# Decoupled DETR For Few-shot Object Detection

Zeyu Shangguan[0000−0003−1435−6959], Lian Huai *, Tong Liu, Yuyu Liu, and Xingqun Jiang

BOE Technology Group Co. Ltd., Beijing, China
{shangguanzeyu,huailian,liutongcto,liuyuyu,jiangxingqun}@boe.com.cn

**Abstract.** The efficient technique for dealing with severe data-hungry issues in object detection, known as Few-shot object detection (FSOD), has been widely explored. However, FSOD encounters some notable challenges such as the model's natural bias towards pre-training data and the inherent defects present in the existing models. In this paper, we introduce improved methods for the FSOD problem based on DETR structures: (i) To reduce bias from pre-training classes (*i.e.* many-shot base classes), we investigate the impact of decoupling the parameters of pre-training classes and fine-tuning classes (*i.e.* few-shot novel classes) in various ways. As a result, we propose a "base-novel categories **de**coupled **DETR** (DeDETR)" network for FSOD. (ii) To further improve the efficiency of the DETR's skip connection structure, we explore varied skip connection types in the DETR's encoder and decoder. Subsequently, we introduce a unified decoder module that dynamically blends decoder layers to generate the output feature. Our model's effectiveness is evaluated using PASCAL VOC and MSCOCO datasets. Our results indicate that our proposed module consistently improves performance by 5% to 10% in both fine-tuning and meta-learning frameworks and has surpassed the top scores achieved in recent studies.

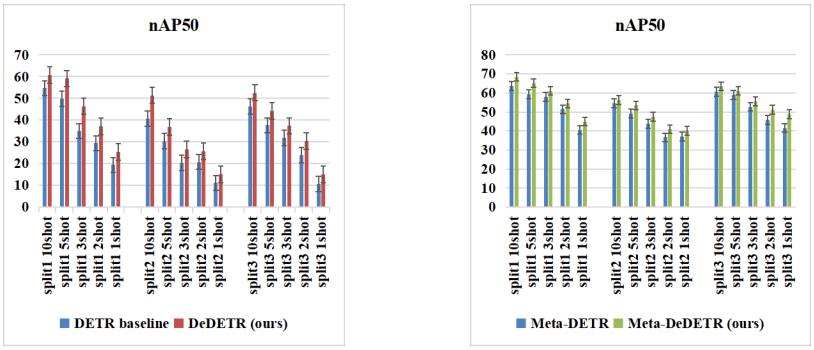
**Keywords:** Few-shot learning · Object detection · Transformer

## 1 Introduction

Few-shot learning (FSL) is designed to create a highly adaptable deep learning model capable of handling situations where training samples are extremely scarce and often previously unobserved. This process imitates human infant learning, as they can rapidly acquire new knowledge with minimal instruction, based on their already extensive prior knowledge. In practice, few-shot learning has a broad range of promising applications, such as industrial defect detection, medical image analysis, archaeological research, landform change detection, environmental protection, *etc.* [14, 21]

Few-shot object detection(FSOD) is an important task in few-shot learning and holds practical significance in various scenarios. Since TFA [32], FSOD has made significant progress based on the Faster RCNN (FRCN) [25] baseline.

* Corresponding author




(a) fine-tune (b) meta-learning

Fig. 1: Our decoupled DETR achieves stable improvements under both fine-tuning and meta-learning paradigms. Histograms demonstrate our experimental results for average precision of novel categories (nAP50) on the three few-shot PASCAL VOC data splits.

There are two commonly used training paradigms, namely, meta-learning and fine-tuning. Current methods utilizing FRCN in meta-learning and fine-tuning training paradigms have achieved competitive results. The emergence of the detection transformer (DETR) [3] in 2020 has further improved the framework for general object detection. This end-to-end set prediction-based object detection framework has not only outperformed traditional anchor-based methods (FRCN, YOLO [1], *etc.*) but has also been widely applied to more sub-tasks of object detection, including instance segmentation and FSOD.

As a result, FSOD based on DETR is considered a new trend, not only due to the simplicity of the DETR framework but also because of its homology to the Transformer, which makes it easier to combine with other Transformer-based tasks, especially for multi-modal language-vision tasks [8]. Meanwhile, recent literature has demonstrated that DETR has achieved outstanding performance in FSOD. Meta-DETR [37] was the first to explore FSOD based on the meta-learning paradigm, while FsDETR [2] was the first to explore FSOD without retraining. However, the development of FSOD based on DETR is still in its early stages.

