Distilling DETR with Visual-Linguistic Knowledge for Open-Vocabulary Object Detection

Liangqi Li[1], Jiaxu Miao[2], Dahu Shi[1,2,3]*, Wenming Tan[1], Ye Ren[1], Yi Yang[2], Shiliang Pu[1]
[1]Hikvision Research Institute [2]Zhejiang University
[3]Key Laboratory of Peace-building Big Data of Zhejiang Province
{liliangqi,shidahu,tanwenming,renye,pushiliang}@hikvision.com
{jiaxumiao,yangyics}@zju.edu.cn

Abstract

Current methods for open-vocabulary object detection (OVOD) rely on a pre-trained vision-language model (VLM) to acquire the recognition ability. In this paper, we propose a simple yet effective framework to Distill the Knowledge from the VLM to a DETR-like detector, termed DK-DETR. Specifically, we present two ingenious distillation schemes named semantic knowledge distillation (SKD) and relational knowledge distillation (RKD). To utilize the rich knowledge from the VLM systematically, SKD transfers the semantic knowledge explicitly, while RKD exploits implicit relationship information between objects. Furthermore, a distillation branch including a group of auxiliary queries is added to the detector to mitigate the negative effect on base categories. Equipped with SKD and RKD on the distillation branch, DK-DETR improves the detection performance of novel categories significantly and avoids disturbing the detection of base categories. Extensive experiments on LVIS and COCO datasets show that DK-DETR surpasses existing OVOD methods under the setting that the base-category supervision is solely available. The code and models are available at https://github.com/hikvision-research/opera.

1. Introduction

Object detection has witnessed rapid progress [10, 9, 28, 13, 20, 2, 45, 33, 32, 35] for years, which aims to localize and categorize objects in an image. However, the object detection model can only perform well on a closed and small set of categories, while cannot detect novel ones which are not trained. Recently, visual-language models [24, 15] (VLM), consisting of an image encoder and a text encoder, have shown impressive zero-shot classification ability after being trained on large-scale loosely aligned image-text pairs. This motivates the community to implement an open-vocabulary object detector [11, 5, 7, 39, 43, 42, 44, 23] which is expected to recognize *arbitrary* categories.

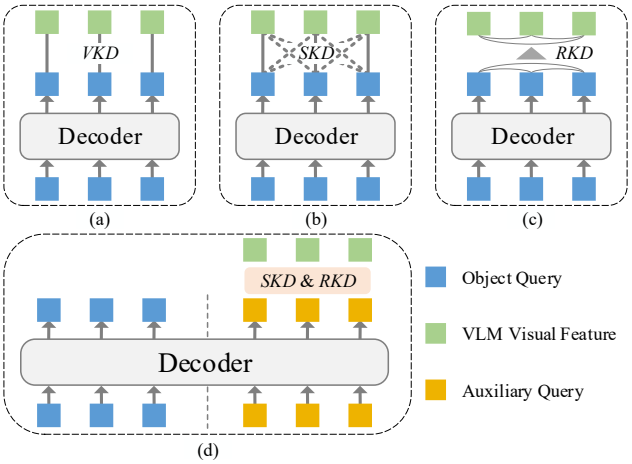

Figure 1. **Comparison of different distillation implementations.** VKD in (a) denotes vanilla knowledge distillation that forces source features to align with target features in a one-to-one manner. SKD in (b) and RKD in (c) are newly proposed for OVOD in this paper. (d) indicates that we implement distillation on auxiliary queries to avoid original detection branch being disturbed.

Distilling the knowledge of novel-category objects from the VLM to the detector is a typical practice [11, 36, 22] to solve the open-vocabulary object detection (OVOD) problem. ViLD [11] is a representative distillation-based approach. Category text embeddings, extracted by the VLM text encoder, serve as the classifier to perform OVOD (known as ViLD-text). A knowledge distillation (KD) module is further introduced to align the object features to visual embeddings extracted by the VLM image encoder. By adopting the vanilla KD as depicted in Figure 1, ViLD evidently improves the performance of novel categories. Besides, ZSD-YOLO[36] and HierKD [22] also adopt knowl-

*Corresponding author.

Method	AP_r	AP_c	AP_f
ViLD-text [11]	10.1	23.9	32.5
ViLD [11]	16.6	24.6	30.3

Table 1. **OVOD performance of ViLD on LVIS benchmark.** AP_r, AP_c and AP_f denote the performance of rare, common and frequent categories in LVIS dataset, respectively. Compared with ViLD-text which does not adopt distillation, there is an evident AP_f drop for ViLD.

edge distillation techniques to improve the novel-category performance based on one-stage detectors [27, 41], instead of the two-stage detector [13] used in ViLD. Recently, end-to-end detectors [2, 45] boost the development of object detection due to their high efficiency and effectiveness. However, distilling the knowledge to an end-to-end detector is less studied in the OVOD field.

