Teach-DETR: Better Training DETR With Teachers

Linjiang Huang, Kaixin Lu, Guanglu Song, Liang Wang, Fellow, IEEE, Si Liu, Senior Member, IEEE, Yu Liu, and Hongsheng Li, Member, IEEE

Abstract—In this paper, we present a novel training scheme, namely Teach-DETR, to better train DETR-based detectors from various types of teacher detectors. We show that the predicted boxes from teacher detectors are effective medium to transfer knowledge of teacher detectors, which could be either RCNN-based or DETR-based detectors, to train a more accurate and robust DETR model. This new training scheme can easily incorporate the predicted boxes from multiple teacher detectors, each of which provides parallel supervisions to the student DETR. Our strategy introduces no additional parameters and adds negligible computational cost to the original detector during training. During inference, Teach-DETR brings zero additional overhead and maintains the merit of requiring no non-maximum suppression. Extensive experiments show that our method leads to consistent improvement for various DETR-based detectors. Specifically, we improve the state-of-the-art detector DINO Zhang et al. 2022 with Swin-Large Liu et al. 2021 backbone, 4-scale feature pyramid and 36-epoch training schedule, from 57.8% to 58.9% in terms of mean average precision on COCO 2017 val set.

Index Terms—Object detection, detection transformer, knowledge transfer.

I. INTRODUCTION

OBJECT detection is a fundamental task in computer vision, which aims to localize the boxes and categories of objects of interest. Due to the significant successes in various fields, deep learning has been the prevailing solution for object detection. The previous development is rapidly advanced by a series of RCNN-based methods, e.g., Faster-RCNN [3], YOLO [4], RetinaNet [5], etc.

Manuscript received 15 March 2023; revised 17 September 2023; accepted 23 September 2023. Date of publication 26 September 2023; date of current version 3 November 2023. This work was supported in part by the National Key R&D Program of China Project under Grant 2022ZD016100, in part by the Centre for PERCEPTUAL and Interactive INTELLIGENCE (CPII) Ltd. under the Innovation and Technology Commission (ITC)’s InnoHK, and in part by the General Research Fund of Hong Kong RGC Project under Grant 14200421. The work of Hongsheng Li is a PI of CPII under the InnoHK. Recommended for acceptance by J. Wang. (Corresponding author: Hongsheng Li.)

Linjiang Huang is with the Multimedia Laboratory, Centre for Perceptual and Interactive Intelligence (CPII), The Chinese University of Hong Kong, Hong Kong, China (e-mail: ljhuang524@gmail.com).

Hongsheng Li is with the Multimedia Laboratory, Centre for Perceptual and Interactive Intelligence (CPII), The Chinese University of Hong Kong, Hong Kong, China, and also with Shanghai AI Laboratory, Shanghai 201210, China (e-mail: hsl@ee.cuhk.edu.hk).

Kaixin Lu is with Shanghai University, Shanghai 200444, China (e-mail: lukaixin@shu.edu.cn).

Guanglu Song and Yu Liu are with Sensetime Research, Shanghai 200233, China (e-mail: guanglusong@foxmail.com; liuyuisanai@gmail.com).

Si Liu is with Beihang University, Beijing 100191, China (e-mail: liusi@buaa.edu.cn).

Liang Wang is with the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100045, China (e-mail: wangliang@nlpr.ia.ac.cn).

Code is available at https://github.com/LeonHLJ/Teach-DETR.

Digital Object Identifier 10.1109/TPAMI.2023.3319387

Recently, DETection TRansformer (DETR) [6] introduces the Transformer [7] to serve as detection heads and achieves impressive performance without the need of non-maximum suppression (NMS). The design of DETR is very different from the RCNN-based detectors [3], [4], [5]. On top of the feature map from a backbone network, a Transformer encoder-decoder regards and assigns object detection as a set prediction problem [8]. Nevertheless, through the rapid development of classical detectors in the past decade, there have been many mature designs and tens of well-learned RCNN-based detectors with promising performance. We would like to ask the following question: given well-trained RCNN-based detectors and DETR-based detectors, can we effectively transfer their knowledge to train a more accurate and robust DETR model? Knowledge distillation (KD) [9] seems to be a plausible way to transfer knowledge from teacher detectors to DETRs. However, our experiments in Table I show that, existing knowledge distillation methods [10], [11], [12] cannot effectively transfer knowledge from RCNN-based detectors to DETR-based detectors. Native KD methods sometimes even result in worse performance to student DETR-based detectors.

In this paper, we propose a novel training scheme, namely Teach-DETR, to construct proper supervision from teacher detectors for training more accurate and robust DETR models. We find that the predicted bounding boxes of teacher detectors can serve as an effective bridge to transfer knowledge from different teacher detector and the DETR-based student detector. Our research highlights the following crucial observations: first, the bounding boxes offer an unbiased representation of objects for all detectors, which remains unaffected by any discrepancies in detector frameworks, label assignment strategies, and so on. Second, incorporating additional box annotations significantly increases the capacity of object queries by training them with various one-to-one assignment patterns, thereby enhancing

| Method | AP | AP50 | AP75 | APs
|--------|----|------|------|------|
| Student (H-Def-DETR) | 52.5 | 71.1 | 57.3 |
| Teacher (Mask-RCNN) | 46.4 | 67.0 | 50.5 |
| FKID [10] | 52.0 (±0.5) | 70.5 | 57.1 |
| DeFeat [11] | 52.2 (±0.3) | 70.8 | 57.1 |
| FGD [12] | 51.7 (±0.8) | 70.0 | 57.0 |
| Teach-DETR | 53.5 (±1.0) | 72.0 | 58.5 |

The student DETR is an H-Deformable-DETR [13] with Swin-small backbone, 300 queries and 12-epoch training schedule. The teacher detector is the Mask-RCNN [14] with Swin-small backbone and 36-epoch training schedule.
training efficiency and detection performance. Lastly, leveraging the predicted bounding boxes of teacher detectors does not introduce additional architectures, as it has a similar format to ground truth (GT) boxes.

Nevertheless, it is non-trivial to integrate the GT annotations with the auxiliary supervisions. Due to the one-to-one matching of query-label assignment of DETR-based detectors, incorporating auxiliary supervisions can be detrimental to maintain DETR’s key signature of enabling inference without NMS. To overcome the challenge of transferring knowledge across different types of detectors, we propose a solution that aligns with our observations and utilizes output boxes and scores of teacher detectors as extra independent and parallel supervision groups for training query outputs.

However, the ambiguity between GT boxes and the teacher’s imperfect boxes poses a significant challenge in training the student DETR model with both types of boxes. To tackle this issue, we propose a teacher-box reweighting strategy that employs the output scores of teacher’s boxes to measure their qualities and assign them different weights for better balancing their losses. Our Teach-DETR is a versatile training scheme and can be integrated with various popular DETR-based detectors without modifying their original architectures. Moreover, our framework has no requirement on teacher architectures, it can be RCNN-based detectors or DETR-based detectors, which is more general to various types of teachers. In contrast, existing DETR distillation methods require both the teacher and student models to be DETR-based models which have much fewer use scenarios than our method. During training, our method introduces no additional parameters and adds negligible computational cost upon the original detector. During inference, our method brings zero additional overhead. We conduct extensive experiments to prove that Teach-DETR can consistently improve the performance of DETR-based detectors. For example, our approach improves the state-of-the-art detector DINO [1] with Swin-Large [2] backbone, 4-scale feature pyramid and 36-epoch training schedule, from 57.8% to 58.9% in terms of mean average precision on COCO 2017 [15] val set, demonstrating effectiveness of our method even when applied to a state-of-the-art high-performance detector.