In prior literature, FSL has been described as an issue of extreme sample imbalance or long-tail due to sufficient base samples in the pre-training stage but insufficient novel samples in the fine-tuning stage [23]. This makes the model tend to have a bias towards the base classes, and our goal is to address this issue. FSCE [29] and FSRC [26] have pointed out that the poor performance of FSOD is more related to inaccurate classification than inaccurate positioning. We have observed that this phenomenon occurs not only in RCNN structures but also in DETR structures. Even with DETR, the focus remains on solving the problem of inaccurate classification. We argue that the extreme sample imbalance of FSOD results in the dominance of old knowledge from data-abundant classes in

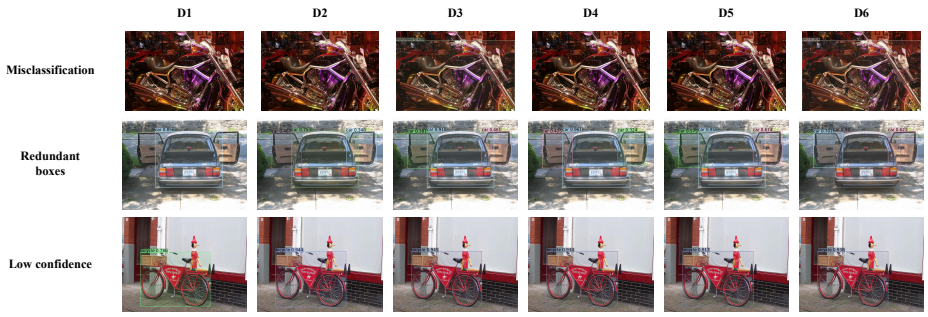



Fig. 2: The intermediate layer of decoder may have better outputs comparing to the last layer of decoder, in terms of classification, boxes regression and prediction confidence.

parameter optimization, even during fine-tuning. This means that the model will always have a certain bias toward the data-abundant classes. To overcome this problem, we have proposed a decoupling module (DeMod), which aims to add an intervene at the low-dimensional feature stage of the model to enhance the model's concentration. Specifically, the old and novel classes have independent feature-extracting modules. Therefore, during training, the basic features learned by the model for old and new categories will not be mixed together, reducing the bias towards old categories.

Furthermore, the inherent defect of the current model poses a significant bottleneck. In DETR, the encoder and decoder are responsible for feature encoding and object decoding, respectively. This involves a process from shallow to deep and then back to shallow. In the traditional transformer structure, the feature transmission from the encoder to the decoder is linear, meaning that the decoder will only use the output of the last layer of the encoder as input. We hypothesize that this connection is inefficient because a shallow encoder might match a shallow decoder better, and vice versa. Therefore, we propose a method of skip connections between the encoder and decoder, which can effectively utilize the intermediate output of previous encoders at each decoder layer. In addition, SQR [4] has pointed out that it is not only the final layer of the decoder that produces the correct prediction results, the output of the middle layer of the decoder sometimes produces better results. Therefore, we also conduct corresponding experiments in the FSOD scenario and indeed find that the intermediate layer of DETR predicted better results than the last layer when fine-tuning, as shown in Fig. 2. Therefore, we attempt to use the decoder output in an adaptive way to decide which layer to emphasize as the output. Specifically, we design an adaptive decoder fusion strategy so that the final output of the decoder module is determined by the weighting of the middle layer, instead of only relying on the output of the last layer. With the help of these modules, we could achieve significant improvement on the commonly used PASCAL VOC and COCO dataset, as shown in Fig. 1. Overall, our main contributions include:

- We propose a decoupling module for novel and base categories that could effectively reduce the **biasing influence** of the existing categories on the new class. This approach leads to the most significant and robust improvement.
- To improve the baseline model, We discuss and simplify the skip connections between the encoder and decoder that does not require an extra learnable module and is competitive with the full skip connections. Also, We design an unified adaptive decoder fusion strategy that dynamically determine the final output based on the weights of all the middle decoder layers without the need for manual decoder layer selection tactics.
- Our approach has been demonstrated to be effective and reliable in both fine-tune and meta-learning paradigms, and our results achieve SOTA on meta-learning.