In this paper, we propose a framework to **d**istill the **k**nowledge from the VLM to a **DETR**-like detector, termed DK-DETR. Nevertheless, the vanilla knowledge distillation adopted by aforementioned methods leads to limited improvement on novel categories. To this end, we propose two ingenious knowledge distillation schemes, namely semantic knowledge distillation (SKD) and relational knowledge distillation (RKD) , as shown in Figure 1. In SKD, the feature alignment between the detector and the VLM image encoder is treated as a pseudo-classification problem instead of a regression problem as in vanilla knowledge distillation. It not only pulls together features belonging to the same object but also pushes away features from different objects. In RKD, considering that the VLM can construct a well-structured feature space among abundant visual entities, we propose to model relationships between objects hidden in the VLM image encoder and distill the relational knowledge to our detector.

Although knowledge distillation can effectively improve the novel-category performance, it negatively affects base categories which are well trained with sufficient ground-truth labels (*e.g.*, base-category performance AP_f in ViLD drops from 32.5 to 30.3 in Table 1). Such a phenomenon can be attributed to training objective inconsistence and domain shift between the VLM and the detector. Under the supervision of ground-truth labels, object features in the detector are trained to localize and recognize base-category objects, but distillation forces object features to align with VLM visual embeddings, which results in feature disturbance. Consequently, we add a group of auxiliary queries in our approach for distillation exclusively, which avoids the performance degradation of base categories.

Equipped with SKD and RKD on auxiliary queries, DK-DETR achieves satisfactory performance on both base and novel categories. Note that both distilling implementations and auxiliary queries are only used for training, and do not introduce any budget at inference. The main contributions of this work are summarized as follows.

- We propose a simple yet effective distilling framework for the end-to-end open-vocabulary object detector and effectively improve novel-category performance.
- To distill the knowledge from the VLM to the detector, the proposed SKD transfers the semantic knowledge explicitly, while RKD exploits implicit relationship information between objects.
- By introducing a group of auxiliary queries, DK-DETR disentangles the training of the detection and distillation, which avoids the performance degradation of base categories.
- DK-DETR surpasses existing OVOD methods on both LVIS and COCO datasets under the setting that only the base-category supervision is available, and also achieves competitive performance when it generalizes to other datasets.

2. Related Work

2.1. Open-vocabulary Object Detection

Current open-vocabulary object detection methods usually acquire the ability to recognize novel-category objects by utilizing pre-trained vision-language models (VLM). From the perspective of how to use the VLM, we can divide current researches about OVOD into three main groups.

Knowledge distillation. It is straightforward to mine the knowledge about novel categories by transferring it from the VLM to the detector. ViLD[11], ZSD-YOLO[36] and HierKD[22] use knowledge distillation to perform the transferring. Such methods effectively improve the detection performance for novel categories. However, ViLD needs an external RPN to generate proposals for distillation in advance, and a cumbersome dual-head structure to assemble scores, which makes it inefficient. ZSD-YOLO distills the knowledge to YOLO [25, 26, 27] more efficiently without offline proposals. HierKD designs hierarchical distillation by using caption annotations upon ATSS [41], which achieves both high performance and efficiency. All these methods adopt vanilla knowledge distillations (VKD) in a one-to-one manner. In this paper, we find VKD only mines limited information and propose two ingenious knowledge distillation schemes.

Exploiting extra data. PromptDet [7] uses CLIP [24] to explore objects of novel categories on an external dataset LAION-400M [30], and produces pseudo bounding box annotations to further train the detector. Detic [44] uses the external ImageNet-21K [4] for joint training, which expands the vocabulary of the detector to a large number of concepts.