II. RELATED WORKS

A. DETR and its Variants

DEtection TRansformer (DETR) [6] is a state-of-the-art method for object detection that utilizes Transformers [7], a type of neural network architecture that excels at handling sequential data. Unlike traditional object detection methods, DETR does not require non-maximum suppression, which makes it more efficient and accurate. However, a major challenge of DETR is its slow convergence [16], [17], which hinders its practical implementation. Fortunately, many researchers are working on addressing this limitation and achieving faster and better detection performance.

Several approaches have been proposed to improve the Transformer layers in DETR. For example, Deformable-DETR [17] introduces a multi-scale deformable attention scheme that sparsely samples a small set of key points to better attend to important regions. Other works [16], [18], [19], [20] argue that the slow convergence of DETR is primarily due to its inability to rapidly focus on regions of interest. To address this issue, some approaches propose to modulate the cross-attentions of DETR, allowing queries to attend to restricted regions. Anchor-DETR [19], for instance, uses anchor points as object queries, with each anchor point responsible for attending to a restricted region. Conditional-DETR [18] decouples each query into a content query and a positional query, with the latter limiting the spatial range for the content queries to focus on nearby regions.

Recently, some approaches [1], [21] have attributed the slow convergence issue to the unstable bipartite matching used in DETR. To stabilize bipartite graph matching and accelerate model convergence, these approaches introduce an auxiliary query denoising task. Other approaches [13], [22] propose to duplicate ground truth boxes to support one-to-many matching in DETR. In contrast to previous approaches, the proposed method suggests a new training scheme to learn better DETR-based detectors from teachers. Our method is orthogonal to previous methods, meaning it can further improve their performance.

B. Transfer Knowledge for Object Detection

For object detection, the commonly used way of leveraging teacher detectors is knowledge distillation (KD) [9]. KD is usually applied to decrease the model complexity while improving the performance of the smaller student model. Compared to KD in classification, KD in object detection should consider more complex structure of boxes and somewhat cumbersome pipelines, and there are more hints, e.g., anchor predictions [23], [24], [25], proposal ranking results [26], object-related features [27], [28], [29], [30], contextual features [11], [12] and relations among features [12], [31], can be used. In contrast, our method does not mean to conduct knowledge distillation or model compression. The core idea of Teach-DETR is to transfer knowledge of various teachers to train a more accurate and robust DETR-based detector. Therefore, the teacher detectors can be smaller or perform worse than the “student” DETR. Besides, our approach could be a meaningful exploration, providing a way to distill knowledge from any detectors to DETRs.

There have been few works that have used knowledge distillation for DETR. For instance, ViDT [32] attempts to use KD between output queries of teacher DETR and student DETR to improve the student’s performance and reduce computational costs. There are also some concurrent preprints that have investigated knowledge distillation between DETRs. Specifically, DETRDistill [33] proposes to distill the predictions of teacher DETR to student DETR layer by layer. They feed the teacher’s query to the student model to provide stable input queries. D^2ETR [34] and KD-DETR [35] use consistent input queries, randomly sampled or copied from teacher queries, as the input queries of both the student’s and teacher’s decoder, to distill class and box predictions, as well as attention weights between teacher and student. However, these methods have the limitation of only distilling knowledge between DETRs of the same architecture but different backbones. In contrast, our method can transfer knowledge from detectors, either RCNN-based or DETR-based, to DETRs. Additionally, our method shows that even worse
teachers (refer to Table VIII) can improve DETRs’ performance, nobody has reported similar results before.

### III. Method

Our approach aims to transfer knowledge of teacher detectors to train a more accurate and robust DETR-based detector. Our methodology involves incorporating the predicted bounding boxes from multiple teacher detectors, which may include RCNN-based detectors, DETR-based detectors, or a combination of both, to act as auxiliary supervision during the DETR training process. Additionally, our experiments show that the proposed teacher-box reweighting plays a crucial role in effectively transferring knowledge from teacher detectors to enhance DETR-based student’s accuracy.

In this section, we first provide an overview of DETR [6] (Section III-A), followed by a detailed description of Teach-DETR (Section III-B). Finally, we examine the factors responsible for the success of Teach-DETR (Section III-C).

#### A. Preliminary

The DEtection TRansformer (DETR) model, introduced in [6], is a state-of-the-art object detection algorithm that leverages a combination of backbone models such as ResNet [36] or Swin Transformer [2] and a Transformer encoder-decoder architecture for accurate box classification and regression.

Initially, the backbone model is used to extract image features, which are then passed to the Transformer encoder for global feature aggregation across each image token. The Transformer decoder utilizes a fixed set of $N$ queries, which undergo self-attention among object queries and cross-attention between queries and image tokens. Subsequently, the queries are fed into the detection heads to produce $N$ box predictions $P = \{p_0, p_1, \ldots, p_N\}$. DETR utilizes the Hungarian matching algorithm [8] to establish bipartite matching $\hat{\sigma}$ between the ground truth boxes and the predictions with the lowest cost. This matching is achieved by minimizing the cost function, defined as

$$\hat{\sigma} = \arg\min_{\sigma \in S_N} \sum_{i} L_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) ,$$

(1)

where $S_N$ is a permutation of $N$ elements and $L_{\text{match}}$ is the box classification and regression cost between the ground truth $y_i$ and the prediction $\hat{y}_{\sigma(i)}$ of the query $\sigma(i)$. To achieve uniform batch sizes, the ground truth data is augmented by padding it with $\emptyset$, which signifies the absence of an object.

#### B. Training DETR With Teachers

**Exploration of DETR With Knowledge Distillation:** We endeavor to enhance the detection performance of DETR-based detectors by utilizing various teacher detectors. While knowledge distillation (KD) techniques seem like a straightforward solution, they have limited efficacy in transferring knowledge between RCNN-based and DETR-based detectors. To address this issue, we employ state-of-the-art KD approaches [10], [11], [12] to perform feature imitation between the features generated by Mask RCNN’s FPN and the multi-scale features generated by Swin-Small backbone. In DeFeat [11], we attempt to distill the classification logits of the predicted boxes. As shown in Table I, the pipelines of RCNN-based and DETR-based detectors differ significantly, making it challenging to transfer knowledge from RCNN-based detectors to DETR-based ones. Existing KD methods may even lead to poorer performance of the student.