## 2 Related works

### 2.1 Few-shot Object Detection

Few-shot object detection (FSOD) has traditionally been divided into two paradigms: meta-learning and fine-tuning. In recent years, zero-shot paradigms have emerged that do not require fine-tuning [2] and those that require retraining [13], both of which have shown positive results. In this paper, we conduct experiments using both the fine-tuning and meta-learning paradigms.

TFA [32] has standardized the evaluation system of FSOD: during the fine-tuning stage, balanced data samples containing both old and new classes are used. Additionally, three different data divisions are implemented in VOC to evaluate the model's stability [32]. Meta-DETR [37] also adopts this category partitioning approach. However, unlike TFA, which fine-tunes on the balanced samples, it fine-tunes the novel class while still strictly adhering to the $n$-shot settings, with the old class having more samples. In this paper, we also follow the evaluation system of Meta-DETR, using uneven fine-tuning data. FSCE [29] and HTRPN [27, 28] argues that classification is a more critical bottleneck than positioning. Our experiments on the DETR baseline confirm this, and therefore, we are also focusing more on addressing misclassification. FSED [7] introduces a transformer-based class encoding approach to increase the inter-class distance, enabling the model to concentrate on the essential feature information. FM-FSOD [10] discusses the gain from large pre-trained foundation models such as large language model (LLM) [30]. CD-FSOD [6] proposes an adaptive support feature fusion for the meta-learning paradigm. Xu *et al.* [36] develops a generalized model using variational autoencoder to produce a large amount of augmented data. Liu *et al.* [23] proposes that the few-shot problem is an extreme data imbalance issue, which is one of the causes of misclassification. We concur with this viewpoint and have developed our base-novel categories decoupling module.

### 2.2 Detection Transformer

The effectiveness of the model directly affects the localization performance on FSOD task [14]. Complex models can easily be overfitted in the few-shot scenario. Hence, exploring a stronger baseline network for the few-shot scenario can achieve remarkable results. As a representation of a generalized vision transformer, the detection transformer (DETR) is showing its robustness and effectiveness for the object detection task. Recent DETR-based works have surpassed R-CNN and YOLO works on common object detection benchmarks [38]. DETR first converts the traditional object detection task from an anchor-based method into an ensemble prediction problem that no longer needs hand-designed modules such as non-maximum suppression (NMS) and region proposals (RPN) [3]. Such an integrated approach makes the end-to-end object detection task more intuitively perceptual. Additionally, its alignment with the transformer structure brings DETR more potential for expansion into visual-language multi-modal research [8]. The variation works of DETR further promote the development of object detection and demonstrate competitive and even stronger generalizing ability compared to R-CNN based methods [38]. Deformable DETR [41] proposes a deformable attention module that significantly improves the perceptual ability of the model by focusing only on the sampling points near a reference point instead of all the sampling points. DAB-DETR [20] redesigns a 4D anchor to replace the 2D anchor points as the position queries. DN-DETR [18] addresses the slow converging problem of DETR by adding a denoising loss that is trained with perturbed ground-truth labels. DINO [38] further improves the DN-DETR by applying contrastive learning and mixed query selection. Meta-DETR [37] applies DETR in the FSOD task for the first time under the meta-learning paradigm. FsDETR firstly tried to realize the non-retraining paradigm based on DETR [2]. FsDETR [2] explored few-shot DETR in a retraining-free manner. Since DETR has reached the best performance on general object detection task, we build our model based on it with both fine-tuning paradigm and meta-learning paradigm.

DETR, as a strong baseline for object detection, still has some aspects to be improved, especially the ways of its skip connection. As SQR [4] argues that the intermediate layers of the DETR decoder may yield better detection results compared to the last layer. Consequently, it establishes a connection between the current decoder layer and previous layers, allowing the decoder module to recollect previous decoder information. Additionally, DINO [38] designs a look forward twice structure to enable the current DETR decoder layer to connect with the subsequent two layers. These two strategies are similar but do not create a fully connected network among all decoder layers, and they do not offer a unified strategy for determining which layer should serve as the output. Therefore, we aim to design a unified decoder module that can automatically determine how to utilize information from different decoder layers.

