

Distilling Knowledge from Large-Scale Image Models for Object Detection

Gang Li^{1,2†}, Wenhai Wang^{3,2}, Xiang Li⁴, Ziheng Li⁵, Jian Yang¹,
Jifeng Dai^{5,2}, Yu Qiao², and Shanshan Zhang^{1✉}

¹Nanjing University of Science and Technology ²OpenGVLab, Shanghai AI Laboratory

³The Chinese University of Hong Kong ⁴Nankai University ⁵Tsinghua University

Abstract. Large-scale image models have made great progress in recent years, pushing the boundaries of many vision tasks, *e.g.*, object detection. Considering that deploying large models is impractical in many scenes due to expensive computation overhead, this paper presents a new knowledge distillation method, which Distills knowledge from Large-scale Image Models for object detection (dubbed DLIM-Det). To this end, we make the following two efforts: (1) To bridge the gap between the teacher and student, we present a frozen teacher approach. Specifically, to create the teacher model via fine-tuning large models on a specific task, we freeze the pretrained backbone and only optimize the task head. This preserves the generalization capability of large models and gives rise to distinctive characteristics in the teacher. In particular, when equipped with DETection TRansformers (DETRs), the frozen teacher exhibits sparse query locations, thereby facilitating the distillation process. (2) Considering that large-scale detectors are mainly based on DETRs, we propose a Query Distillation (QD) method specifically tailored for DETRs. The QD performs knowledge distillation by leveraging the spatial positions and pair-wise relations of teacher’s queries as knowledge to guide the learning of object queries of the student model. Extensive experiments are conducted on various large-scale image models with parameter sizes ranging from 200M to 1B. Our DLIM-Det improves the student with Swin-Tiny by 3.1 mAP when the DINO detector with Swin-Large is used as the teacher. Besides, even when the teacher has 30 times more parameters than the student, DLIM-Det still attains a +2.9 distillation gain.

Keywords: Object detection · Knowledge distillation · Large-scale image model

1 Introduction

It is proved that scaling up model sizes is a promising direction to make neural networks reach human performance [33, 42, 48, 52]. Recently, great progress has been achieved in large-scale image models [11, 33, 34, 41], pushing performance boundaries of many vision tasks, including object detection, semantic segmentation, *etc.* These large-scale models demand for high performance computing device, making them hard to deploy in commonly used hardware platforms. Therefore, Knowledge Distillation (KD) is necessary

[†] This work was done while G. Li was an intern at Shanghai AI Laboratory.

✉ Corresponding author (shanshan.zhang@njust.edu.cn).

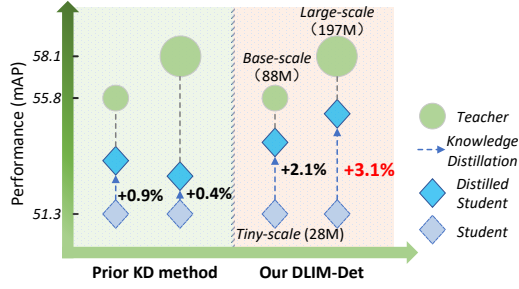


Fig. 1: Comparison between previous KD [3] and our DLIM-Det method for distilling *large-sized* teacher models. Previous methods struggle to distill knowledge from large models, whereas our method works well. Experiments are conducted on Swin Transformer [34].

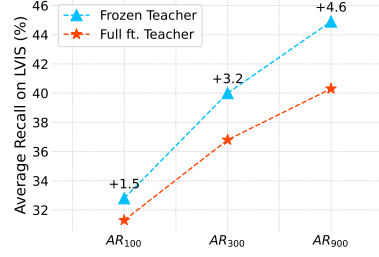


Fig. 2: Average Recall (AR) comparison on LVIS labels for the frozen teacher and fully fine-tuned teacher. Higher AR numbers indicate better generalization capacity.

to extract powerful and informative knowledge from the large model (teacher) to boost the compact model (student). The teacher knowledge serves as auxiliary supervisions and helps the student achieve better performance than trained only with ground-truth labels [21, 37]. However, existing KD methods [3, 8, 15, 50, 51] for object detection mainly focus on medium- or small-sized models, leaving the distillation from large-scale models rarely explored.

Distilling knowledge from large-scale image models is not trivial. As observed in previous works [35, 36], the discrepancy between the teacher and student becomes severe as the model size of teacher model increases, which hinders the effectiveness of KD methods. As shown in Fig. 1, when directly applying the prior KD method (*e.g.*, DETRDistill [3]) to a detector (*e.g.*, DINO [54]), the performance even degrades as the teacher model becomes stronger. Specifically, though the Swin-Large [34] has $2\times$ parameters than the Swin-Base, the latter is a better teacher and outperforms the former by +0.5 mAP in terms of KD gains. To address this problem, previous works [35, 39, 56] design a teacher assistant strategy, which generates multiple intermediate-sized teacher candidates to bridge the gap between the teacher and student. However, such a multi-step KD method induces additional training costs and severely affects the efficiency. In this paper, we propose that when obtaining the teacher model through fine-tuning the large-scale image backbone on a specific dataset, freezing the backbone and only optimizing the task head can empower the teacher with richer transferable knowledge compared to fully fine-tuning. Large-scale models exhibit increased generalizability due to their strong capacity and training on extensive datasets, however, there is a risk of overfitting when fine-tuned on specific datasets. Through freezing the pre-trained backbone, we maintain the generalizability of large-scale models, resulting in improved distillation performance. The teacher with pretrained large-scale backbone frozen is denoted as the *frozen teacher*.

To quantitatively assess the generalizability of both the frozen and fully fine-tuned (ft.) teachers, we compute the Average Recall (AR) on the LVIS annotations [16], and the results are presented in Fig. 2. Swin-Large [34] based DINO [54] serves as the teacher model. Notably, both teachers are trained only with supervision from the 80 object categories in COCO. The LVIS dataset, while sourcing images from COCO, features

a re-annotated and broader spectrum of 1203 categories. As results shown, the frozen teacher outperforms the fully ft. teacher by a margin of 4.6% in AR_{900} , indicating that the frozen teacher generates more generalizable features. Strengthening the generalizability of the teacher model endows it with specific characteristics that facilitate the distillation process. In particular, when equipped with detection transformers, sparse query locations emerge in the frozen teacher.