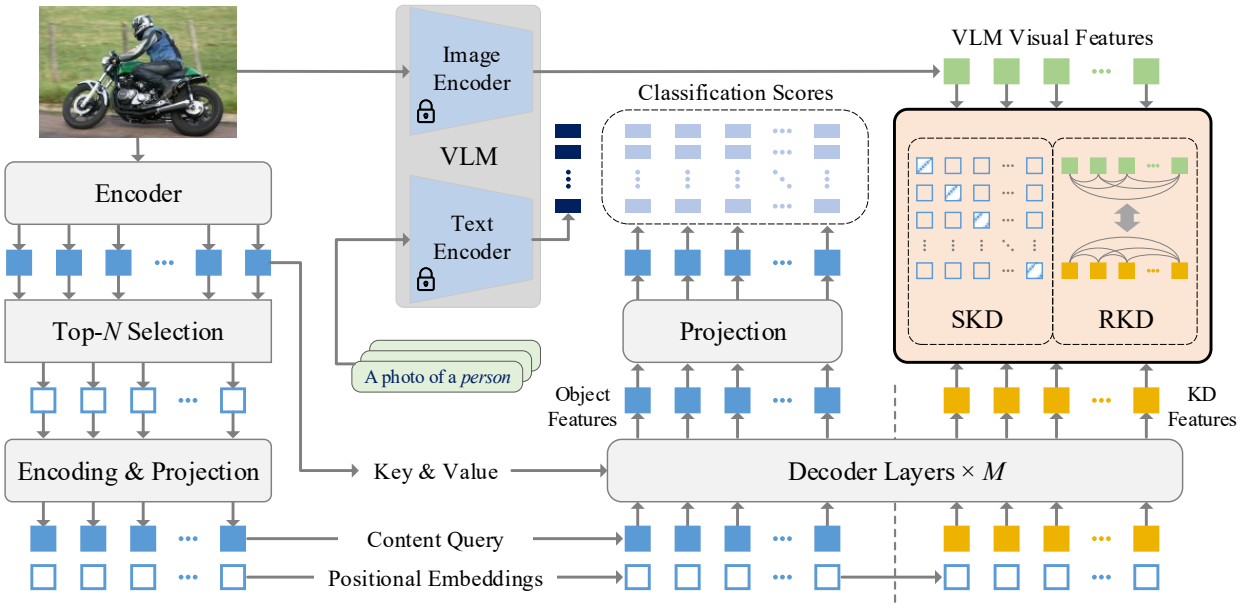

Figure 2. **The overall architecture of DK-DETR.** Given an input image, we extract feature tokens and feed them into an encoder followed by a decoder to generate object features as in Deformable DETR [45]. These object features are projected and used to calculate cosine similarities with text embeddings from the VLM text encoder to perform detection for base and novel categories. Moreover, we introduce two kinds of knowledge distillation schemes based on a newly added distillation branch along with a group of auxiliary KD queries. The distillation branch transfers the knowledge, which is beneficial for novel categories, from the VLM image encoder to the detector. Note that the distillation branch is only used for training and introduces no cost during inference.

VL-PLM [42] trains a class-agnostic foreground detector to mine novel-category objects in LVIS [12] itself. By adopting external data or pseudo labels, these methods improve the performance of novel categories considerably and are orthogonal to distillation-based approaches.

Text prompt tuning. The VLM such as CLIP [24] uses a human-designed prompt, *i.e.* "`a photo of a [CLASS]`", along with categories names to produce text embeddings. DetPro [5] and PromptDet [7] think that the human-designed prompt is not suitable enough for classes and scenes to be detected. And they propose a prompt engineering pipeline to train several learnable prompt embeddings on the detection dataset before training the detector.

2.2. Transformer in Object Detection

Transformer [34] has been widely applied in natural language processing. Recently, DETR [2] opens up the opportunity for employing transformers in object detection tasks. Many follow-up researches attempted to speed up training convergence of DETR. Deformable DETR [45] proposes deformable attention modules which only attend to certain sampling points. DN-DETR [18] and DINO [40] present denoising training methods with noise-added ground-truth labels. Group DETR [3] and Hybrid Matching [16] introduce auxiliary queries to convert the one-to-one matching in DETR to a one-to-many matching, which improves both the training efficiency and detection performance.

3. Method

For the open-vocabulary object detection task, the model is designated to detect objects of an *arbitrary* category. Following prior works [11, 36], categories of an off-the-shelf object detection dataset are divided into two subsets, namely base categories C_{base} and novel categories C_{novel}, in which only base categories are used for training.

In this section, we introduce four main components: assembling DETR-based detector with text embeddings, auxiliary knowledge distillation branch, semantic knowledge distillation and relational knowledge distillation, as shown in Figure 2.

3.1. Overall Architecture

We build our open-vocabulary detection framework on Deformable DETR [45] detector. Given an image I, the encoder outputs refined multi-scale feature tokens as memory features. These feature tokens denoting potential objects are fed into a classification head and a regression head to generate objectness confidence scores and coarse bounding boxes. We select top-N tokens according to their confidence scores and choose corresponding bounding boxes $B = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N\}$ as initial anchor boxes.

Through a sinusoidal encoding [34] and a projection layer, these anchor boxes are used to produce content queries $Q^{obj} = \{\mathbf{q}_1^{obj}, \mathbf{q}_2^{obj}, \ldots, \mathbf{q}_N^{obj}\} \in \mathbb{R}^{N\times D}$ (also known as object queries in DETR [2]) and positional embeddings for the following decoder. N content queries along with positional embeddings and memory features are fed into six decoder layers to get N object embeddings $F^{obj} \in \mathbb{R}^{N\times D}$ representing object features of N potential objects in image I. Then F^{obj} are fed into a text-based classifier (Section 3.2) to produce classification scores that correspond to C_{base} for training, or $C_{base} \cup C_{novel}$ for inference. In particular, a projection layer is employed to object features to align with the text embedding dimension. For clarity, the pipeline from object queries Q^{obj} to classification scores is called the detection branch in our method.

Additionally, to explore the rich knowledge in the pre-trained vision-language model (VLM), we introduce an auxiliary knowledge distillation (KD) branch (Section 3.3) and propose two ingenious knowledge distillation schemes, namely semantic knowledge distillation (Section 3.4) and relational knowledge distillation (Section 3.5). Note that the distillation branch is only used for the training stage, which does not introduce any extra computational cost during inference.