**Auxiliary Supervisions From Teacher Detectors:** In light of the above challenges, we propose to leverage the predicted bounding boxes of teachers as the auxiliary supervisions for better training DETRs. There are two main reasons for using the predicted boxes for knowledge transfer. First, the bounding box is the unbiased representation of results of all detectors, which would not be affected by the discrepancies in different detector frameworks, so it can serve as a good medium to transfer the knowledge of teachers to student DETRs. Second, introducing more box annotations would largely unleash the capacity of object queries by providing more positive supervisions [13], [22], and thus improve training efficiency and detection performance. The whole pipeline of our Teach-DETR is shown in Fig. 1.

![Fig. 1. Pipeline of our proposed Teach-DETR. The auxiliary supervisions from multiple teacher detectors are used to conduct bipartite matching and weighting independently. The final training loss is the sum of the losses of $K$ teachers and the original losses of the GT boxes.](image)

However, since the one-to-one matching is the critical design of DETRs to discard the NMS during inference, naively introducing the auxiliary supervisions would result in ambiguous training targets. How to balance the contributions between different sets of teachers’ output boxes and GT boxes is a problem. In Table II, we show that simply concatenating the auxiliary boxes and the GT boxes greatly deteriorates the performance.

We try to address the above issue by conducting the one-to-one matching between GT boxes and object queries, and between each teacher’s boxes and object queries independently. The matchings are also properly weighted according to the teachers’ confidences. Specifically, for an input image, given $K$ teacher detectors, we can obtain $K$ sets of detection boxes, for the $i$th...
teacher, it contains for example $M$ predicted bounding boxes $\{\hat{y}_1, \ldots, \hat{y}_m\}$. Each of these predicted boxes $\hat{y}_j$ contains the estimated box size and location, the predicted category, and the predicted confidence score $\hat{y}_j$. We apply one-to-one assignment between queries not only to the GT, but also to the auxiliary supervision of each teacher detector independently and thus we can obtain $K$ groups of matching

$$
\hat{y}_1 = \arg \min_{\sigma_1 \in \mathcal{A}_1} \sum_{i} \mathcal{L}_{\text{match}}(\hat{y}_i, \hat{y}_{\sigma_1(i)}) ,
$$

$$
\vdots
$$

$$
\hat{y}_k = \arg \min_{\sigma_k \in \mathcal{A}_k} \sum_{i} \mathcal{L}_{\text{match}}(\hat{y}_i, \hat{y}_{\sigma_k(i)}) ,
$$

where $\hat{y}_k$ is the matching result of the $k$th teacher’s auxiliary supervisions. The final training loss is the sum of the losses of $K$ teachers and those of the GT boxes.

Eq. (2) exhibits a semblance to the formulation employed in Group-DETR [22], for the multi-group matching, albeit with a noteworthy distinction. Our approach features a solitary group of object queries, while encompassing multiple groups of bounding boxes, including both GT boxes and auxiliary boxes. We apply one-to-one matching between each group of box annotations and the single group of object queries, respectively. In contrast, Group-DETR employs multiple groups of object queries alongside a solitary group of boxes, i.e., GT boxes. Therefore, they apply one-to-one matching between each group of object queries and the GT boxes, respectively. Consequently, the one-to-one matching mechanism delineated by (2) operates in a manner that is “inversely” oriented compared to Group-DETR’s operation.

Teacher-Box Reweighting: Despite its potential benefits, the coexistence of ground truth (GT) boxes and imperfect teacher boxes poses a significant challenge in training the student DETR model. To tackle this issue, we propose a solution that leverages the confidence scores of boxes predicted by the teacher detectors. This approach not only balances the supervision signal from the two types of boxes but also mitigates the negative impact of low-quality auxiliary boxes.

Essentially, confidence scores can serve as proxies for the quality of bounding boxes to some extent. Therefore, we use them as weighting factors to adjust the corresponding losses for each box’s classification and regression. This mechanism enables us to selectively emphasize or de-emphasize the influence of teacher boxes based on their confidence scores, thereby improving the overall performance of the student DETR. The final loss $L^i_j$ for each auxiliary box $y^i_j$ can be expressed as follows:

$$
L^i_j = \hat{y}^i_j \cdot (\lambda_{\text{cls}} L^i_{\text{cls},j} + \lambda_{\text{box}} C^i_{\text{box},j}) ,
$$

where $L_{\text{cls}}$ and $L_{\text{box}}$ denote the classification loss and box regression loss, respectively, including $\ell_1$ loss and GIoU loss [6]. The balancing weights $\lambda_{\text{cls}}$ and $\lambda_{\text{box}}$ determine the importance of each loss component. For negative queries, we assign a moderate score of 0.5.

As illustrated in Table II, incorporating auxiliary boxes can boost the performance of Swin-Small $H$-Deformable-DETR [13] by 0.6%. Moreover, our method, which employs confidence scores, further improves the average precision by 1.0%, with the help of a single Swin-Small Mask RCNN [14] teacher for providing auxiliary supervision.

C. Why the Auxiliary Supervisions Help?

The implementation of auxiliary supervisions by teacher detectors contributes to a significant improvement in the performance of DETR-based detectors. In this study, we aim to assess the effectiveness of this technique through experimental analysis. Our baseline is $H$-Deformable-DETR [13], which employs a Swin-S backbone, 300 queries, and a 12-epoch schedule. We introduce auxiliary supervisions by leveraging predicted boxes from a Swin-S Mask RCNN [14].

Enrich Supervisions: Referring to Fig. 2, we observe that some teacher’s boxes contain newly-annotated objects, such as the keyboard of a laptop, which can enhance the annotations. Our statistical analysis reveals that roughly 15% of the teacher’s boxes comprise newly-annotated boxes that do not overlap with the GT boxes. To exploit these newly-annotated boxes as auxiliary supervisions, denoted as Newly-annotated boxes, we conducted an experiment in Table III. Our findings indicate that the newly-annotated boxes alone can enhance the student’s performance.

To corroborate our observation, we further conducted an experiment that exclusively utilized the newly-annotated boxes for training the student.
Moreover, the confidence scores of the teacher’s boxes can encapsulate the teacher’s knowledge regarding the quality of the projected boxes. To this end, we replace the confidence scores with the IoUs of the teacher’s boxes to their closest GT boxes. For those teacher’s boxes that do not intersect with any GT box, we maintain their confidence scores. As revealed in Table III, after replacing the confidence scores, the performance gain decreases by 0.2%, indicating that teachers can evaluate the quality of predicted boxes more effectively, and utilizing IoU scores is not optimal.

**Intrinsic Characteristics of Objects:** Given that teacher detectors are trained using GT boxes as a reference, they are able to capture not only the machine-perceived object locations but also some intrinsic characteristics of the objects, including their level of detection difficulty. Fig. 3 illustrates that when we employ the AP value of a category to indicate its detection difficulty, the lower the AP value, the greater the ratio of auxiliary boxes to GT boxes. This suggests that objects that are easy to detect, such as the “mouse” depicted in Fig. 2, are more likely to have only one auxiliary box, whereas more challenging objects may have multiple auxiliary boxes that are more widely dispersed.