Advanced large-scale detectors [5, 41, 54, 58] are predominately based on detection transformers, where object queries absorb contextual information for each instance then output predictions. Object queries from the frozen teacher exhibit sparser distribution compared to those from the fully fine-tuned teacher. This sparsity characteristic enables the student model to circumvent unnecessary investments in modeling complex query relations, particularly those leading to excessively compact query locations. To fully exploit sparse query locations from the frozen teacher, we introduce a Query Distillation method. Our Query Distillation aligns object queries between the teacher and student models based on spatial positions and relations. This method consists of two key components: Query Position Distill (QPD) and Query Relation Distill (QRD). Initially, we establish a one-to-one set matching between object queries of two models. Subsequently, the proposed QPD allows the teacher’s queries to serve as auxiliary localization supervisions during student model training. Aside from QPD, we further propose Query Relation Distillation to lead the student model to mimic pair-wise relations between teacher’s queries. Combining Query Distillation and the frozen teacher can significantly alleviate the learning burden for the student, thereby contributing to a more efficient knowledge distillation process.

To conclude, we propose a new object detection knowledge distillation method, called **DLIM-Det**, tailored for Large-scale Image Models ranging from 200M to 1B parameters. Leveraging Query Distillation and the frozen teacher approach, DLIM-Det effectively transfers knowledge from large-scale models to compact ones. Extensive experiments on both the Transformer-based (Swin Transformer [34]) and CNN-based (InternImage [41]) models demonstrate the effectiveness of our approach. Results indicate our Query Distillation method works well for detection transformers, surpassing existing KD methods. Additionally, the frozen teacher approach bridges the gap between teacher and student, facilitating efficient distillation from large-scale models. Notably, when the teacher model is Swin-Large, DLIM-Det improves Swin-Tiny based DINO [54] by 3.1%; even for the giant InternImage-Huge model (1B+ parameters), it improves the InternImage-Tiny student by 2.9%, achieving a remarkable 56.1 mAP.

2 Related Work

Knowledge Distillation. As a classical model compression technique, Knowledge Distillation (KD) has been applied to many vision tasks [14, 17, 24, 25, 32, 47, 53]. In this paper, we focus on the KD for object detection. Using the intermediate features of the teacher as knowledge, feature distillation offers deep supervisions for the student [4, 15, 37, 51]. However, feature distillation is usually vulnerable to semantic misalignment between features of two models, and suffers from performance degradation when the discrepancy between two models is too large. Besides, to implement relation-based

distillation, previous works [8, 27, 49] design delicate modules or schemes to capture relations between different instances [8], anchors [27], patches [17], and pixels [49]. At last, the response-based KD methods are also widely investigated in previous works [8, 15, 22, 56], where the probabilistic scores are matched between two models. With the development of detection transformers (DETRs), some recently proposed KD methods [3, 7] are tailored for DETRs. DETRDistill [3] transfers the label assignment results from the teacher to student in order to stabilize the bipartite matching, while D3ETR [7] feeds the teacher queries to the student decoder and then trains the decoder using matching results from the teacher. Although these KD methods achieve decent distillation performance for medium-sized teacher models (such as ResNet-101 [20]), few of them explore the distillation from large-scale image models. In contrast to these works, we focus on distillation from large-scale models with parameters spanning from 200M to 1B. Through our method, we show that larger models lead to greater distillation gains.

Object Detection. Object detection is a fundamental task in the computer vision field [28, 46]. DETECTION TRANSFORMERS (DETRs) achieve great progress recently and gradually dominate object detection. With outstanding scalability, current mainstream large-scale detectors [5, 54, 58] typically rely on DETRs. To address the inefficiency issue of the original DETR [2], Deformable DETR [57] and Adamixer [13] introduce the local attention mechanism to replace the original global attention, resulting in faster model convergence. Some works [31, 43] attempt to insert the spatial prior into the object queries and lead each query to attend to a certain region. DN-DETR [26] designs a query denoising scheme to improve the stability of bipartite graph matching. Based on DN-DETR, DINO [54] presents a contrastive way for denoising training and achieves the state-of-the-art performance. Building on SoTA performance, we employ DINO as the default detector in this paper.

Large-scale Image Models. Some recent works [10, 33, 34, 41, 42, 48] scale up image models to reach 3 billion parameters, demonstrating a very powerful capacity of large-scale image models. Swin Transformer [33], built on the transformer architecture, incorporates techniques like post-norm and scaled cosine attention to enhance the stability of large vision models. Beit-3 [42], adopting the ViT [10] architecture, explores scaling up models through multi-modal pretraining. InternImage [41], utilizing a CNN base, integrates deformable convolution as a core operator to establish large-scale CNN models. After acquiring large-scale pretrained models, there has been significant interest in efficiently fine-tuning these models for downstream tasks. Notably, approaches such as VPT [23] insert learnable parameters into the input space while keeping most of the pretrained model frozen. Similarly, AdaptFormer [6] introduces lightweight learnable modules into the MLP block. Lin et al. [30] investigate the pretraining tasks that best align with the frozen setting. Different from these works, our paper aims to extract transferable knowledge from large models and use it to enhance tiny-scale models via the KD.

3 Method

In this section, we provide a concise overview of knowledge distillation and the detection transformer in Sec. 3.1. Subsequently, we introduce the frozen teacher method in Sec. 3.2, followed by an explanation of Query Distillation in Sec. 3.3.

3.1 Preliminary

Knowledge Distillation (KD). Standard KD enforces the student to align its probabilistic scores ($p^{(s)}$) with those of the teacher via the Kullback–Leibler (KL) divergence [21, 55]:

$$\mathcal{L}_{KD} = p^{(t)} \log\left(\frac{p^{(t)}}{p^{(s)}}\right). \quad (1)$$

The teacher’s logits ($p^{(t)}$) serve as soft label, providing non-target classes not present in the ground truth labels. Despite being weak, this non-target class information significantly enhances the student’s generalization. To amplify this effect, prior works typically apply a high temperature to soften the $p^{(t)}$.

Detection Transformers (DETRs). In DETRs, the backbone extracts image features, which are then inputted into the transformer encoder along with positional encodings. The enhanced image features serve as the key and value, providing contextual information for the subsequent transformer decoder. The input for the decoder is N embeddings, known as object queries, denoted as $Q = \{q_0, q_1, \dots, q_{N-1}\}$. Each query will be decoded into class probability p and box coordinate b via the transformer decoder.