3.2. Text-based Classifier

Original Deformable DETR maps semantic categories to discrete integral labels to perform category classification. However, for novel categories beyond the integral-label set, such a detector is incapable of classifying them. Inspired by the pre-trained vision-language model CLIP [24], we propose to define the label space by the category name itself. Specifically, given a category and its name, we first feed the name into a human-designed language template such as "`a photo of a [CLASS]`" to form a sentence. Then the sentence is fed into the VLM text encoder $\mathcal{T}(\cdot)$ to extract the category-conditioned text embedding $\mathbf{t}$. Next, the text embedding $\mathbf{t}$ is used to compute a cosine similarity with the visual feature $\mathbf{f}$ of an object, namely

$$cos(\mathbf{f}, \mathbf{t}) = \frac{\mathbf{f}\cdot\mathbf{t}}{||\mathbf{f}||\cdot||\mathbf{t}||}, \tag{1}$$

to measure the correlation between them, where $||\cdot||$ denotes the L_2 normalization, and $cos(\cdot)$ denotes the cosine similarity. Finally, the confidence score is calculated as

$$s = \sigma(cos(\mathbf{f}, \mathbf{t})/\tau), \tag{2}$$

where $\sigma(\cdot)$ denotes the sigmoid function and τ is a temperature.

Given a novel category with its linguistic category name, Deformable DETR can be readily assembled with the text-based classifier to perform OVOD task. And the modified detector is served as an open-vocabulary object detection baseline in this paper.

3.3. Auxiliary Distillation Branch

When the baseline detector generalizes to novel categories directly, the detection performance is unsatisfying because of annotation absence. We propose to distill the knowledge from the VLM image encoder $\mathcal{V}(\cdot)$, which has seen a large number of images of different categories during text-image pre-training, to our detector. A naive implementation of KD as in ViLD [11] is to align object features from the detector to VLM visual embeddings of these objects. Given coarse bounding boxes B of these objects generated by the detector encoder, we can crop regions from the image I and feed them into the VLM image encoder $\mathcal{V}(\cdot)$ to extract visual embeddings $V = \mathcal{V}(crop(I, B))$. However, training objectives of the VLM and the detector are inconsistent, so this naive knowledge distillation would introduce interference to disturb detector features. In this paper, a group of auxiliary learnable embeddings is introduced in the distillation branch to serve as KD queries $Q^{kd} = \{\mathbf{q}_1^{kd}, \mathbf{q}_2^{kd}, \ldots, \mathbf{q}_N^{kd}\} \in \mathbb{R}^{N\times D}$ as shown in Figure 2. Q^{kd}, corresponding to Q^{obj}, shares the same positional embeddings. The distillation branch and the detection branch share the same network weights. Fed with KD queries and positional embeddings, the decoder produces KD features $F^{kd} \in \mathbb{R}^{N\times D}$ for the following distillation. Note that attention masks from KD queries Q^{obj} to object queries Q^{kd} in self-attention modules are blocked to avoid object features being influenced.

3.4. Semantic Knowledge Distillation

Semantic knowledge distillation (`SKD`) aims to directly distill knowledge from the VLM image encoder $\mathcal{V}(\cdot)$ into the detector. A straightforward loss to perform the alignment between two kinds of features is L_1 loss as in ViLD. However, L_1 loss only aligns features in a one-to-one manner, which does not exploit sufficient information in the VLM. And the strict alignment also increases the training difficulty.

Supervised by L_1 loss, formulated as

$$L_1(F^{kd}, V) = \frac{1}{N}\sum_i |F^{kd}_{i,:} - V_{i,:}|, \tag{3}$$

all the elements of a single feature vector $F^{kd}_{i,:}$ should be exactly same with $V_{i,:}$. However, to transfer the knowledge from the VLM image encoder to our detector, we only need to maximize the similarity between two features from the same object. In this paper, we reformulate the regression problem in vanilla knowledge distillation as a pseudo-classification problem. As shown in Figure 3, a pair of KD feature and VLM visual embeddings of an identical object are regarded as a positive pair, otherwise a negative pair. The cosine similarity is regarded as the classification score. The binary-cross-entropy (BCE) loss is adopted to punish

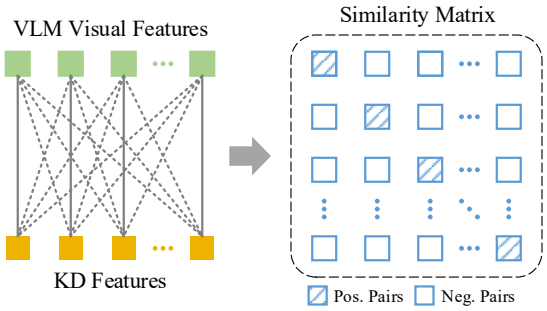

Figure 3. **Semantic Knowledge Distillation.** Instead of distilling two groups of features in a one-to-one manner, we cast the alignment between KD features and VLM visual embeddings into a pseudo-classification problem. Two kinds of features from the same object are regarded as a positive pair, otherwise a negative pair.