Consequently, during training, the student detector prioritizes losses associated with those harder objects that have a greater number of auxiliary boxes. As demonstrated in Fig. 3(c), the Teach-DETR method substantially improves the category APs of these more challenging categories. As shown in Table III, if we simultaneously replace the labels and the confidence scores as above, the auxiliary information from teachers can only provide the box positions. Even if, it still achieves 0.6% gain over the baseline model, demonstrating that the positions of auxiliary boxes alone is an informative source for supervising the student DETR.

### IV. Experiments

#### A. Setup

**Dataset:** We showcase the efficacy of Teach-DETR on the COCO 2017 dataset [15]. Similar to most prior approaches, we fine-tune the model on the training set and present the mean average precision (AP) on the val set with IoU thresholds ranging from 0.5 to 0.95, with a step size of 0.05. We also provide an optional report on the AP of diverse object scales, comprising small (AP$_S$), medium (AP$_M$), and large (AP$_L$). To assess the generalizability of Teach-DETR, we present the results of its compatibility on the LVIS v1.0 [37], using mean average precision as the evaluation metric.

**Implementation Details:** To ensure a fair comparison, we directly adopt the training settings of the student detectors, such as batch size, learning rate, optimizer, etc.

The auxiliary boxes from teachers are obtained in an offline manner. Since the efficiency of bipartite matching significantly affects the Teach-DETR’s efficiency, we limit the number of auxiliary boxes per teacher to below 50 by selecting those with high confidence scores, particularly for LVIS v1.0 [37], as the detectors tend to generate numerous low confident boxes (below 0.01) for each image. Even if we use the confidence scores to weigh the losses of auxiliary boxes, too many low-confident boxes are still harmful.

| Method                          | AP   | AP$_{50}$ | AP$_{75}$ |
|---------------------------------|------|-----------|-----------|
| Student                         | 52.5 | 71.1      | 57.3      |
| Student + Teach-DETR            | 53.5 | 72.0      | 58.2      |
| Newly-annotated boxes           | 52.7 | 71.5      | 57.9      |
| Replace labels with GT labels   | 53.2 | 71.5      | 58.2      |
| Replace scores with IoUs        | 53.3 | 71.6      | 58.5      |
| Replace labels & scores         | 53.1 | 71.3      | 58.2      |

Table III: Results on COCO val set for validating the effectiveness of the auxiliary supervisions.
For $\mathcal{H}$-Deformable-DETR [13], we apply auxiliary supervisions to the one-to-one matching branch only, as using them for the one-to-many matching branch does not improve the results. For DN-DETR [21] and DINO [1], we use auxiliary boxes only for the matching part.

Teacher Detectors: For COCO, we test four teacher detectors, i.e., Mask RCNN [14], RetinaNet [5], FCOS [39] and $\mathcal{H}$-Deformable-DETR [13]. These detectors are designed in different manners, including the two-stage and ROI-based detector, one-stage detector, anchor-free detector, and DETR-based detector. The detection results of these teachers are shown in Table IV. Besides, the performances of the three used teacher detectors in LVIS v1.0 [37] are shown in Table V.

### B. Ablation Study

In this section, we aim to explore diverse approaches to utilizing auxiliary supervisions. We present our findings based on the COCO 2017 val1.0 dataset, using $\mathcal{H}$-Deformable-DETR [13] as the student model with Swin-Small Backbone, 300 queries, and 12-epoch training schedule. Unless otherwise specified, we leverage the auxiliary supervisions derived from Swin-Small Mask RCNN [14] with a 36-epoch training schedule. The respective performances are presented in Table IV. It is worth noting that we solely employ the auxiliary supervisions in the one-to-one matching branch of $\mathcal{H}$-Deformable-DETR [13]. Our experiment reveals that utilizing auxiliary supervisions for the one-to-many matching branch brings no further improvement.

Teacher’s Boxes versus Noisy Boxes: As demonstrated in Section III-B, the utilization of solely the teacher’s box positions has shown notable improvements. However, an intriguing question arises: could incorporating manually generated boxes as auxiliary supervisions augment performance even further? In this study, we take inspiration from DN-DETR [21] and introduce noise to the GT boxes to create a set of noisy boxes. We generate three groups of noisy boxes, with hyper-parameters set to 0.4 for both center shifting and box scaling. These noisy boxes retain the same data format as the teacher’s boxes, including size, location, category, and confidence scores. However, unlike the teacher’s boxes, they do not contain any teacher information. Instead, the confidence scores are manually set as the IoUs to GT boxes.

Table VI presents a comparison of different designs. Surprisingly, unlike the teacher’s boxes, incorporating the noisy boxes actually results in a performance decline upon the baseline model. This highlights the indispensability of utilizing auxiliary boxes provided by teachers.

Offline Teachers versus Online Teachers: According to the results presented in Table VI, the use of online teachers yields comparable results to offline teachers. However, it should be noted that the online approach incurs a significant increase in both training memory and computational cost. Therefore, to mitigate these challenges, we opt to leverage offline teachers to generate the necessary supervision boxes, which are then loaded from memory during the training phase.

Hard Labels versus Soft Labels: In Teach-DETR, we leverage the hard labels of the predicted bounding boxes, which correspond to the category ID with the highest probability. As illustrated in Table VI, employing soft labels, i.e., predicted class probabilities of boxes, results in a significant decline in detection performance. This phenomenon arises because RCNN-based detectors, like Mask RCNN [14], generally adopt the softmax operation and cross-entropy loss. In contrast, DETR-based detectors prefer the sigmoid operation and binary cross-entropy loss or focal loss [5]. Aligning the predictions of these two models is challenging. On the contrary, hard labels can avoid such inconsistencies.

Utilizing Auxiliary Boxes of Different IoUs: In Section III-B, we demonstrated that incorporating the newly-annotated boxes alone can lead to performance gains over the baseline detector. Here, we delve deeper into the effects of different auxiliary boxes and categorize them based on their Intersection over Union (IoU) scores with respect to the ground truth (GT) boxes. For Swin-S Mask RCNN [14], predicted boxes with IoU scores of 0.0, (0.0, 0.5), and (0.5, 1.0) with respect to GT boxes constitute roughly 15%, 51%, and 34% of the total number of auxiliary boxes, respectively.

To determine which IoU ranges of boxes contribute most to the DETR student, we partition the auxiliary boxes into three groups based on their IoU scores with the GT boxes: boxes with IoU scores greater than 0.1, 0.3, and 0.5. The results are presented in Fig. 4. It is evident that the most significant improvement is achieved using boxes that have significant overlap with GT boxes. Specifically, employing auxiliary boxes with IoU scores

| Method          | Backbone | #epoch | AP  |
|-----------------|----------|--------|-----|
| Mask RCNN [14]  | Swin-S   | 36     | 40.7|
| RetinaNet [5]   | Swin-S   | 36     | 46.4|
| FCOS [39]       | Swin-S   | 36     | 49.4|
| $\mathcal{H}$-Deformable-DETR [13] | Swin-S | 12 | 52.5|
| $\mathcal{H}$-Deformable-DETR [13] | Swin-L | 12 | 55.9|
| DINO [1]        | Swin-L   | 12     | 57.3|

Swin-Large is pretrained on the ImageNet-22k [38].
greater than 0.3 yields the same performance as utilizing all auxiliary boxes. However, incorporating auxiliary boxes with IoU scores between 0.1 and 0.3 leads to a drop in performance, indicating that such auxiliary boxes are potentially noisy. Tightening the restriction further to only leverage auxiliary boxes with IoU scores greater than 0.5 results in a substantial drop in performance of 0.4%, likely due to the insufficient number of boxes to stabilize training and provide complementary information from the teachers.