3.2 Frozen Teacher Method

Inspired by the NLP domain, where scaling up language models achieves remarkable success [9, 12], some recent works scale up vision models to more than 1 billion parameters and set new records across various vision tasks [33, 41, 42, 48]. It is natural to explore the possibility of transferring the knowledge from these large models to more compact ones using the Knowledge Distillation (KD) technique. However, empirical observations (Fig. 1) suggest that as the model size of the teacher increases, the discrepancy between the teacher and student becomes more severe, limiting the effectiveness of the KD.

In this paper, we observed that when fine-tuning large-scale pretrained models on a specific dataset to create the teacher model, freezing the backbone can endow the teacher with richer transferable knowledge compared to fully fine-tuning. The overall framework of our DLIM-Det is illustrated in Fig. 3. The pretraining-then-finetuning paradigm is the cornerstone for modern neural networks [18, 19, 45], especially for large-scale models. Pretraining on massive data is essential for achieving top performance in the downstream tasks [41]. Although fully fine-tuning the entire model may enhance task-specific performance, it also heightens the risk of overfitting and diminishes the model’s generalizability. We denote the teacher produced by the frozen setting as the *frozen teacher*. This frozen teacher maintains the generalizability of the large-scale models. As evidenced by the Average Recall (AR) metrics on LVIS labels shown in Fig. 2, the frozen teacher exhibits an increased ability to recognize a broader array of non-target objects

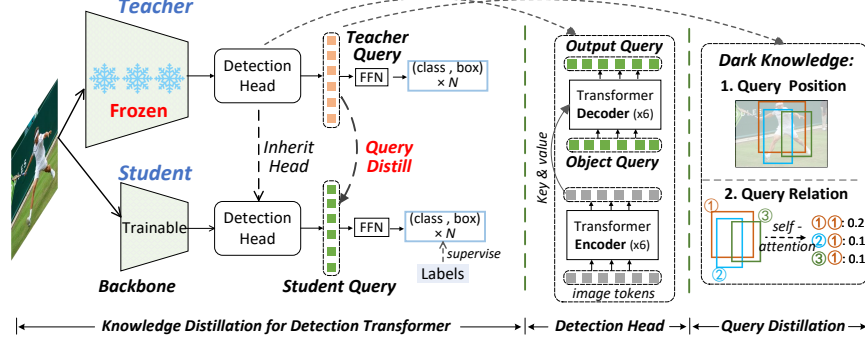


Fig. 3: The framework of our DLIM-Det. Freezing the backbone when fine-tuning large-sized models on a specific dataset to create the teacher. The alignment between object queries of the teacher and student is conducted based on positions and relations.

Table 1: The query density with different IoU thresholds.

Methods	IoU: 0.5		IoU: 0.7	
	Swin-L	Intern-L	Swin-L	Intern-L
Full finetuning	38.1	33.6	15.2	13.0
Frozen	21.3	21.8	7.5	8.3

while avoiding excessive specialization on the ground truths. Enhanced generalizability endows the teacher with specific characteristics that facilitate the distillation process.

In the following, we investigate how these distinctive characteristics enhance the efficiency of the distillation process. When equipped with DETRs, we compare the spatial distributions of object queries under both the frozen and fully fine-tuned settings. We observe that *the frozen teacher locates object queries more dispersely and sparsely* than the fully fine-tuned teacher. To quantify the dispersion of query spatial distribution, we introduce a metric called Query Density. Query density represents the number of object queries with an Intersection over Union (IoU) greater than τ surrounding each ground truth. We average the query density over all GT objects in the COCO 2017 validation set [29], with τ set to 0.5 and 0.7, respectively. The results are presented in Tab. 1. As results show, the backbone-frozen strategy significantly affects the query density, reducing it by at least 1/3. The fully fine-tuned teacher exhibits high query density values, and each GT object is surrounded by more than 30 queries with IoUs larger than 0.5. Among these compact queries, only one query should predict the object, while the others should predict no object. One example is visualized in Fig. 5. The detector needs to model their relations, identifying the optimal localized query with high confidence and suppressing the others with low confidence. In the case of sparse query locations, an instance is surrounded by fewer object queries, making their relations much easier to model.

When using the fully fine-tuned teacher, the student model needs to allocate unnecessary capacity to model complex query relations in order to mimic the compact query locations of the teacher. In contrast, the sparse query locations provided by the frozen teacher are easier for the student to learn, reducing the learning burden. To fully leverage the sparse query of the frozen teacher, we design a Query Distillation method, which is described in Sec. 3.3.

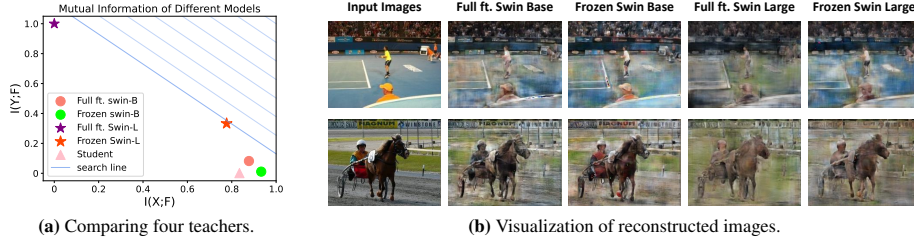


Fig. 4: Comparison of different teacher models based on mutual information. (a) Multiple teacher models and their mutual information concerning the input ($I(X; F)$) and target ($I(Y; F)$). The blue lines represent the process of searching for the model whose representation maximizes $I(X; F) + I(Y; F)$. “Full ft.” denotes full fine-tuning. (b) The reconstructed image qualitatively reflects $I(X; F)$.

Taking into account the aspect of generalization and drawing inspiration from prior research [40], which introduced an optimal teacher model selection algorithm based on mutual information, we also substantiate the superiority of the frozen teacher from the perspective of the mutual information principle. Specifically, the entropy of teacher’s representation ($H(F)$) can be formulated as:

$$H(F) = I(X; F) + I(Y; F) + Z, \quad (2)$$

$I(X; F)$ and $I(Y; F)$ represent task-relevant information regarding the input image X and ground truth target Y , respectively. Z refers to nuisance, which is task-irrelevant information. Following standard practice [40, 44], we employ the detection performance (*e.g.* mAP) to quantify $I(Y; F)$. To measure the mutual information concerning the input ($I(X; F)$), we feed the intermediate features¹ into a decoder network (typically, U-Net [38]) to reconstruct the input image X . The average reconstruction loss is used to quantify $I(X; F)$. Different values of $I(X; F)$ can be clearly distinguished based on the quality of reconstructed images, as illustrated in Fig. 4b. The theoretical proofs and training loss curves can be found in the Appendix.