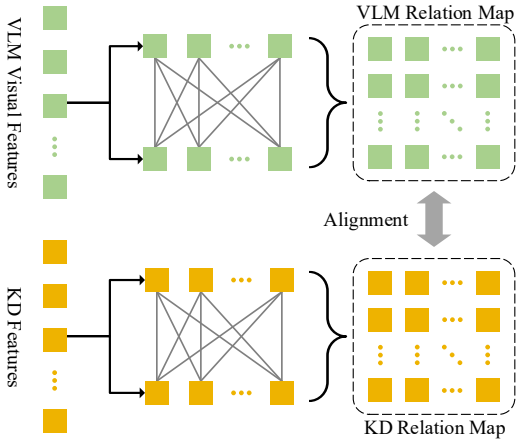

Figure 4. **Relational Knowledge Distillation.** Through calculating the cosine similarity, pair-wise object relationships within each of the VLM features and KD features are modeled to produce a VLM relation map and a KD relation map, respectively. We guide the KD relation map to align with the VLM relation map.

positive samples with label 1, and negative samples with label 0. Specifically, the BCE-based loss is formulated as

$$L_{skd} = -\frac{1}{N}\sum_{i} log(\sigma(\frac{cos(F^{kd}_{i,:}, V_{i,:})}{\tau_s})) - \frac{1}{N}\sum_{\substack{i,j \\ i\neq j}} log(1-\sigma(\frac{cos(F^{kd}_{i,:}, V_{j,:})}{\tau_s})), \quad (4)$$

where $cos(\cdot)$ denotes the cosine similarity as in Equation 1 and τ_s is a temperature. Compared with L_1 loss as in Equation 3, the BCE-based semantic knowledge distillation has two advantages. On the one hand, the loss in Equation 4 does not punish features output by the distillation branch to be exactly the same as the corresponding visual embeddings from the VLM image encoder $\mathcal{V}(\cdot)$, and this would reduce training difficulty. On the other hand, L_1 loss only pulls two kinds of features from an identical object close, while our `SKD` loss pushes the features from different objects far away, and this would exploit some implicit relational information. Furthermore, we propose to explicitly explore the relational information in the next subsection.

3.5. Relational Knowledge Distillation

Semantic knowledge distillation exploits knowledge by aligning semantic features of the distillation branch to visual embeddings of the VLM image encoder. Apart from such direct feature alignment, we propose to explore hidden knowledge contained in the VLM from another dimension. The relation between two objects reflects their correspondence, and it is valuable for open-vocabulary recognition. For example, a tiger may be closer to a cat than a dog in the VLM embedding space. If we distill such knowledge to our detector, it can help the detector to avoid recognizing a tiger as a dog during inference.

Relational knowledge distillation aims to model and distill the relationship between individual objects in image I. Specifically, as shown in Figure 4, given VLM visual embeddings V and KD features F^{kd}, the relationship represented by a pair-wise similarity matrix is calculated for VLM features and KD features, respectively. Relationships are denoted by a VLM relation map $R^v = \bar{F}\bar{F}^T \in \mathbb{R}^{N\times N}$ and a KD relation map $R^{kd} = \bar{V}\bar{V}^T \in \mathbb{R}^{N\times N}$, where $\bar{F}$ is L_2-normalized F^{kd}, and $\bar{V}$ is L_2-normalized V. These two similarity matrices capture the pair-wise correlations among objects, and we guide the matrix R^{kd} from the distillation branch to align with R^V from the VLM image encoder. The distillation process is formulated as

$$L_{rkd} = -\frac{1}{N}\sum_{i=1}^{N} KL(\sigma(\frac{R^{kd}_{i,:}}{\tau_r})||\sigma(\frac{R^{v}_{i,:}}{\tau_r})), \quad (5)$$

where τ_r is a temperature, and $KL(\cdot)$ stands for Kullback-Leibler divergence. It should be noticed that the relationship between objects in a single image is just one kind of correlations we can capture. There are other relationships that can assist our model as well, such as the relationship between objects from different images, which will be further discussed in ablation experiments.

3.6. Loss Functions

Apart from the semantic KD loss and the relational KD loss mentioned above, we use a classification loss function L_{cls} for our text-based classifier as in Deformable DETR [45], and classification scores are calculated as in Equation 2. Besides, both L_1 loss (denoted as L_{box}) and generalized IoU loss [29] L_{iou} are adopted for bounding box regression

as in DETR [2]. Formally, the overall loss function of our DK-DETR can be formulated as:

$$L = \lambda_{cls}L_{cls} + \lambda_{box}L_{box} + \lambda_{iou}L_{iou} + \lambda_{skd}L_{skd} + \lambda_{rkd}L_{rkd}, \quad (6)$$

where λ_{cls}, λ_{box}, λ_{iou}, λ_{skd} and λ_{rkd} are the loss weights, respectively.

4. Experiments

4.1. Dataset

We conduct experiments on LVIS v1 [12] and COCO [21] datasets, which are commonly used in open-vocabulary object detection, to evaluate the performance of our method. Besides, to imitate open-vocabulary applications in the real world, we also demonstrate the generalization ability of our DK-DETR on COCO validation set, Objects365 [31] validation set and Pascal VOC [6] test set by the model trained on LVIS.