Given that utilizing all auxiliary boxes yields the best performance and to avoid introducing additional hyper-parameters, we utilize all auxiliary boxes in our approach and subsequent experiments.

**Utilizing Auxiliary Boxes of Different Scores:** In addition to the Intersection over Union (IoU) metric, the confidence score assigned to a bounding box can also serve as a valuable indicator of its quality. As illustrated in Fig. 5, we apply varying thresholds to remove low-confidence boxes generated by the Swin-Small Mask RCNN. Given that the auxiliary boxes undergo non-maximum suppression, they already exhibit high confidence scores. Consequently, additional culling based on confidence scores does not necessarily yield superior performance. Instead, this approach may inadvertently discard useful auxiliary boxes, such as those that have been recently annotated, resulting in performance degradation when the threshold is too restrictive.

**Applying Auxiliary Supervisions to Different Layers:** In the two-stage architecture of \( H \)-Deformable-DETR [13], the GT boxes are utilized at the Transformer encoder output and the output queries of both the final layer and intermediate layers of the decoder. In this study, we explore the potential benefits of extending the use of auxiliary boxes to these layers.

Table VII reveals that incorporating auxiliary supervisions to all three types of layers can lead to improved performance, particularly when applied to the intermediate layers of the Transformer decoder. In light of the computational efficiency of leveraging auxiliary boxes, we adopt this strategy across all layers to achieve optimal results.

**Detection Performance of Teacher Detectors:** As previously mentioned, the use of teacher detectors for enhancing detection performance is not a novel concept in knowledge distillation. Nonetheless, our experimental results illustrate that our approach diverges from knowledge distillation as the student detectors can outperform the teacher detectors.

To ascertain whether the quality of the teacher detector impacts the performance of the DETR student, we conducted a series of experiments in Table VIII(a). Specifically, we focused on teacher detectors with the same architecture but differing detection capabilities, achieved through modifications to their backbones. We utilized the Mask RCNN [14] with three different backbones, namely ResNet50, Swin-Small, and Swin-Large. Despite some performance enhancement when using the R50 Mask RCNN as the teacher, there still exists a significant performance gap compared to using the auxiliary supervision of Swin-Small Mask RCNN. Additionally, adopting teacher boxes from Swin-Large Mask RCNN only resulted in comparable performance to that of Swin-Small Mask RCNN. These results suggest a correlation between the teacher’s performance and the improvement in DETR performance. However, the amount of information available from boxes limits the extent of this enhancement.
We then investigate the impact of teacher architectures in Table VIII(b). In addition to the Mask RCNN [14], we select RetinaNet [5], FCOS [39], $\mathcal{H}$-Deformable-DETR [13] and DINO [1] as teachers. First, we observe that the performance of the teacher detectors’ architecture does not impact DETR performance when the teacher’s detection capability was sufficiently high. Second, even when using teachers with the same architecture as the student, such as $\mathcal{H}$-Deformable-DETR, the performance improvement was still evident. Third, when we employ much stronger teacher detectors than the student, such as Swin-Large $\mathcal{H}$-Deformable-DETR and Swin-Large DINO, the performance gain is incremental compared to the results with those moderately-capable teachers. The conclusion is similar to that derived from Table VIII(a), predictions generated by a strong detector align closely with the ground truth (GT) boxes, thereby constraining supplementary information provided from auxiliary boxes, available online. Furthermore, bounding boxes have inherent limitations in their ability to provide comprehensive insights sourced from teachers.

Finally, we assess the expandability of our approach in Table VIII(c). Building on the Mask RCNN, we gradually incorporated the auxiliary supervision of RetinaNet, FCOS, and $\mathcal{H}$-Deformable-DETR. Our results indicate that the auxiliary supervisions from different teachers can complement each other. Introducing more teachers consistently improves the performance of the student, resulting in an AP of 54.2%.

Early Box Fusion by NMS: In our method, we utilize the auxiliary boxes from multiple teachers in a parallel and independent manner. Here, we employ an “early” fusion approach by means of Non-Maximum Suppression (NMS) to fuse the boxes from multiple teachers, forming a single group of auxiliary supervision. In Fig. 6, we illustrate the results obtained through different hyper-parameter configurations of NMS while fusing the boxes from two teachers, namely Swin-S Mask RCNN and Swin-S RetinaNet.

Initially, we examine the effects of utilizing different NMS thresholds (indicated by the red line). As evident from the results, the performance is highly sensitive to the NMS thresholds. Specifically, when the threshold is too relaxed, the presence of numerous redundant auxiliary boxes leads to unfavorable performance degradation. To overcome the distribution difference of confidence scores between the two teachers’ boxes, we opt to fix the NMS threshold at 0.5 and introduce an offset to the confidence scores of RetinaNet’s boxes (represented by the blue dashed line). However, the addition of an offset fails to yield any improvement. The above experiments demonstrate the arduous task of fine-tuning hyper-parameters for NMS. This issue becomes even more pressing when we employ four teachers, as showcased in Table VIII. Despite extensive hyper-parameter tuning, we achieve an AP of 53.9%, which is inferior to our solution.

Self-Taught Teach-DETR: Table VIII presents a noteworthy finding: the Swin-$\mathcal{H}$-Deformable-DETR, a well-trained teacher with the same architecture as the student, can offer valuable auxiliary boxes. This observation raises a crucial question: can the same detector under various training stages enhance the proposed strategy? Fig. 7 showcases the use of Swin-$\mathcal{H}$-Deformable-DETR at epochs 2, 5, 11, and 12 as teacher models, and the results demonstrate the potential benefit of different training stages for our method. Additionally, consistent with previous findings, superior teachers consistently yield superior students.

Building on the above results, we explore the utilization of the mean teacher approach [40], which employs exponential moving average with a momentum of 0.999, to provide auxiliary supervision in an online fashion. Remarkably, this approach yields a 0.5% improvement, underscoring the effectiveness of our proposed method.

C. Analyses

Teach-DETR Can Improve Stability of DETR: In Fig. 8, we employ the DN-DETR algorithm [21] to compute the instability score (IS) of bipartite matching during the training process. Specifically, we record the matching outcomes by assigning an index to each positive query, indicating the corresponding ground-truth (GT) box, while labeling negative queries with index $-1$, signifying the “no object” category. The IS is determined by calculating the difference between the matching results of two consecutive epochs. It is worth noting that incorporating auxiliary boxes from teachers, denoted as $\mathcal{H}$-Def + Aux in Fig. 8, reduces the IS of $\mathcal{H}$-Deformable-DETR [13] by 0.2. When considering the auxiliary boxes, we determine that an object query is matched with GT box $i$ if the intersection over union (IoU) between the auxiliary box matched by the query and GT box $i$ is greater than 0.5. This approach significantly reduces the IS. Conversely, although utilizing noisy boxes reduces the IS in the early stages, as represented by $\mathcal{H}$-Def + Noisy $\dagger$, its
Fig. 8. Instability scores (IS) \cite{21} of various DETRs on COCO 2017 training set. These detectors are with same settings of R50 backbone, 12-epoch training schedule and auxiliary boxes from R50 Mask RCNN (if applicable). If the IoU between the auxiliary box matched by an object query and the GT box $i$ is greater than 0.5, we consider the object query is matched with the GT box $i$.