After defining $I(X; F)$ and $I(Y; F)$, we compare base-scale and large-scale models under both full fine-tuning and frozen settings, with results depicted in Fig. 4a, where mutual information ($I(X; F)$ and $I(Y; F)$) is normalized to ensure the scale matching. Among the four teacher models, the frozen Swin-L attains the maximum of task relevant information $I(X; F) + I(Y; F)$, achieving a delicate trade-off between $I(X; F)$ and $I(Y; F)$. A model with high values of $I(X; F) + I(Y; F)$ can effectively suppress nuisance Z and serves as an informative teacher. Though Base-scale models exhibit higher values of $I(X; F)$ compared to Large-scale ones, they lag in terms of $I(Y; F)$, which may result in inferior KD performance. Empirical experiments (Tab. 3) support our assumption: the frozen Swin-L achieves the best distillation performance among the four teacher models. To conclude, the teacher model selection algorithm [40], grounded in mutual information, further supports our frozen teacher method.

¹ The features are from the last layer of Transformer Encoder in DETRs.

3.3 Query Distillation

Query Distillation aligns the spatial positions and pair-wise relations of queries between the teacher and student, consisting of Query Relation Distill and Query Position Distill. Initially, we establish the one-to-one matching for object queries of two models, where the cost matrix is computed based on classification and localization costs. Given the cost matrix, Hungarian Matching is used to solve the graph matching, yielding the optimal permutation function $\hat{\sigma}$:

$$\hat{\sigma} = \arg \min_{\sigma \in \mathfrak{S}_N} \sum_i^N \mathcal{L}_{match}(q_i^{(t)}, q_{\sigma(i)}^{(s)}), \quad (3)$$

where N equals the number of predicted objects. Among all permutations of N elements $\mathfrak{S}_N$, $\hat{\sigma}$ is the one with the lowest cost. Based on the class probability p and box coordinate b decoded from each object query, we define the cost between the teacher query $q^{(t)}$ and student $q^{(s)}$ as follows:

$$\mathcal{L}_{match}(q^{(t)}, q^{(s)}) = \alpha_1 \mathcal{L}_{KL}(p^{(t)}, p^{(s)}) + \alpha_2 \mathcal{L}_{L1Loss}(b^{(t)}, b^{(s)}) + \alpha_3 \mathcal{L}_{GIoU}(b^{(t)}, b^{(s)}), \quad (4)$$

where $\alpha_1, \alpha_2, \alpha_3$ are set to 2, 5, 2 empirically.

Query Position Distill. Instead of matching the probabilistic scores of two models, we observe that the spatial position of object queries can serve as informative knowledge. The position of an object query refers to the bounding box decoded from the query, represented by a vector: $b \in [x, y, w, h]$, where x and y denote the center coordinates, while w and h represent the width and height, respectively. Given the matching σ between object queries of two models, we formulate the procedure of Query Position Distill as follows:

$$\mathcal{L}_{QPD} = \sum_i^N \beta_1 \mathcal{L}_{L1}(b_i^{(t)}, b_{\sigma(i)}^{(s)}) + \beta_2 \mathcal{L}_{GIoU}(b_i^{(t)}, b_{\sigma(i)}^{(s)}), \quad (5)$$

where β_1, β_2 are set to 5 and 2, empirically. Through the thorough experiments (listed in Tab. 8b), we empirically find that besides the foreground queries, even poorly localized teacher queries that cover background regions still contain valuable knowledge to distill.

Due to the one-to-one label assignment, only a small subset of object queries receives localization supervision in each image. For example, on the COCO dataset [29], which typically contains an average of 7 instances per image, only 7 out of the 900 queries are involved in the regression loss. This sparsity of localization supervision impairs the training of detection transformers. By the proposed Query Position Distill, the spatial positions of teacher’s queries serve as additional supervisions to provide richer information than the sparse ground truth labels.

Query Relation Distill. In previous works, relation-based KD methods have been widely investigated. These methods aim to capture various relations between different entities such as instances [8, 22], pixels [49], anchors [27] or patches [17]. To model these relations, intricate modules, *e.g.*, GcBlock [1], and complex relation mechanisms, *e.g.* Rank Mimicking [27], have been designed, which not only lack interpretability but



Fig. 5: An example to show the position, score and relation of object queries. Query relations are obtained from self-attention. Only 3 queries are visualized for a clear presentation.

also introduce additional training costs. In contrast, the transformer decoder of DETRs explicitly models the pair-wise relations between object queries through the self-attention. We denote by $A = [a_{ij}]_{N \times N}$ the attention map from the self-attention, where N denotes the number of queries. L1 Loss is employed to narrow the distance between $A^{(t)}$ and $A^{(s)}$:

$$\mathcal{L}_{QRD} = \frac{1}{N} \sum_i \sum_j^N |A_{i,j}^{(t)} - A_{\sigma(i),\sigma(j)}^{(s)}|, \quad (6)$$

where σ denotes the permutation function to match the queries between the teacher and student. Query relations are critical for end-to-end detection. As shown in Fig. 5, precise query relations allow the best localized query to pop out with high classification scores, while the others will be suppressed with low scores.

The overall training loss $\mathcal{L}_{overall}$ is:

$$\mathcal{L}_{overall} = \gamma_1 \mathcal{L}_{task} + \gamma_2 \mathcal{L}_{QPD} + \gamma_3 \mathcal{L}_{QRD}, \quad (7)$$

where $\gamma_1, \gamma_2, \gamma_3$ are set to 1, 2, 1, respectively.

We also find that the parameters of the detection head in the teacher model serve as a source of knowledge. Even for heterogeneous DETRs, the basic blocks in their heads are consistent, including attention and feed forward network (FFN). Through empirical investigations, we prove that inheriting the parameters of these basic blocks from the teacher head to student head is convenient and beneficial.