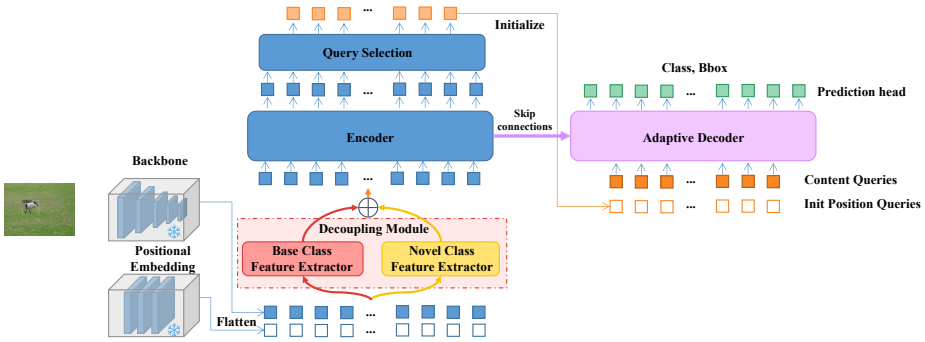



Fig. 3: Overview of our DeDETR. Based on the DETR baseline, the decoupling module is marked in red block. The skip connections operator and the adaptive decoder are marked in pink.

## 3 Proposed Methods

### 3.1 Preliminary

The task of few-shot object detection aims to first pre-train the model on the base classes ($C_B$) where there are sufficient training samples, and then fine-tune it on both base and novel classes ($C_N$) that has only a few training samples for each class. Finally, the fine-tuned model is evaluated on the entire test dataset that includes $C_B \cup C_N$. In the PASCAL VOC few-shot training set, categories containing 1, 2, 3, 5, and 10 instances are named as 1 to 10-shots; 15 categories are selected as base classes, and the other 5 categories are considered novel classes. Similarly, in COCO datasets, the base classes contain 60 categories and the novel classes contain 20 categories under 10 and 30-shot settings. In the meta-learning paradigm, the $n$-shot support set means there are $n$ labeled instances from each few-shot category for training.

### 3.2 Overview

The overview of our method is in Fig. 3. The input image is first sent to both the feature pyramid network backbone and a positional embedding layer simultaneously to extract the visual and positional feature embedding. These flatten features are passed through our **decoupling modules** to generate base-novel-categories-specified features according to the composition of the current training batch. Next, the decoupling module features are sent to the vanilla DETR encoder layers to obtain layer-wise memories and position queries. The encoder memories are then processed by our **skip connections operator** and serve as the input to the decoder layers. Finally, our proposed **adaptive decoder module** integrates each separate encoder layer to generate the final output features, which are sent to the prediction head consisting of two multi-layer perceptrons (MLP) to obtain the classification and box regression results.

Table 1: The value selection strategies for $w$. $N_b$ and $N_n$ indicate the number of base and novel class instances respectively.

| $w$ | Training (Case1) | Training (Case2) | Training (Case3) | Evaluating |
|---|---|---|---|---|
| Hard | 1 | 0 | Fixed $0 \sim 1$ | Fixed $0 \sim 1$ |
| Soft | 1 | 0 | $\frac{N_b}{N_b+N_n}$ | Fixed $0 \sim 1$ |
| Learnable | 1 | 0 | Learnable | Learnable |

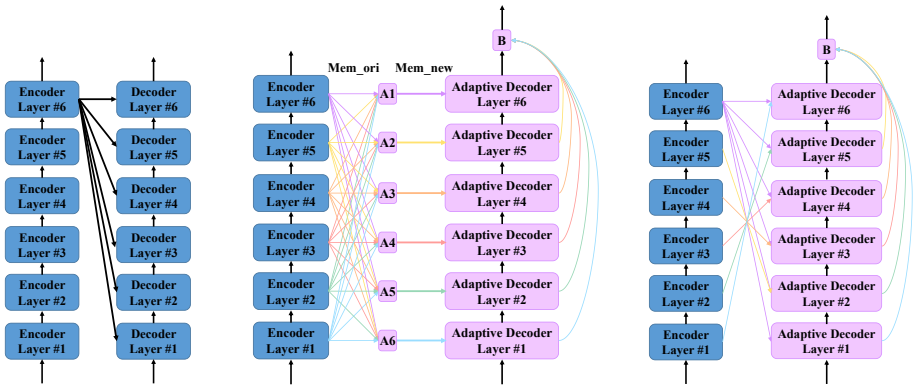
### 3.3 Decoupling Modules (DeMod)

We suggest that during the baseline fine-tuning, the undiscriminating feature fusion of base class and novel class samples will decrease the effect from the novel class in terms of weights updating. Only a few samples from the novel class could hardly push a large model like DETR toward a suitable optimum without any specific operation. Therefore, we propose assigning separate weight sets to function as customized feature extractors for the novel and base classes, which we call decoupling modules.