Fig. 9. Performance curves of Swin-Small H-Deformable-DETR \cite{13}, Swin-Large DINO \cite{1}, and them with auxiliary supervisions. The auxiliary boxes are collected from Swin-S Mask RCNN \cite{3}, Swin-S FCOS \cite{39}, Swin-S RetinaNet \cite{5}, and Swin-S H-Deformable-DETR \cite{13}.

The performance curves of state-of-the-art object detectors, namely Swin-Small H-Deformable-DETR \cite{13} and Swin-Large DINO \cite{1}, along with two detectors augmented with auxiliary boxes from four teachers are presented in Fig. 9. The negligible increase in training time induced by (AP) and total training time (in hours) of various models. When we employ Swin-Small H-Deformable-DETR (1x) as the teacher, the Swin-Small H-Deformable-DETR student achieves impressive AP scores of 54.4\% and 55.0\% at 24 epochs and 36 epochs, respectively. These performances come at a training cost of approximately 57.3 hours and 76.5 hours, respectively. Notably, these training durations are similar to those of the original Swin-Small H-Deformable-DETR, which achieved AP scores of 54.2\% at 36 epochs (56.7 hours) and 54.0\% at 48 epochs (75.6 hours). Therefore, taking into account the training time required for teachers, Teach-DETR also proves itself capable of enhancing detection performance under a comparable training time.

Notably, Teach-DETR requires more teachers to reach higher detection results. For example, when utilizing four different teachers, the Swin-Small H-Deformable-DETR can obtain AP of 54.2\% at 12 epochs and takes almost 195 hours for training. There exists a trade-off between the performance gain and training efficiency. However, it should be noted that existing DETR-oriented distillation methods do not support multi-teacher distillation. Moreover, compared with existing distillation methods, such as DETRDistill \cite{33}, which requires more than 80 hours to distill between Swin-Small H-Deformable-DETR for 36 epochs, Teach-DETR demands only 57.3 hours, which is more training efficient.

Performance Curve: The performance curves of state-of-the-art object detectors, namely Swin-Small H-Deformable-DETR \cite{13} and Swin-Large DINO \cite{1}, along with two detectors augmented with auxiliary boxes from four teachers are presented in Fig. 9. The negligible increase in training time induced by
that the incorporation of Teach-DETR has a positive impact on Swin-Small $H$-Deformable-DETR’s detection performance, enhancing its ability to make correct decisions. This improvement is achieved by leveraging the knowledge from the teacher detectors, which serve as a reference for the student detector. Such an approach is highly effective in addressing the challenges associated with object detection tasks, especially in scenarios where accurate detection is critical.

D. Compatibility With State-of-the-Art DETRs

In Table IX, we begin by verifying the compatibility of Teach-DETR on the COCO 2017 val dataset [15] using several typical DETR-based detectors. Specifically, we report results on three DETR-based detectors, namely Conditional DETR [18], DAB-DETR [20], and DN-DETR [21], as well as four Deformable-DETR-based methods with two-stage settings, including the vanilla Deformable-DETR [17], Deformable-DETR + tricks [1], [13], DINO [1], and $H$-Deformable-DETR [13]. In the case of $H$-Deformable-DETR [13], we report the results based on different backbones and query numbers.

The results show that Teach-DETR consistently improves baseline detectors, achieving a 2.6% gain on DN-DETR [21], which includes a box denoising task. For DINO [1], one of the top-performance DETR-based detectors, Teach-DETR can improve its 4-scale feature pyramid model by 1.1%. Additionally, for some lightweight backbones such as MobileNet-v2 [43] and ResNet18, Teach-DETR can significantly improve their performance by over 2.0%. As the capability of the backbones increases, Teach-DETR demonstrates its effectiveness in boosting performance, even with the Swin-Large backbone pre-trained on ImageNet-22k, achieving a 0.9% boost for $H$-Deformable-DETR [13].

In addition, we further evaluate the compatibility of Teach-DETR based on two atypical DETR-based detectors, namely YOLOS [41] and ViDT [32]. YOLOS [41] extends the Vision Transformer (ViT) to object detection by combining the backbone and the Transformer encoder to form an efficient architecture. Its decoder is a lightweight architecture, making it an encoder-only DETR-based detector. On the other hand, ViDT [32] integrates the backbone and the Transformer encoder together but retains the Transformer decoder, thus qualifying as an encoder-free DETR-based detector. Despite their atypical architectures, Teach-DETR consistently improves the performance of both detectors, demonstrating its universality.

Furthermore, to further validate the effectiveness of Teach-DETR, in Table X, we evaluate its performance on the LVIS v1.0 dataset [37] using three representative DETRs, including DAB-DETR [20], DN-DETR [21], and $H$-Deformable-DETR [13]. Similarly, Teach-DETR significantly improves the performance of these DETRs, showcasing its powerful generalization ability.

E. Comparison With State-of-the-Art Methods

In Table XI, we compare Teach-DETR with two concurrent methods DETRDistill [33] and KD-DETR [35]. As evidenced by our results, our proposed method yields a 2.4% improvement...
TABLE IX
RESULTS OF DIFFERENT DETR WITH PROPOSED TEACH-DETR ON COCO 2017 val.1, SET