LVIS. LVIS dataset is annotated with object detection and instance segmentation labels, and it contains ∼120k images over 1203 categories with a long-tail distribution. The categories are divided into "frequent", "common" and "rare" according to the appearing frequency in the training set. For the open-vocabulary setting, we follow ViLD [11] to use frequent and common categories (including 886 categories) as the base categories C_{base}, and hold out rare categories (including 337 categories) as the novel categories C_{novel}. As for validating, we evaluate all of 1203 classes on the validation set.

COCO. COCO 2017 is a common benchmark for general object detection. There are 80 categories and ∼118k images in COCO. We follow [1] to divide it into 48 base categories and 17 novel categories and keep the remaining 15 categories unused. As for generalization evaluation by the model trained on LVIS, we use all 80 categories.

Objects365. Objects365 is a brand new dataset for object detection research, and it has 365 categories.

Pascal VOC. Pascal VOC is an earlier object detection benchmark that contains only 20 categories.

4.2. Implementation Details

We use Deformable DETR [45] as the base detector of our method, and utilize SOIT head [38] to predict mask results for instance segmentation evaluation. ResNet-50 [14] is selected as the backbone to compare with state-of-the-art methods fairly.

We augment the input image with large-scale jittering [8] of image side range $[102, 2048]$ and random horizontal flipping. Then the image is padded to 1024×1024. We use 16 NVIDIA Tesla V100 GPUs to train models for 70 epochs with a total batch size of 128. Note that SKD and RKD are not used in the first 40 epochs which can be regarded as a burn-in stage. AdamW [17] optimizer with base learning rate of 8×10^{-4}, momentum of 0.9 and weight decay of 0.05 is adopted. The learning rate is decayed at the 55th epoch. Parameters of the ResNet-50 backbone are initialized with models pre-trained on ImageNet dataset [4]. Classifier temperature τ, SKD temperature τ_s, RKD temperature τ_r are set to 0.05, 0.05 and 0.2 respectively. Loss weights used in SKD and RKD are set to $\lambda_{skd} = 0.1$ and $\lambda_{rkd} = 2$ respectively. The query number is set to $N = 300$ as in Deformable DETR [45].

Although we can align N KD features to VLM visual embeddings for distillation, it should be noticed that our target of distillation is to exploit the information about potential novel objects. Consequently, after bipartite matching in the detection branch, we exclude objects belonging to base categories C_{base}, and sample queries with high objectness confidence scores which are more likely to contain foreground objects. In our experiments, we select 20 queries with the highest confidence scores for distillation.

4.3. Main Results

Open-vocabulary LVIS benchmark. We firstly make comparisons with state-of-the-art methods on LVIS [12] dataset, as shown in Table 2. Without any bells and whistles, using the same ResNet-50 as the backbone, our DK-DETR achieves 20.5 AP_r scores for segmentation. It improves the baseline (*i.e.* DK-DETR$_{\text{w/o KD}}$) by 4.1 points and significantly outperforms other methods under the setting that only the base-category supervision is available.

Compared with VL-PLM [42], Detic [44] and PromptDet [7], we do not need extra data or labels for training. Compared with ViLD [11], we both use distillation technology but ViLD just mines superficial knowledge by the vanilla distillation as shown in Figure 1 (a). Note that our method has no negative effects on the performance of base categories (AP_c and AP_f) because we introduce a group of auxiliary queries for distillation. Although both frameworks use Deformable DETR as the base detector, our DK-DETR outperforms OV-DETR [39] by a large margin. Our method can generate predictions of all categories at once. In contrast, OV-DETR only produces the output for one category at one inference forward, which is quite inefficient and can not be deployed to real-world applications.

Open-Vocabulary COCO benchmark. Table 3 shows the open-vocabulary detection performance of our DK-DETR as well as many other state-of-the-art methods on COCO [21]. Following ViLD, we only use bounding box annotations without instance segmentation masks to train the detector, and report detection results only. DK-DETR achieves 32.3 AP_{novel} scores and outperforms all state-of-the-art methods except VL-PLM [42] which generates pseudo labels for novel categories. DK-DETR improves our baseline by 23.3 AP_{novel} scores, and the performance

Method	Base Detector	Segmentation		Detection	
		AP_r	AP_c / AP_f / AP	AP_r	AP_c / AP_f / AP
w/ novel-category supervision*					
VL-PLM [42]	Faster R-CNN [19]	17.2	23.7 / 35.1 / 27.0	-	- / - / -
Detic [44]		17.8	26.3 / 31.6 / 26.8	-	- / - / -
PromptDet† [7]		21.4	23.3 / 29.3 / 25.3	21.8	24.3 / 32.4 / 27.1
w/o novel-category supervision					
ViLD [11]	Faster R-CNN [19]	16.6	24.6 / 30.3 / 25.5	16.7	26.5 / 34.2 / 27.8
RegionCLIP [43]		-	- / - / -	17.1	27.4 / 34.0 / 28.2
DetPro† [5]		19.8	25.6 / 28.9 / 25.9	20.8	27.8 / 32.4 / 28.4
OV-DETR [39]	Deformable DETR [45]	17.4	25.0 / 32.5 / 26.6	-	- / - / -
DK-DETR$_{\text{w/o KD}}$		16.4	28.9 / 35.3 / 29.3	17.3	32.0 / 40.0 / 32.9
DK-DETR		**20.5**	28.9 / 35.4 / 30.0	**22.2**	32.0 / 40.2 / 33.5