Specifically, we build two separate deformable self-attention modules (structure from [41]) for the base and novel classes. These are added as input to the transformer encoder. The visual embedding and position embedding are sent to DeMod and processed simultaneously through the base and novel feature-extracting branches. We then check the sample composition of the current training batch and perform a conditional weighting operation: (Case 1) If the current training batch contains only base classes, the output features will come solely from the base feature extracting branch; (Case 2) if the current training batch contains only novel classes, the output features will come solely from the novel feature extracting branch; (Case 3) if the current batch contains both base and novel classes, the output features will be the weighted summation of base and novel feature embedding; as shown in Eq.1, where $x$ is the input feature; $f_{DeMod}(x)$ is the output feature; $f_{b_{pmt}}(x)$ and $f_{n_{pmt}}(x)$ are the base and novel feature embedding respectively; $w$ is the weight of summation.

$$f_{DeMod}(x) = w \cdot f_{b_{pmt}}(x) + (1 - w) \cdot f_{n_{pmt}}(x) \tag{1}$$

The selection of $w$ is based on the occurrence of Case 1, 2, or 3 within a training batch. The value of $w$, as shown in Tab. 1, is set to 1 and 0 for Case 1 and Case 2, respectively. For Case 3, we investigate three methods. (1) Hard coefficient: the weighting $w$ is fixed as a constant between 0 to 1 for both training and evaluation. (2) Soft coefficient: the weight $w$ is determined by the ratio of the number of base instances and novel instances during training, and fixed as a constant during evaluating (here we set it as 0.6 empirically). (3) Learnable coefficient: $w$ is a learnable parameter between 0 to 1 for both training and evaluating through backpropagation. Our experimental results indicate that the soft




(a) DETR baseline. (b) Learnable skip connections. (c) Soft skip connections.

Fig. 4: Three type of connection between encoder and decoder. Block A1 to A6 indicates the parameters as in Eq. 2. Block B refers to the parameters as in Eq. 4.

coefficient strategy performs best, please refer to Sec. 4.6. For the propagation of the 3 cases, please see the Appendix.

### 3.4 Skip Connections Between Encoder and Decoder

The encoder and decoder module of classic DETR are usually composed of 6 self-attention layers respectively. The output of the encoder's last layer (*i.e.* memory embedding), serves as the input for each decoder layer, as depicted in Fig. 4a. There have been a lot of works that discuss the possible ways to connect the encoder layers and decoder layers deeply. As the transformer encoder translates the low-layer features into high-layer features, and the decoder interprets these high-layer features back into low-layer features, creating skip connections between the encoder and decoder would be intuitive and trivial. Lai *et al.* proposes skip connections between the encode and decoder by collecting the outputs of all encoder layers and concatenating them with the output of the decoder layer in a weighted manner [15]. We follow this setting and explore a comparable structure.

We investigate two types of skip connections: learnable connections and soft connections. The learnable connection method includes a set of learnable parameters for the encoder output, as shown in Eq. 2. For each decoder layer, the new input memory embedding is the weighted combination of the original memory embedding from all encoder layers, as shown in Fig. 4b, where $Mem\_new^{\{j\}}$ is the updated encoder memory for decoder layer $j$, and $Mem\_ori^{\{i\}}$ is the original encoder memory from the encoder layer $i$; $A_{ij}$ represents the normalized learnable parameter with a 6 x 6 shape. Each decoder layer $j$ has 6 parameters that weighting the $Out\_enc^{\{i\}}$.