4 Experiments

4.1 Experimental settings

Datasets and Backbone. We conduct the experiments on the MS COCO 2017 dataset [29], which is a commonly used object detection dataset. The dataset is split into the `train2017` set with 118k images and the `val2017` set with 5k images. All models are trained on the `train2017` set, and evaluated on the `val2017` set. Results are reported with two backbones: Swin Transformer [34] and InternImage [41]. We employ the COCO-style average precision (mAP) as the metric. DINO [54] is the default detector.

Implementation details. All models, except for the Intern-Huge model, are trained on $8 \times 32G$ V100 GPUs with a batch size of 16. In contrast, the Intern-Huge models are

Table 2: Effectiveness of each component.

Methods			Swin Transformer			InternImage		
			mAP	AP ₅₀	AP ₇₅	mAP	AP ₅₀	AP ₇₅
Teacher: Large Scale			58.1	77.1	63.5	56.9	75.8	61.9
Student: Tiny Scale			51.3	69.0	56.0	53.2	71.5	58.0
Query Distill	Frozen Teacher	Inherit Head						
✓			51.6 (+0.3)	69.5	56.4	54.2 (+1.0)	72.5	59.2
✓	✓		52.2 (+0.9)	70.1	57.0	54.7 (+1.5)	72.9	59.8
✓		✓	53.7 (+2.4)	71.7	58.8	55.2 (+2.0)	73.7	60.3
✓	✓	✓	54.4 (+3.1)	72.1	59.4	55.9 (+2.7)	74.1	61.4

Table 3: Distillation performance comparison between Large- and Base-scale teacher models. Numbers in brackets refer to distillation gains against the original student (Swin-T, 51.3 mAP).

Teacher	Freeze Backbone	Distilled Student		
		mAP	AP ₅₀	AP ₇₅
Large-scale	✓	54.4 (+3.1)	72.1 (+3.1)	59.4 (+3.4)
Base-scale	✓	53.5 (+2.2)	71.5 (+2.5)	58.5 (+2.5)
Base-scale	✗	53.1 (+1.8)	71.1 (+2.1)	58.0 (+2.0)

trained on 80GB A100 GPUs. Student models are trained with 12 epochs with AdamW as the optimizer. The initial learning rate is set to 0.0001, which is divided by 10 at the 11th epoch. Following the default practice of DINO [54], the number of object query is set to 900.

4.2 Ablation Studies

In ablation studies, unless otherwise stated, we adopt the tiny-scale models with around 30M parameters as the student.

Effects of each component. In Tab. 2, we verify the effectiveness of each component, including the frozen teacher method, Query Distillation, and head inheriting. When facing large-scale teacher models, Query Distillation produces a modest increase of +0.3/1.0 mAP for Swin/Intern, respectively. Subsequently, we replace the fully fine-tuned teacher with the frozen teacher and perform Query Distillation again. The results demonstrate that the frozen teacher significantly amplifies the distillation gains. Specifically, compared to the fully fine-tuned teacher, the frozen teacher yields an additional increase of +0.6/0.5% for Swin/Intern models, respectively. Moreover, inheriting the transformer’s parameters of the teacher’s head produces noticeable improvements, suggesting that the head parameters of the teacher contain some general knowledge that is somewhat independent of the backbone structure. Building on head parameter inheriting, the frozen teacher outperforms the fully fine-tuned teacher by +0.7/0.7% for Swin/Intern models. These experimental findings validate that the frozen teacher is easier to distill than the previously employed fully fine-tuned teacher. Finally, by combining the three components, DLIM-Det, achieves significant improvements, surpassing the student baseline by +3.1/2.7% for the Swin/Intern models, respectively.

Distillation on medium- and large-sized teacher models. In Tab. 3, we compare the distillation performance under Base- and Large-scale teacher models. With the help of the frozen teacher method, a large-scale teacher surpasses its base-scale counterpart by +0.9 mAP. These results affirm the core premise of this paper: larger teacher models lead

Table 4: The frozen setting affects the detection performance.

Methods	Freeze Backbone	Schedule	mAP	AP ₅₀	AP ₇₅
Swin-Large	×	3x	58.1	77.1	63.5
	✓	3x	56.6 (-1.5)	74.9	61.9
Intern-Large	×	2x	57.8	76.5	63.1
	✓	2x	55.8 (-2.0)	74.4	61.0

Table 5: Applying different training schedules to the Frozen and Full ft. settings as to obtain teachers with identical performance. The KD contains Query Distillation and the inheriting head.

Methods	Freeze Backbone	Schedule	mAP	AP ₅₀	AP ₇₅
Teacher:					
T1: Swin-Large	×	1x	56.9	75.6	62.2
T2: Swin-Large	✓	3x	56.7	74.9	61.9
Student:					
Swin-Tiny	-	1x	51.3	69.0	56.0
KD from T1	-	1x	53.1 (+1.8)	70.9	57.9
KD from T2	-	1x	54.4 (+3.1)	72.1	59.4

Table 6: Effectiveness of Query Position Distill (QPD) and Query Relation Distill (QRD). “T/S” denotes the Teacher/Student.

Methods	InternImage			Swin Transformer		
	mAP	AP ₅₀	AP ₇₅	mAP	AP ₅₀	AP ₇₅
T: Base	54.6	73.6	59.3	55.8	74.4	60.7
S: Tiny	53.2	71.5	58.0	51.3	69.0	56.0
+ QPD	54.2 (+1.0)	72.4	59.1	51.6 (+0.3)	69.5	56.2
+ QRD	54.2 (+1.0)	72.9	59.1	51.8 (+0.5)	70.2	56.5
+ QPD & QRD	54.5 (+1.3)	73.2	59.5	52.1 (+0.8)	69.9	56.7

to greater distillation gains. Additionally, the frozen teacher method generalizes well to base-scale models. When using the base-scale model as a teacher, the frozen teacher surpasses the fully fine-tuned teacher by 0.4 mAP.

Fair comparison between the fully fine-tuned teacher and frozen teacher. Freezing the backbone typically results in performance degradation to some extent compared to fully fine-tuning. As shown in Tab. 4, under the same training schedule, the frozen teacher lags behind the fully fine-tuned teacher by around 2 mAP. One might argue that, owing to lower detection performance, the frozen teacher narrows the gap between the teacher and student, thereby attaining better distillation results. In Tab. 5, we conduct experiments to challenge this perspective. Concretely, we assign long and short training schedules to the frozen and fully fine-tuned teachers, resulting in teachers with nearly identical performance. By ensuring the same gap between the teacher and student, we find that the frozen teacher surpasses the fully fine-tuned teacher by a large margin, **+1.3** mAP. These results highlight the intrinsic advantages of the frozen teacher approach, such as sparse query locations.