| Method                        | Backbone | #query | #epochs | AP   | AP50 | AP75 | APs  | APtd | APed |
|-------------------------------|----------|--------|---------|------|------|------|------|------|------|
| Conditional-DETR-DC5 [18]     | R101     | 9      | 300     | 45.0 | 65.9 | 48.4 | 26.1 | 46.8 | 62.8 |
| Conditional-DETR-DC5 + TD     | R101     | 300    | 46.7 (~1.7) | 66.9 | 50.8 | 28.3 | 50.9 | 63.5 |
| DAB-DETR-DC5 [20]             | R101     | 300    | 45.8 (~2.7) | 68.1 | 52.8 | 30.5 | 52.6 | 64.4 |
| DAB-DETR-DC5 + TD             | R101     | 300    | 47.3        | 67.5 | 50.8 | 28.6 | 51.5 | 65.0 |
| DN-DETR-DC5 [21]              | R101     | 300    | 49.9 (~2.6) | 69.5 | 54.2 | 31.7 | 53.8 | 66.7 |
| YOLOv5 [41]                   | Det-TS [42] | 100   | 150     | 35.6 | 53.9 | 36.0 | 14.3 | 39.0 | 54.9 |
| YOLOv5 + TD                    | Det-TS [42] | 100   | 150     | 38.0 (~2.4) | 58.4 | 39.5 | 17.4 | 40.1 | 56.3 |
| ViTDE [32]                    | Swin-S   | 100    | 50     | 47.2 | 67.5 | 51.1 | 28.7 | 50.1 | 64.3 |
| ViTDE + TD                     | Swin-S   | 100    | 50     | 49.0 (~1.8) | 68.7 | 53.4 | 32.1 | 52.0 | 64.9 |
| Deformable-DETR [17]          | Swin-S   | 300    | 36     | 50.7 | 70.7 | 54.8 | 32.4 | 54.3 | 67.2 |
| Deformable-DETR + TD          | Swin-S   | 300    | 36     | 53.2 (~2.5) | 72.3 | 56.3 | 37.3 | 56.9 | 68.5 |
| Deformable-DETR [17] + tricks | Swin-S   | 300    | 36     | 53.8 | 72.8 | 56.9 | 36.5 | 57.5 | 69.0 |
| Deformable-DETR + tricks + TD | Swin-S   | 300    | 36     | 55.5 (~1.7) | 74.2 | 61.1 | 40.1 | 59.4 | 70.5 |
| DINO+ [1]                     | Swin-L (IN-22k, 384) | 900   | 36     | 57.8 | 76.5 | 63.2 | 40.6 | 61.8 | 75.9 |
| DINO+ + TD                    | Swin-L (IN-22k, 384) | 900   | 36     | 58.9 (~1.1) | 77.4 | 65.0 | 42.7 | 62.9 | 74.9 |
| H-Deformable-DETR [13]        | MobileNetv2 | 300   | 36     | 44.4 | 61.4 | 48.6 | 28.4 | 47.2 | 57.2 |
| H-Deformable-DETR + TD        | MobileNetv2 | 300   | 36     | 46.5 (~2.1) | 63.1 | 50.0 | 29.8 | 48.2 | 57.9 |
| H-Deformable-DETR [13]        | R18      | 300    | 45.4        | 62.2 | 49.5 | 29.2 | 48.0 | 58.8 |
| H-Deformable-DETR + TD        | R18      | 300    | 47.6 (~2.2) | 64.2 | 51.9 | 30.6 | 49.1 | 59.3 |
| H-Deformable-DETR [13]        | R50      | 300    | 50.0 | 63.8 | 54.4 | 32.9 | 52.7 | 65.3 |
| H-Deformable-DETR + TD        | R50      | 300    | 51.9 (~1.9) | 70.1 | 57.0 | 35.2 | 54.8 | 66.3 |
| H-Deformable-DETR [13]        | Swin-S   | 300    | 54.2        | 73.0 | 59.1 | 36.8 | 57.9 | 69.6 |
| H-Deformable-DETR + TD        | Swin-S   | 300    | 55.8 (~1.6) | 74.3 | 61.4 | 39.0 | 59.8 | 69.9 |
| H-Deformable-DETR [13]        | Swin-L (IN-22k) | 300   | 36     | 57.1 | 75.6 | 62.6 | 40.3 | 61.0 | 72.8 |
| H-Deformable-DETR + TD        | Swin-L (IN-22k) | 300   | 36     | 58.0 (~0.9) | 76.6 | 66.9 | 42.0 | 62.0 | 73.4 |
| H-Deformable-DETR [13]        | Swin-L (IN-22k) | 300   | 36     | 57.6 | 76.3 | 63.4 | 41.3 | 61.9 | 73.7 |
| H-Deformable-DETR + TD        | Swin-L (IN-22k) | 300   | 36     | 58.5 (~0.9) | 77.4 | 64.8 | 42.5 | 62.5 | 73.8 |

The auxiliary boxes are collected from Swin-S Mask RCNN [3], Swin-S FCOS [39], Swin-S RetinaNet [5] and Swin-S H-Deformable-DETR [13], whose performances are shown in Table IV. All the Deformable-DETR-based methods are two-stage. Weight decay of all H-Deformable-DETRs is 0.0001. ~ Tricks denotes dropout rate 0 within transformer, mixed query selection and look forward twice [1], [13]. * using top 300 predictions for evaluation. TD is short of Teach-DETR.

TABLE X
COMPARISON OF RESULTS WITH TWO CONCURRENT METHODS ON LVIS V1.0 [37]

| Method                        | Backbone       | #epochs | AP   |
|-------------------------------|----------------|---------|------|
| DAB-DETR-DC5 [20]             | R101           | 50      | 23.7 |
| DAB-DETR-DC5 + TD             | R101           | 50      | 25.6 |
| DN-DETR [13]                  | R101           | 50      | 30.3 |
| DN-DETR + TD                  | R101           | 50      | 32.1 |
| H-Deformable-DETR [13]        | Swin-S         | 36      | 35.7 |
| H-Deformable-DETR [13]        | Swin-S         | 36      | 37.0 |

Both of two detectors use 300 queries. TD is short of Teach-DETR.

TABLE XI
COMPARISON OF DETECTION PERFORMANCE ON COCO WITH TWO KNOWLEDGE DISTILLATION METHODS

| Method                        | AP   | AP50 | AP75 | Training Time |
|-------------------------------|------|------|------|---------------|
| T: R101-Deformable-DETR       | 44.6 | 63.7 | 48.5 | -             |
| S: R50-Deformable-DETR        | 43.8 | 62.6 | 47.7 | 61            |
| DETRDistill[33]               | 46.6 | 65.6 | 50.7 | 102           |
| KD-DETR [35]                  | 46.1 | 65.6 | 50.6 | 95            |
| Teach-DETR                   | 45.2 | 64.7 | 50.6 | 66            |

The student is trained with 50 epochs. * The teachers used in Teach-DETR are the same as those in Table IX.

for R50 Deformable-DETR. While this is marginally lower than the performance of DETRDistill [33] and on par with that of KD-DETR [35], it’s important to note that these two methods are tailored for knowledge distillation between DETRs with identical architectures but different backbones. In contrast, our approach is capable of transferring knowledge from either RCNN-based or DETR-based detectors, thereby enhancing the performance of DETRs in a more versatile manner. We also conducted evaluations on the training times (minutes/epoch) of the aforementioned methods with 8 Tesla A100 GPUs. In this regard, our method offers a significant advantage, thanks to the minimal computation cost of bipartite matching and the offline acquisition of auxiliary boxes, which do not increase training time. In contrast, DETRDistill [33] and KD-DETR [35] incur additional time costs for the teacher forward operation and the use of additional queries for the student’s decoder. Taken together, compared to other methods, Teach-DETR stands out for its versatility and superior training efficiency.

In Table XII, we compare Teach-DETR with KD methods designed for RCNN-based detectors. Despite leveraging the strong Swin-L H-Deformable-DETR as the teacher, these approaches lead to comparable or even degraded performance compared to our method.