$$Mem\_new^{\{j\}} = \sum_{i}^{i=6} A_{ij} \cdot Mem\_ori^{\{i\}} \tag{2}$$

For the soft skip connections, the new memory embedding only comes from one of the intermediate layers and the last layer of the encoder, as shown in Fig. 4. For example, for a decoder layer $D_j$, the new memory embedding is a weighted summation of the last encoder layer $E_6$ and the corresponding intermediate layer $E_i$, where $i = 6-j$. As shown in Eq. 3, in which $Mem\_new^{\{j\}}$ is the new encoder memory for decoder layer $j$; $Mem\_ori^{\{i\}}$ is the original encoder memory from encoder layer $i$; and weight $A = 1 - 0.15l_i$, where $l_i$ represents an arithmetical layer number (integer), from 0 to 5. Our experiments indicate that the soft skip connections has more advantages over the learnable skip connections, please refer to Sec. 4.6.

$$Mem\_new^{\{j\}} = A \cdot Mem\_ori^{\{6\}} + (1-A) \cdot Mem\_ori^{\{i\}} \tag{3}$$

### 3.5 Adaptive Decoder Selection

As we mentioned in Sec. 1 and Fig. 2, the output of the 5 intermediate layers of the decoder could possibly get better detection results than the last layer. Therefore, we intend to design a scheme that could let the model determine which layer is the final output. Specifically, we design a set of learnable parameters that could be applied to weighting the decoder layers, as shown in Fig. 4c 4b. In detail, we assign a set of normalized coefficients to integrate all of the decoder layer outputs, as shown in Eq. 4, where $Dec\_new$ represents the new decoder output; $Dec\_ori^{\{j\}}$ is the original output from decoder layer $j$; and $B_j$ is the learnable coefficient for each decoder layer $j$. Our experiment results in 4.5 indicate that our adaptive decoder selection is efficient in improving the model's performance.

$$Dec\_new = \sum_{j}^{j=6} B_j \cdot Dec\_ori^{\{j\}} \tag{4}$$

## 4 Experiments

### 4.1 Datasets

Following previous works, we evaluate our few-shot object detection model on the two commonly used datasets: COCO and PASCAL VOC [24, 29, 32, 37]. For COCO dataset, 60 categories are selected as base categories for pre-training, while the other 20 categories are novel categories. For PASCAL VOC dataset, 15 categories are base categories while the remaining 5 categories are defined as novel categories. Specifically, the few-shot PASCAL VOC dataset has three category splits for the purpose of eliminating the contingency while evaluating the model, each data split contains different base and novel category combinations.

Table 2: Few-shot object detection performance on PASCAL VOC dataset. The novel classes nAP50 are evaluated on three separate splits. Our proposed DeDETR reaches new SOTA in most of the scenarios of meta-learning. **The highest nAP50 for each column** are in black bold text, and the second highest scores is in blue text with underline. Sign † indicates the meta-learning paradigm. Sign ⋆ indicates the utilization of an imbalanced few-shot data set.