Effectiveness of the Query Distillation. The Query Distillation is specifically designed for detection transformers, including Query Position Distill (QPD) and Query Relation Distill (QRD). We ablate each of them step by step in Tab. 6. Here, we employ the medium-sized models as the teacher, to exclude the effects of the large gap between the teacher and student. Besides the experiments on the advanced network architecture, such

Table 7: Applying Query Position Distill (QPD) and Query Relation Distill (QRD) to classical backbones (*e.g.* ResNet [20]).

Methods	Resnet		
	mAP	AP ₅₀	AP ₇₅
T: ResNet101	50.0	67.7	54.4
S: ResNet50	48.9	66.4	53.4
+ QPD	49.5 (+0.5)	67.2	53.9
+ QRD	49.5 (+0.5)	67.5	54.1
+ QPD & QRD	49.8 (+0.9)	67.6	54.2

Table 8: Study on distillation positions and teacher query selection.

(a) Study on distillation positions. Teacher: Intern-Base (54.6 mAP); Student: Intern-Tiny (53.2 mAP)

KD on decoder layer						Methods		
1	2	3	4	5	6	mAP	AP ₅₀	AP ₇₅
✓	✓	✓				54.1	72.6	59.0
			✓	✓	✓	53.8	72.3	58.8
✓	✓	✓	✓	✓	✓	54.5	73.2	59.5
✓						54.1	72.6	59.0
		✓				53.5	72.1	58.3
					✓	52.4	70.9	57.0
T: Intern-Base						54.6	73.6	59.3
S: Intern-Tiny						53.2	71.5	58.0
No Sample (900)						54.2	72.6	59.1
Random Sample (300)						54.5	73.2	59.5
FG queries (IoU > 0.3)						54.2	72.5	59.2
BG queries (IoU < 0.3)						54.2	72.8	59.2

Table 9: Applying our DLIM-Det to the huge model (InternImage-Huge) with 1B parameters to verify the scalability. Numbers in brackets refer to distillation gains against the original student.

Teacher	#Params	Distilled Student		
		mAP	AP ₅₀	AP ₇₅
Intern-Huge	1.08B	56.1 (+2.9)	74.4 (+2.9)	61.5 (+3.5)
Intern-Large	223M	55.9 (+2.7)	74.1 (+2.6)	61.4 (+3.4)

as Swin Transformer and InternImage (Tab. 6), we also report results on ResNet [20] in Tab. 7. Our Query Distillation method yields consistent improvements on three different backbones, at least +0.8 mAP. Furthermore, combining the QPD and QRD together achieves better performance than using them individually, verifying the effectiveness of each component.

Study on the distillation positions. To investigate the impact of Query Distillation on different decoder layers in DETRs, we conducted experiments, and the results are presented in Tab. 8a. When applying Query Distillation to all 6 decoder layers, we observe the best performance, surpassing the counterparts where it was applied to the first three or last three decoder layers. Furthermore, we examine the performance when applying Query Distillation to only one decoder layer. The results show that distillation on shallow decoder layers outperforms distillation on deeper layers.

Selection of teacher queries. Using our Query Distillation method, we further conduct experiments to investigate the impact of different subsets of teacher queries on knowledge transfer. Results are reported in Tab. 8b. First, we perform KD on all teacher queries, obtaining +1.0% gains. Then we randomly sample 300 queries and achieve further improvements, surpassing the 900 query baseline by +0.3%. To explore alternative query selection strategies, we define the queries whose IoUs with GT are greater than 0.3 as foreground queries, otherwise as background queries. Distilling knowledge from either pure foreground (namely IoU > 0.3 in Tab. 8b) or pure background queries results in inferior performance compared to random sampling. This suggests that both foreground and background queries contain valuable knowledge, and distilling knowledge from a combination of them yields optimal results.

Table 10: Experiments on heterogeneous teacher-student pairs. Heterogeneous components are underlined. Numbers in brackets denote distillation gains against the original student.

Teacher	Student	Distilled Student		
		mAP	AP ₅₀	AP ₇₅
DINO-Swin-L	Deformable DETR-Swin-T	49.6 (+3.5)	68.3 (+3.4)	53.7 (+3.9)
DINO- <u>Intern</u> -H	DINO- <u>Swin</u> -T	53.3 (+2.0)	71.3 (+2.3)	58.3 (+2.3)

Table 11: Comparison with state-of-the-art knowledge distillation methods.

Methods	Swin Transformer						InternImage					
	mAP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	mAP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
Teacher: Large	58.1	77.1	63.5	41.2	62.1	73.9	57.8	76.5	63.1	41.6	61.3	73.0
Student: Tiny	51.3	69.0	56.0	34.5	54.4	66.1	53.2	71.5	58.0	35.5	56.6	69.3
FGD [49]	51.1 (-0.2)	69.2	55.8	34.3	53.8	65.6	52.8 (-0.4)	71.1	57.6	36.4	56.4	66.8
MGD [50]	51.1 (-0.2)	69.8	55.8	34.9	54.2	66.1	54.2 (+1.0)	72.8	59.1	38.2	57.6	69.4
FitNet [37]	51.3 (+0.0)	70.0	55.9	34.7	54.5	66.3	53.8 (+0.5)	72.4	58.7	38.2	57.4	68.2
DETRDistill [3]	51.7 (+0.4)	69.4	56.4	35.0	54.8	66.0	54.7 (+1.5)	72.8	59.8	38.5	58.2	69.4
Our DLIM-Det	54.4 (+3.1)	72.1	59.4	38.1	57.4	69.7	55.9 (+2.7)	74.1	61.4	40.3	59.4	70.8

Scalability to huge teacher models. We apply DLIM-Det to InternImage-Huge with 1.08 billion parameters. From Tab. 9, we observe that DLIM-Det achieves satisfactory performance even when the teacher has more than 30 times the parameters of the student. Here, the student is represented by Intern-Tiny with 30M parameters, achieving a mAP of 53.2. Specifically, Intern-Huge produces a +2.9 distillation gain on Intern-Tiny based student and surpasses Intern-Large by 0.2 mAP (56.1 v.s. 55.9).

