61.7**	**47.2**	25.3	**46.5**	**60.7**

Ablations on loss and positive/negative samples for classification and regression branches.

The outputs predicted by the teacher include the classification probabilities and the regression locations. The soft label of classification probability is rich in semantic information, while it is unclear whether the location information of negative samples predicted by the teacher is beneficial to the student. Therefore, we explore the performance of the instance distillation with positive/negative samples on the regression branch, respectively. We pre-define the maximum classification probability of the teacher's prediction less than 0.1 as the background. What's more, we explore the importance of the classification branch in the distillation process. The experimental results are shown in Table 4. We can find that using regression position and classification probability alone as extra supervision yields an improvement of 1.3 mAP and 0.3 mAP, respectively. This indicates that the location information is more difficult to learn for the DETRs detectors, making it more important in the distillation phase. We also explore the effectiveness of foreground and background regional

distillation on the student. The experimental results show that the background location information can bring more gain than the foreground location information, which is not straightforward. The reason can be two-fold: (i) though the queries are assigned as background, it can still be viewed as an interaction from the teacher model, providing hints for the model learning process, (ii) the number of background queries is much larger than that of the foreground ones, providing more useful gradient for the student to learn from the teacher.

Table 4: Ablations on loss of progressive instance distillation for classification and regression.

Pos Cls	Neg Cls	Pos Reg	Neg Reg	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
				42.3	61.2	45.6	25.3	44.8	58.2
✓				42.6	**62.0**	45.9	24.9	45.4	58.9
		✓		43.1	61.4	46.4	25.1	45.9	59.5
			✓	43.4	61.5	47.0	25.2	46.2	60.1
		✓	✓	43.6	61.6	47.2	25.2	46.3	60.6
✓		✓	✓	**43.7**	61.7	**47.2**	**25.3**	**46.5**	**60.7**

4.3.3 Ablation on Attention-agnostic Feature distillation

Ablations on Distillation for feature zoning planning strategy.

Distillation methods based on feature imitating are the mainstream at present. *Are these most advanced methods suitable for the structure of DETRs*? We conduct detailed experiments on the feature zoning strategy selection in Table 5. We first conduct the FGD (Yang et al., 2022a) and FKD (Zhang & Ma, 2020), which both use their weighting method to distinguish between the foreground and the background. The result is surprisingly lower than its vanilla baseline. We believe that mandatory regional division and weighting affect the attention mechanism of the DETR itself. MGD (Yang et al., 2022b) uses a global random mask policy to force the student to generate the teacher's features. This randomness mask can accelerate the convergence in the early stage, but the final effect is equal to the baseline. Inside GT Box refers to only distilling the features inside GT boxes, but its result is still lower than FitNet (Romero et al., 2014). This shows that although the former has received some guidance from the teacher, it still loses equilibrium during distillation. To avoid the problem, we convert the distillation object to the query embeddings after cross-attention aggregation.

Table 5: Comparison on several regionally decoupled weighted distillation approaches.

Method	Processing Strategy	Epoch4 AP	Epoch8 AP	Epoch12 AP
-	-	35.0	38.7	42.3
FGD	Focal and Global Decoupling	34.4	39.1	40.7(-1.6)
FKD	Attention-guided and Non-local Distillation	35.9	39.5	42.2(-0.1)
MGD	Global Random Mask	36.3	39.8	42.3(+0.0)
Inside GT Box	Imitation Regions in GT	35.6	39.3	42.6(+0.3)
FitNet	Imitation Whole Feature Equally	36.4	39.6	42.9(+0.6)
Ours	-	**36.6**	**40.0**	**43.3**(+1.0)

5 Conclusion

In this paper, we introduce a simple knowledge distillation framework for DETR-style detectors, named DETRDistill. We explore the methods in three ways: progressive instance distillation, attention-agnostic feature distillation, and teacher-assisted assignment distillation. Extensive experiments demonstrate the effectiveness and generalization of our approach. Notably, it achieves considerable gains on various transformer-based detectors compared with current state-of-the-art knowledge distillation methods. We hope that DETRDistill can serve as a solid baseline for DETR-based knowledge distillation for future research.

REFERENCES

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.

Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5008–5017, 2021a.

Zehui Chen, Chenhongyi Yang, Qiaofei Li, Feng Zhao, Zheng-Jun Zha, and Feng Wu. Disentangle your dense object detector. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4939–4948, 2021b.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 764–773, 2017.

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

Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of detr with spatially modulated co-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3621–3630, 2021.

Ziteng Gao, Limin Wang, Bing Han, and Sheng Guo. Adamixer: A fast-converging query-based object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5364–5373, 2022.

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

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.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 784–799, 2018.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13619–13627, 2022.

Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 6356–6364, 2017.

Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3651–3660, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

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.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627–9636, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4933–4942, 2019.

Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2567–2575, 2022.

Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4643–4652, 2022a.

Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. *arXiv preprint arXiv:2205.01529*, 2022b.

Louis E Yelle. The learning curve: Historical review and comprehensive survey. *Decision sciences*, 10(2):302–328, 1979.

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.

Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2020.

Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9759–9768, 2020.

Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Ming-Ming Cheng. Localization distillation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9407–9416, 2022.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

A APPENDIX

A.1 MORE EXPERIMENTAL RESULTS

A.1.1 DISTILLING ON LIGHTWEIGHT BACKBONES WITH DIFFERENT SETTINGS

The goal of knowledge distillation is to transfer as much knowledge as possible from a large model to a lightweight model for the purpose of deployment on the edge. With this in mind, we also apply DETRDistill with small backbones including ResNet-18 and MobileNetV2 (Sandler et al., 2018) on AdaMixer and Deformable DETR. For AdaMixer, the teacher model uses 300 queries, while the student model uses 100 queries for training, and all other settings are kept consistent with the paper. For Deformable DETR, the teacher model and the student model adopt the same setting except for the backbone. The results are shown in Table 6. Our distillation method has achieved the best performance on both backbones. We achieve 2.0/2.6mAP improvements on AdaMixer, and 3.3/3.5mAP respectively on Deformable DETR. It is worth noting that Deformable DETR with ResNet-18 as the backbone has nearly the performance of the teacher after distillation, proving promising in resource-constrained applications.

Table 6: Experimental results of DETRDistill on smaller backbones: ResNet-18 and MobileNetV2 on COCO validation subset.

Detector	Setting	Query	Backbone	Epoch	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
AdaMixer	Teacher	300	ResNet-101	12	45.3	64.6	49.2	27.3	48.3	61.9
	Student	100	ResNet-18	12	38.3	56.9	40.9	20.9	40.2	53.9
	Ours	-	-	-	**40.3**(+2.0)	**58.0**	**43.3**	**22.8**	**42.4**	**56.9**
	Teacher	300	ResNet-101	12	45.3	64.6	49.2	27.3	48.3	61.9
	Student	100	MobileNetV2	12	36.6	55.2	38.9	20.4	38.5	51.4
	Ours	-	-	-	**39.2**(+2.6)	**56.7**	**42.2**	**22.0**	**41.5**	**55.4**
Deformable DETR	Teacher	300	ResNet-101	50	44.8	64.1	48.9	26.5	48.3	59.6
	Student	300	ResNet-18	50	40.0	58.0	43.3	23.0	42.9	53.7
	Ours	-	-	-	**43.3**(+3.3)	**61.3**	**47.2**	**25.0**	**46.1**	**57.1**
	Teacher	300	ResNet-101	50	44.8	64.1	48.9	26.5	48.3	59.6
	Student	300	MobileNetV2	50	38.8	56.8	42.0	23.2	41.7	51.9
	Ours	-	-	-	**42.3**(+3.5)	**60.4**	**46.0**	**23.5**	**45.5**	**56.2**

A.1.2 SELF-DISTILLATION

Self distillation is a special case of knowledge distillation where the models of teacher and student are aligned, with the only aim of improving the performance of the model. The teacher and student all use ResNet-50 as the backbone. We compare the self-distillation performance of our method with FGD (Yang et al., 2022a) and MGD (Yang et al., 2022b) based on Adamixer and Deformable DETR. Table 7 shows that our DETRDistill achieves a gain of 1.3 mAP and 2.3 mAP over the baselines. While FGD and MGD do not bring any improvement and even cause a decline in results.

Table 7: Self-distillation performance on COCO validation subset. ResNet-50 is adopted as teacher and student backbone.

Detector	Setting	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
AdaMixer	T & S	42.3	61.2	45.6	25.3	44.8	58.2
	FGD	42.0	60.9	45.2	23.5	44.9	58.5
	MGD	41.7	60.7	44.9	23.0	44.7	58.6
	Ours	**43.6**(+1.3)	**62.0**	**46.9**	**26.7**	**46.2**	**59.1**
Deformable DETR	T & S	44.1	63.2	47.9	27.0	47.4	58.3
	FGD	44.2	63.1	48.0	26.3	47.8	58.5
	MGD	44.2	63.3	48.2	26.9	47.7	58.6
	Ours	**46.4**(+2.3)	**65.3**	**50.4**	**28.9**	**49.9**	**60.0**

A.2 DISCUSSION

A.2.1 WHY THE DISTILLATION ON LOCATION REGRESSION CAN IMPROVE THE PERFORMANCE OF DETRS?

Previous distillation methods rarely considered the supervision of location regression. LD (Zheng et al., 2022) uses GFL (Li et al., 2020) to express location distribution and transfer the localization knowledge from the teacher to the student. Moreover, they also proposed a valuable localization region surrounding positive samples that can aid in distillation. Our experiment also proves that regression supervision on negative samples in DETR can greatly improve performance, which is an advantage that GT supervision cannot bring. We infer the reason that the prediction in DETRs are quite sparse (only 100/300 queries in each image) compared with the conventional CNN-based detectors, leading to limited positive samples for knowledge transfer. Besides, negative samples can also be viewed as the response from the teacher model, as a knowledge passing agent to transfer useful information. Therefore, it is necessary to obtain the knowledge of negative samples from the teacher.

A.2.2 HOW DIFFERENT ATTENTION MECHANISMS AFFECT THE PERFORMANCE OF FEATURE IMITATION WITH PARTITION STRATEGY?

The results in Table 5 show that distilling the whole feature map in DETRs is better than only in the foreground region, which is different from the traditional CNN detectors. There are also some details worth noting, as shown in Table 1. FGD severely damages the performance in AdaMixer, but it has little impact on Deformable DETR. This is mainly due to differences in the attention sampling methods. In Deformable DETR, it learns deformable local offset for attention aggregation, while detectors, such as DETR and AdaMixer, implement a global sampling strategy, which is more likely to be susceptible to long-range dependence.