Table 2. **Comparison on LVIS dataset.** AP_r is the main metric for evaluation. Rare categories are held out as novel categories for testing. Common and frequent categories serve as base categories for training. All methods are trained with ResNet-50 backbone. *: VL-PLM uses pseudo-label of novel categories; Detic uses the external ImageNet-21K for joint training; PromptDet uses an extra LAION-400M [30] dataset. †: specific text prompt tuning for the VLM text encoder.

Method	Base Dec.	AP_{novel}	AP_{base}
w/ novel-category supervision			
Detic[44]	Faster R-CNN [19]	24.1	44.7
PromptDet[7]		26.6	50.6
VL-PLM[42]		34.4	60.2
w/o novel-category supervision			
ZSD-YOLO[36]	YOLO [26]	13.6	31.7
HierKD[22]	ATSS [41]	20.3	51.3
ViLD[11]	Faster R-CNN [19]	27.6	59.5
RegionCLIP[43]		31.4	57.1
OV-DETR[39]	Deformable DETR [45]	29.4	61.0
DK-DETR$_{\text{w/o KD}}$		9.0	61.2
DK-DETR		**32.3**	61.1

Table 3. **Comparison on COCO open-vocabulary object detection dataset.** AP_{novel} is the main metric for evaluation. All methods are trained with ResNet-50 backbone. "Base Dec." denotes the base detector.

of base categories is barely influenced.

Generalization Ability. To demonstrate the generalization ability [37] of the open-vocabulary object detection model, following ViLD [11], we directly evaluate the LVIS-trained model on COCO, Objects365 [31] and Pascal VOC [6] datasets by just replacing the classifier with text embeddings, which is produced by feeding categories names of those datasets into the VLM text encoder. As in Table 4, our DK-DETR shows convincing performance and robust generalization ability.

4.4. Ablation Study

In this subsection, we perform a number of ablation experiments to analyze the components of the proposed method. In consideration of efficiency, we carry out all ablation experiments on LVIS [12] dataset based on ResNet-50 [14] backbone without segmentation mask annotations. A Deformable DETR detector assembling with a text-based classifier serves as our baseline for the open-vocabulary object detection. All models are trained for 36 epochs.

Effectiveness of SKD, RKD and Auxiliary Queries. We propose two kinds of knowledge distillation methods in this paper, namely semantic knowledge distillation (SKD) and relational knowledge distillation (RKD), based on a group of auxiliary queries. SKD enforces features from the detector to be consistent with those of the VLM, while RKD helps the detector to imitate the relationship between individual objects in the VLM. As shown in Table 5, SKD improves AP_r by 3.0 points and RKD improves AP_r by 2.8 points. When combining both SKD and RKD, our method significantly outperforms the baseline by 3.4 points for AP_r. Note that if we implement SKD and RKD without auxiliary queries (the fourth row in Table 5), base-category performance AP_c and AP_f drop noticeably. This indicates that auxiliary queries are important for our method to avoid detection features being disturbed.

Loss type of SKD. Table 6 shows the influence of SKD loss type. With a group of auxiliary queries, using L_1 loss for distillation could improve the detection performance of novel categories by 1.9 points, and there is no performance

Method	Base Detector	COCO			Objects365			Pascal VOC	
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP_{50}	AP_{75}
ViLD [11]	Faster R-CNN [19]	36.6	55.6	39.8	11.8	18.2	12.6	72.2	56.7
DetPro [5]		34.9	53.8	37.4	12.1	**18.8**	12.9	74.6	57.9
OV-DETR [39]	Deformable DETR [45]	38.1	**58.4**	41.1	-	-	-	**76.1**	59.3
DK-DETR		**39.4**	54.3	**43.0**	**12.4**	17.3	**13.4**	71.3	**60.9**

Table 4. **Generalization ability on other datasets.** Following the experimental setting in ViLD [11], we evaluate the LVIS-trained model on COCO validation set, Objects365 validation set and Pascal VOC 2007 test set for comparisons of generalization performance. All methods are trained with ResNet-50 backbone.

SKD	RKD	Aux.Q	AP_r	AP_c / AP_f / AP
			17.0	30.0 / 36.9 / 30.5
✓		✓	20.0 (+3.0)	30.8 / 37.4 / 31.5
	✓	✓	19.8 (+2.8)	30.9 / 37.6 / 31.5
✓	✓		19.7 (+2.7)	27.3 / 35.6 / 29.2
✓	✓	✓	**20.4** (+3.4)	30.9 / 37.6 / 31.7

Table 5. **Ablation experiments:** effectiveness of proposed two types of distillations and auxiliary queries. Aux.Q denotes auxiliary queries. The first row serves as a baseline that only assembles Deformable DETR with a text-based classifier to perform open-vocabulary object detection.