V. CONCLUSION

In this paper, we introduce Teach-DETR, a novel approach for enhancing the performance of DETR-based object detectors by leveraging knowledge from pre-trained teacher detectors.
We demonstrate that the predicted boxes generated by teacher detectors serve as a valuable medium for transferring knowledge, thereby facilitating the training of more accurate and robust DETR models. Our approach allows for the integration of predicted boxes from multiple teacher detectors, each of which provides parallel supervision to the student DETR. We properly weigh the matchings based on the teachers’ confidence scores for the predicted boxes. Furthermore, during training, Teach-DETR incurs negligible computational costs, memory usage, and imposes no constraints on the teacher architectures, making it adaptable to various types of teacher detectors. Extensive experiments illustrate that Teach-DETR consistently enhances the performance of various DETR-based object detectors. Overall, our proposed approach provides a robust and generalizable solution for improving object detection with minimal additional computational, memory overhead and architecture requirements.

REFERENCES

[1] H. Zhang et al., “DINO: DETR with improved denoising anchor boxes for end-to-end object detection,” 2022, arXiv:2203.03605.

[2] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 10012–10022.

[3] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. 28th Int. Conf. Neural Inf. Process. Syst., 2015, pp. 91–99.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 779–788.

[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2017, pp. 2908–2918.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in Proc. Eur. Conf. Comput. Vis., Springer, 2020, pp. 213–229.

[7] A. Vaswani et al., “Attention is all you need,” in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 6000–6010.

[8] H. W. Kuhn, “The Hungarian method for the assignment problem,” Nav. Res. Logistics Quart., vol. 2, no. 1/2, pp. 83–97, 1955.

[9] G. Hinton et al., “Distilling the knowledge in a neural network,” 2015, arXiv:1503.02531.

[10] L. Zhang and K. Ma, “Improved object detection with feature-based knowledge distillation: Towards accurate and efficient detectors,” in Proc. Int. Conf. Learn. Representations, 2020.

[11] S. Yuan et al., “Distilling object detectors via decoupled features,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 4643–4652.

[12] D. Jia et al., “DETRs with hybrid matching,” 2022, arXiv:2207.13080.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2017, pp. 2961–2969.

[14] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis., Springer, 2014, pp. 740–755.

[15] P. Gao, M. Zheng, X. Wang, J. Dai, and H. Li, “Fast convergence of DETR with spatially modulated co-attention,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 3621–3630.

[16] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable DETR: Deformable transformers for end-to-end object detection,” in Proc. Int. Conf. Learn. Representations, 2021.

[17] D. Meng et al., “Conditional DETR for fast training convergence,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 3651–3660.

[18] Y. Wang, X. Zhang, T. Yang, and J. Sun, “Anchor DETR: Query design for transformer-based detector,” in Proc. 36th AAAI Conf. Artif. Intell., 2022, pp. 2567–2575.

[19] S. Liu et al., “DAB-DETR: Dynamic anchor boxes are better queries for DETR,” in Proc. Int. Conf. Learn. Representations, 2022.

[20] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, “DN-DETR: Accelerate DETR training by introducing query denoising,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 13619–13627.

[21] Q. Chen, X. Chen, G. Zeng, and J. Wang, “Group DETR: Fast training convergence with decoupled one-to-many label assignment,” 2022, arXiv:2207.13085.

[22] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, “Learning efficient object detection models with knowledge distillation,” in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 742–751.

[23] R. Sun, F. Tang, X. Zhang, H. Xiong, and Q. Tian, “Distilling object detectors with task adaptive regularization,” 2020, arXiv:2006.13108.

[24] Z. Zheng et al., “Localization distillation for dense object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 9407–9416.

[25] G. Li, X. Li, Y. Wang, S. Zhang, Y. Wu, and D. Liang, “Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation,” in Proc. 36th AAAI Conf. Artif. Intell., 2022, pp. 1306–1313.

[26] Q. Li, S. Jin, and J. Yan, “Mimicking very efficient network for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6556–6564.

[27] T. Wang, L. Yuan, X. Zhang, and J. Feng, “Distilling object detectors with fine-grained feature imitation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4933–4942.

[28] X. Dai et al., “General instance distillation for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 7842–7851.

[29] D. Zhixing et al., “Distilling object detectors with feature richness,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2021, pp. 5213–5224.

[30] L. Yao, R. Pi, H. Xu, W. Zhang, Z. Li, and T. Zhang, “G-DeKD: Towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 3591–3600.

[31] H. Song et al., “ViDT: An efficient and effective fully transformer-based object detector,” in Proc. Int. Conf. Learn. Representations, 2021.

[32] J. Chang, S. Wang, G. Xu, Z. Chen, C. Yang, and F. Zhao, “DETRdiss: A universal knowledge distillation framework for DETR-families,” 2022, arXiv:2211.10156.

[33] X. Chen, J. Chen, Y. Liu, and G. Zeng, “DETR: Decoder distillation for detection transformer,” 2022, arXiv:2211.09768.

[34] Y. Wang et al., “Knowledge distillation for detection transformer with consistent distillation points sampling,” 2022, arXiv:2211.08071.

[35] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[36] A. Gupta, P. Dollar, and R. Girshick, “LVIS: A dataset for large vocabulary instance segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 5351–5359.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248–255.

[38] Z. Tian, C. Shen, H. Chen, and T. He, “PCOS: Fully convolutional one-stage object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 9627–9636.

[39] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 1195–1204.

[40] Y. Fang et al., “You only look at one sequence: Rethinking transformer in vision through object detection,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2021, pp. 26183–26197.

[41] H. Trouvion, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, “Training data-efficient image transformers & distillation through attention,” in Proc. 35th Int. Conf. Mach. Learn., PMLR, 2021, pp. 10347–10357.

[42] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4510–4520.

[43] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “FitNets: Hints for thin deep nets,” 2014, arXiv:1412.6550.
Linjiang Huang received the BEng and MEng degrees from Huazhong University of Science and Technology, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences, in 2014, 2017 and 2020, respectively. He is currently a post-doctoral fellow with Multimedia Laboratory, The Chinese University of Hong Kong and CPI, Hong Kong. His main research interests include object detection, action recognition, machine learning, and computer vision.

Kaixin Lu is now working toward the undergraduate degree with Shanghai University. Her major research interests include machine learning, pattern recognition, and computer vision.

Guanglu Song received both the MEng degree from Beihang University, in 2017. He is now a researcher with SenseTime X-Lab. His major research interests include machine learning, pattern recognition, and computer vision.

Liang Wang (Fellow, IEEE) received both the BEng and MEng degrees from Anhui University, in 1997 and 2000, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2004. Currently, he is a full professor of the Hundred Talents Program with the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He is an IAPR fellow.

Si Liu (Senior Member, IEEE) received the PhD degree from the Institute of Automation, Chinese Academy of Sciences. She is currently a full professor with Beihang University. She has been a research assistant and postdoc with the National University of Singapore. Her research interests include computer vision and multimedia analysis.

Yu Liu received the PhD degree from The Chinese University of Hong Kong, in 2021. He is currently a senior director with SenseTime Group. His major research interests include machine learning, pattern recognition, and computer vision.

Hongsheng Li (Member, IEEE) received the BS degree in automation from the East China University of Science and Technology, and the MS and PhD degrees in computer science from Lehigh University, in 2006, 2010 and 2012, respectively. He is currently an associate professor with the Department of Electronic Engineering, The Chinese University of Hong Kong. His research interests include computer vision, medical image analysis, and machine learning.