Distillation on heterogeneous architectures. To verify the generalization of DLIM-Det, we conduct experiments on heterogeneous teacher and student models, including heterogeneous detectors and heterogeneous backbones. For heterogeneous detectors, with DINO [54] as the teacher and Deformable DETR [57] as the student, DLIM-Det improves the student from 46.1 to 49.6 mAP, as reported in Tab. 10. In the case of heterogeneous backbones, where Intern-Huge/Swin-Tiny are employed as the teacher/student, DLIM-Det obtains a considerable improvement of +2.0 mAP. These results show the strong generalization of DLIM-Det.

4.3 Comparison with State-of-the-art Methods

Main results on two backbones. In Tab. 11, we compare our DLIM-Det with several popular KD methods. FitNet [37] is a classical feature distillation method, which introduces adaptation layers to bridge the semantic gap between the teacher and student. However, FitNet fails to improve the performance of the transformer-based backbone (Swin Transformer), showing no improvements. MGD [50] is inspired by the masked image modeling [18] and randomly masks pixels of the student’s features. It then leads the student to generate the teacher’s features in the masked regions. MGD obtains a noticeable improvement of +1.0% for InternImage, but struggles to improve the Swin Transformer. FGD [49] leverages the GcBlock [1] to capture the relations between pixels from feature maps. However, FGD does not perform well on both backbones, indicating that the hand-crafted relation mechanism may conflict with the intrinsic relation modeling of detection transformers. In fact, FGD even hurts the performance of InternImage by -0.4%. The recently proposed DETRDistill [3] is designed for distilling DETRs. In addition to performing customized logits and feature distillation, it also distills the label

assignment results of teacher. Notably, our DLIM-Det outperforms it by a large margin, at least +1.2%. We explain that the Query prior Assignment Distillation in DETRDistill serves a role akin to the De-Noising group in DINO, thereby it only yields minor improvements in our experiments. Moreover, another contribution of DETRDistill [3] is a feature imitation method, which is vulnerable to significant discrepancy between the teacher and student, as confirmed by other works [37, 49, 50].

These experimental findings suggest that while previous works can achieve decent distillation gains for medium-sized teacher, they encounter challenges in distilling large-scale teacher models. In contrast, DLIM-Det demonstrates significant distillation performance specifically tailored for large-scale models.

Table 12: Comparison with the Teacher Assistant (TA) on two backbones. The TA consists of two steps, thus the overall training time is the sum of the training time for both steps.

Methods	Swin Transformer				InternImage			
	mAP	AP ₅₀	AP ₇₅	Train Time	mAP	AP ₅₀	AP ₇₅	Train Time
<i>Baseline:</i>								
Teacher: Large	58.1	77.1	63.5	-	57.8	76.5	63.1	-
Intermediate Teacher: Base	55.8	74.4	60.7	-	54.6	73.6	59.3	-
Student: Tiny	51.3	69.0	56.0	-	53.2	71.5	58.0	-
TA step1 (Large→Base)	57.5 (+1.7)	76.1	62.8	27.3h	56.8 (+2.2)	75.2	62.1	34.5h
TA setp2 (Base→Tiny)	53.7 (+2.4)	71.7	58.9	19.2h	55.7 (+2.5)	74.0	60.7	23.6h
Our DLIM-Det (Large→Tiny)	54.4 (+3.1)	72.1	59.4	20h	55.9 (+2.7)	74.1	61.4	24.8h

Comparison with the Teacher Assistant. The Teacher Assistant (TA) [35, 39, 56] is proved effective in reducing the gap between the teacher and student models. Following the previous work [35], we use the base-scale model with approximately 90M parameters as the TA model. The TA model is distilled from the teacher model and then serves as the teacher for the student model. The results in Tab. 12 demonstrate that our DLIM-Det, which directly distills knowledge from large-scale models, achieves superior performance and higher efficiency compared to the TA method. Specifically, by using only half of the training time, DLIM-Det outperforms Teacher Assistant by 0.7/0.2 mAP on Swin/Intern models, respectively.

5 Conclusion

In this paper, we introduce DLIM-Det, a novel knowledge distillation method specifically tailored for large-scale image models in the context of object detection. To bridge the gap between the teacher and student, we propose the frozen teacher method, which involves freezing the pretrained backbone during teacher creation. The frozen teacher preserves the generalization capacity of large-scale models and gives rise to distinctive characteristics, such as sparse query locations. Through the proposed Query Distillation method, we harness the full advantages offered by the frozen teacher, thereby facilitating the distillation process. Extensive experiments provide robust evidence that larger teacher models can indeed lead to greater distillation gains through DLIM-Det.

Acknowledgement. This work is partially supported by National Natural Science Foundation of China (No. 62322602), Natural Science Foundation of Jiangsu Province, China (No. BK20230033), and National Natural Science Foundation of China (No. 62172225).

References

1. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF international conference on computer vision workshops. pp. 0–0 (2019) [8](#), [13](#)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 213–229. Springer (2020) [4](#)
3. Chang, J., Wang, S., Xu, H.M., Chen, Z., Yang, C., Zhao, F.: Detrdistill: A universal knowledge distillation framework for detr-families. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6898–6908 (2023) [2](#), [4](#), [13](#), [14](#)
4. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems* **30** (2017) [3](#)
5. Chen, Q., Chen, X., Zeng, G., Wang, J.: Group detr: Fast training convergence with decoupled one-to-many label assignment. *arXiv preprint arXiv:2207.13085* (2022) [3](#), [4](#)
6. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems* **35**, 16664–16678 (2022) [4](#)
7. Chen, X., Chen, J., Liu, Y., Zeng, G.: D²etr: Decoder distillation for detection transformer. *arXiv preprint arXiv:2211.09768* (2022) [4](#)
8. Dai, X., Jiang, Z., Wu, Z., Bao, Y., Wang, Z., Liu, S., Zhou, E.: General instance distillation for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7842–7851 (2021) [2](#), [4](#), [8](#)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018) [5](#)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020) [4](#)
11. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636* (2022) [1](#)
12. Floridi, L., Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines* **30**, 681–694 (2020) [5](#)
13. Gao, Z., Wang, L., Han, B., Guo, S.: 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) [4](#)
14. Guo, G., Han, L., Wang, L., Zhang, D., Han, J.: Semantic-aware knowledge distillation with parameter-free feature uniformization. *Visual Intelligence* **1**(1), 6 (2023) [3](#)
15. Guo, J., Han, K., Wang, Y., Wu, H., Chen, X., Xu, C., Xu, C.: Distilling object detectors via decoupled features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2154–2164 (2021) [2](#), [3](#), [4](#)
16. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5356–5364 (2019) [2](#)
17. Hao, Z., Guo, J., Jia, D., Han, K., Tang, Y., Zhang, C., Hu, H., Wang, Y.: Learning efficient vision transformers via fine-grained manifold distillation. In: *Advances in Neural Information Processing Systems* (2021) [3](#), [4](#), [8](#)

18. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022) [5](#), [13](#)
19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) [5](#)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) [4](#), [12](#)
21. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) [2](#), [5](#)
22. Huang, T., You, S., Wang, F., Qian, C., Xu, C.: Knowledge distillation from a stronger teacher. arXiv preprint arXiv:2205.10536 (2022) [4](#), [8](#)
23. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022) [4](#)
24. Jia, Z., Sun, S., Liu, G., Liu, B.: Mssd: multi-scale self-distillation for object detection. Visual Intelligence **2**(1), 8 (2024) [3](#)
25. Kang, Z., Zhang, P., Zhang, X., Sun, J., Zheng, N.: Instance-conditional knowledge distillation for object detection. Advances in Neural Information Processing Systems **34**, 16468–16480 (2021) [3](#)
26. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: 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) [4](#)
27. Li, G., Li, X., Wang, Y., Zhang, S., Wu, Y., Liang, D.: Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1306–1313 (2022) [4](#), [8](#)
28. Li, G., Li, X., Zhang, S., Yang, J.: Towards more reliable evaluation in pedestrian detection by rethinking “ignore regions”. Visual Intelligence **2**(1), 4 (2024) [4](#)
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) [6](#), [8](#), [9](#)
30. Lin, Y., Liu, Z., Zhang, Z., Hu, H., Zheng, N., Lin, S., Cao, Y.: Could giant pre-trained image models extract universal representations? Advances in Neural Information Processing Systems **35**, 8332–8346 (2022) [4](#)
31. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329 (2022) [4](#)
32. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J.: Structured knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2604–2613 (2019) [3](#)
33. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022) [1](#), [4](#), [5](#)
34. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) [1](#), [2](#), [3](#), [4](#), [9](#)
35. Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 5191–5198 (2020) [2](#), [14](#)

36. Park, D.Y., Cha, M.H., Kim, D., Han, B., et al.: Learning student-friendly teacher networks for knowledge distillation. *Advances in neural information processing systems* **34**, 13292–13303 (2021) [2](#)
37. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550* (2014) [2](#), [3](#), [13](#), [14](#)
38. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18. pp. 234–241. Springer (2015) [7](#)
39. Son, W., Na, J., Choi, J., Hwang, W.: Densely guided knowledge distillation using multiple teacher assistants. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9395–9404 (2021) [2](#), [14](#)
40. Wang, C., Yang, Q., Huang, R., Song, S., Huang, G.: Efficient knowledge distillation from model checkpoints. *Advances in Neural Information Processing Systems* **35**, 607–619 (2022) [7](#)
41. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al.: Internimage: Exploring large-scale vision foundation models with deformable convolutions. *arXiv preprint arXiv:2211.05778* (2022) [1](#), [3](#), [4](#), [5](#), [9](#)
42. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442* (2022) [1](#), [4](#), [5](#)
43. Wang, Y., Zhang, X., Yang, T., Sun, J.: Anchor detr: Query design for transformer-based detector. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 36, pp. 2567–2575 (2022) [4](#)
44. Wang, Y., Ni, Z., Song, S., Yang, L., Huang, G.: Revisiting locally supervised learning: an alternative to end-to-end training. *arXiv preprint arXiv:2101.10832* (2021) [7](#)
45. Wei, F., Gao, Y., Wu, Z., Hu, H., Lin, S.: Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems* **34**, 22682–22694 (2021) [5](#)
46. Xu, T., Zhu, X.F., Wu, X.J.: Learning spatio-temporal discriminative model for affine subspace based visual object tracking. *Visual Intelligence* **1**(1), 4 (2023) [4](#)
47. Yang, C., Zhou, H., An, Z., Jiang, X., Xu, Y., Zhang, Q.: Cross-image relational knowledge distillation for semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12319–12328 (2022) [3](#)
48. Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., Gao, J.: Focal attention for long-range interactions in vision transformers. *Advances in Neural Information Processing Systems* **34**, 30008–30022 (2021) [1](#), [4](#), [5](#)
49. Yang, Z., Li, Z., Jiang, X., Gong, Y., Yuan, Z., Zhao, D., Yuan, C.: Focal and global knowledge distillation for detectors. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4643–4652 (2022) [4](#), [8](#), [13](#), [14](#)
50. Yang, Z., Li, Z., Shao, M., Shi, D., Yuan, Z., Yuan, C.: Masked generative distillation. In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*. pp. 53–69. Springer (2022) [2](#), [13](#), [14](#)
51. Yao, L., Pi, R., Xu, H., Zhang, W., Li, Z., Zhang, T.: G-detkd: towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 3591–3600 (2021) [2](#), [3](#)
52. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432* (2021) [1](#)

- 53. Zeng, Z., Yang, F., Liu, H., Satoh, S.: Improving deep metric learning via self-distillation and online batch diffusion process. *Visual Intelligence* **2**(1), 1–13 (2024) [3](#)
- 54. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605* (2022) [2](#), [3](#), [4](#), [9](#), [10](#), [13](#)
- 55. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4320–4328 (2018) [5](#)
- 56. Zheng, Z., Ye, R., Hou, Q., Ren, D., Wang, P., Zuo, W., Cheng, M.M.: Localization distillation for object detection. *arXiv preprint arXiv:2204.05957* (2022) [2](#), [4](#), [14](#)
- 57. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020) [4](#), [13](#)
- 58. Zong, Z., Song, G., Liu, Y.: Detsr with collaborative hybrid assignments training. *arXiv preprint arXiv:2211.12860* (2022) [3](#), [4](#)