SKD Loss	AP_r	AP_c / AP_f / AP
-	17.0	30.0 / 36.9 / 30.5
L_1	18.9 (+1.9)	30.8 / 37.3 / 31.4
BCE	**20.0** (+3.0)	30.8 / 37.4 / 31.5

Table 6. **Ablation experiments:** impact of semantic distillation loss type. The first row serves as the baseline. Auxiliary queries are used for distillation.

drop for base categories. If we replace the L_1 loss with BCE-based distillation loss, the performance can be further improved by 1.1 points. This validates the effectiveness of our SKD implementation, which not only pulls together features belonging to the same object, but also pushes away features from different objects.

Implementation of RKD. Relational knowledge distillation aims to model and imitate the relationship between individual objects from the input image. Rather than SKD that directly transfers the knowledge from the VLM image encoder, RKD carries out another way to exploit the information of novel categories. A simple relationship between objects can be formulated as the cosine similarity between any pair of objects within the same input image, which we call "image". As mini-batch input is commonly used for model training, we can also model the relationship between two objects that come from different images in a mini-batch, which is denoted as "batch". Furthermore, we can combine

Relation type	AP_r	AP_c / AP_f / AP
-	17.0	30.0 / 36.9 / 30.5
image	19.5 (+2.5)	30.8 / 37.4 / 31.3
batch	19.2 (+2.2)	31.2 / 37.6 / 31.6
image+batch	**19.8** (+2.8)	30.9 / 37.6 / 31.5

Table 7. **Ablation experiments:** impact of the relational type for RKD. The first row serves as the baseline. "image" denotes distilling relationships between objects from the same input image, and "batch" indicates that objects come from different images in a batch for training.

these two types of RKD. As shown in Table 7, both two types of RKD could effectively improve AP_r scores and there is merely slight performance difference between "image" and "batch". Combining two types of RKD leads to a little more performance boost for novel categories, so we use this combined variant for RKD by default.

5. Conclusion

This paper presents a simple yet effective distilling framework for the end-to-end open-vocabulary object detector, termed DK-DETR. With a group of auxiliary queries, we propose two ingenious knowledge distillation schemes based on a DETR-like detector, and effectively transfers the knowledge from the VLM to the detector. The distillation branch in DK-DETR is only used for training, and would not introduce any cost during inference. DK-DETR surpasses existing OVOD methods under the setting that the base-category supervision is solely available.

Acknowledgments

This work is supported by National Key R&D Program of China under Grant No. 2023YFE0204200, the Fundamental Research Funds for the Central Universities under Grant No. 226-2022-00051, and National Key R&D Program of China under Grant No. 2022ZD0160101.

References

[1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 6

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 2, 3, 4, 6

[3] Qiang Chen, Xiaokang Chen, Gang Zeng, and Jingdong Wang. Group detr: Fast training convergence with decoupled one-to-many label assignment. *arXiv preprint arXiv:2207.13085*, 2022. 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 6

[5] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 1, 3, 7, 8

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 6, 7

[7] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *The European Conference on Computer Vision*, 2022. 1, 2, 3, 6, 7

[8] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2918–2928, 2021. 6

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1

[11] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1, 2, 3, 4, 6, 7, 8

[12] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 3, 6, 7

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1

[16] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. *arXiv preprint arXiv:2207.13080*, 2022. 3

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 9, 2015. 6

[18] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022. 3

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 7, 8

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[22] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14074–14083, 2022. 1, 2, 7

[23] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. *ECCV*, 2022. 1

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[26] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2, 7

[27] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[29] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[30] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2, 7

[31] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 6, 7

[32] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11069–11078, 2022. 1

[33] Dahu Shi, Xing Wei, Xiaodong Yu, Wenming Tan, Ye Ren, and Shiliang Pu. Inspose: instance-aware networks for single-stage multi-person pose estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3079–3087, 2021. 1

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4

[35] Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9697–9706, 2023. 1

[36] Johnathan Xie and Shuai Zheng. Zsd-yolo: Zero-shot yolo detection using vision-language knowledgedistillation. *arXiv preprint arXiv:2109.12066*, 2021. 1, 2, 3, 7

[37] Yi Yang, Yueting Zhuang, and Yunhe Pan. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering*, 22(12):1551–1558, 2021. 7

[38] Xiaodong Yu, Dahu Shi, Xing Wei, Ye Ren, Tingqun Ye, and Wenming Tan. Soit: Segmenting objects with instance-aware transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3188–3196, 2022. 6

[39] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 106–122. Springer, 2022. 1, 6, 7, 8

[40] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 3

[41] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020. 2, 7

[42] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 159–175. Springer, 2022. 1, 3, 6, 7

[43] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 1, 7

[44] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 350–368. Springer, 2022. 1, 2, 6, 7

[45] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1, 2, 3, 5, 6, 7, 8