| Method \ Shot | Backbone | Split1 | | | | | Split2 | | | | | Split3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 10 | 1 | 2 | 3 | 5 | 10 | 1 | 2 | 3 | 5 | 10 |
| TFA w/ cos [32] | FRCN-R101 | 39.8 | 36.1 | 44.7 | 55.7 | 56.0 | 23.5 | 26.9 | 34.1 | 35.1 | 39.1 | 30.8 | 34.8 | 42.8 | 49.5 | 49.8 |
| MPSR [35] | FRCN-R101 | 41.7 | - | 51.4 | 55.2 | 61.8 | 24.4 | - | 39.2 | 39.9 | 47.8 | 35.6 | - | 42.3 | 48.0 | 49.7 |
| FSCE [29] | FRCN-R101 | 44.2 | 43.8 | 51.4 | 61.9 | 63.4 | 27.3 | 29.5 | 43.5 | 44.2 | 50.2 | 37.2 | 41.9 | 47.5 | 54.6 | 58.5 |
| Retentive R-CNN [5] | FRCN-R101 | 42.4 | 45.8 | 45.9 | 53.7 | 56.1 | 21.7 | 27.8 | 35.2 | 37.0 | 40.3 | 30.2 | 37.6 | 43.0 | 49.7 | 50.1 |
| DeFRCN [24] | FRCN-R101 | 40.2 | 53.6 | 58.2 | 63.6 | 66.5 | 29.5 | 39.7 | 43.4 | 48.1 | 52.8 | 35.0 | 38.3 | 52.9 | 57.7 | 60.8 |
| FSOD-UP [33] | FRCN-R101 | 43.8 | 47.8 | 50.3 | 55.4 | 61.7 | 31.2 | 30.5 | 41.2 | 42.2 | 48.3 | 35.5 | 39.7 | 43.9 | 50.6 | 53.5 |
| KFSOD [39] | FRCN-R101 | 44.6 | - | 54.4 | 60.9 | 65.8 | 37.8 | - | 43.1 | 48.1 | 50.4 | 34.8 | - | 44.1 | 52.7 | 53.9 |
| FSRC [26] | FRCN-R101 | 45.5 | 43.4 | 51.1 | 61.4 | 64.0 | 28.4 | 31.3 | 45.0 | 46.1 | 51.6 | 38.8 | 45.1 | 48.4 | 55.5 | 59.0 |
| LVC [13] | FRCN-R101 | **54.5** | 53.2 | 58.8 | 63.2 | 65.7 | 32.8 | 29.2 | **50.7** | 49.8 | 50.6 | 48.4 | 52.7 | 55.0 | 59.6 | 59.6 |
| ⋆ DETR baseline (Our Impl.) | DETR-R101 | 19.4 | 29.4 | 35.0 | 49.8 | 54.7 | 11.1 | 20.8 | 20.4 | 30.4 | 40.6 | 10.6 | 23.9 | 31.9 | 37.6 | 46.3 |
| ⋆ **DeDETR (Our)** | DETR-R101 | 25.3 | 37.2 | 46.4 | 59.1 | 60.8 | 15.1 | 25.6 | 26.5 | 36.9 | 51.4 | 15.1 | 30.4 | 37.3 | 44.1 | 52.6 |
| † TIP [16] | FRCN-R101 | 27.7 | 36.5 | 43.3 | 50.2 | 59.6 | 22.7 | 30.1 | 33.8 | 40.9 | 46.9 | 21.7 | 30.6 | 38.1 | 44.5 | 50.9 |
| † CME [17] | FRCN-R101 | 41.5 | 47.5 | 50.4 | 58.2 | 60.9 | 27.2 | 30.2 | 41.4 | 42.5 | 46.8 | 34.3 | 39.6 | 45.1 | 48.3 | 51.5 |
| † DC-Net [12] | FRCN-R101 | 33.9 | 37.4 | 43.7 | 51.1 | 59.6 | 23.2 | 24.8 | 30.6 | 36.7 | 46.6 | 32.3 | 34.9 | 39.7 | 42.6 | 50.7 |
| † CGDP [19] | FRCN-R2101 | 40.7 | 45.1 | 46.5 | 57.4 | 62.4 | 27.3 | 31.4 | 40.8 | 42.7 | 46.3 | 31.2 | 36.4 | 43.7 | 50.1 | 55.6 |
| † Meta Faster R-CNN [9] | FRCN-R101 | 40.2 | 30.5 | 33.3 | 42.3 | 46.9 | 26.8 | 32.0 | 39.0 | 37.7 | 37.4 | 34.0 | 32.5 | 34.4 | 42.7 | 44.3 |
| † FCT [11] | PVTv2-B2-Li | 49.9 | **57.1** | 57.9 | 63.2 | 67.1 | 27.6 | 34.5 | 43.7 | 49.2 | 51.2 | 39.5 | **54.7** | 52.3 | 57.0 | 58.7 |
| ⋆† Meta-DETR [37] | DETR-R101 | 40.6 | 51.4 | 58.0 | 59.2 | 63.6 | 37.0 | 36.6 | 43.7 | 49.1 | 54.6 | 41.6 | 45.9 | 52.7 | 58.9 | 60.6 |
| ⋆† FM-FSOD [10] | ViT-S | 41.6 | 49.0 | 55.8 | 61.2 | 67.7 | 34.7 | 37.6 | 47.6 | 52.5 | **58.7** | 39.5 | 47.8 | 54.4 | 57.8 | 62.6 |
| ⋆† **Meta-DeDETR (Our)** | DETR-R101 | 44.9 | 54.5 | **61.1** | **65.1** | **68.5** | **40.1** | **41.0** | 47.5 | **53.4** | 56.2 | **48.8** | 51.2 | **55.6** | **61.1** | **63.5** |

As we mentioned in Sec. 2.1, TFA [32] uses balanced $n$-shot base-novel data set where the number of instances for the novel and base classes are same; Meta-DETR [37] evaluates the model based on imbalanced data set where the number of instances for base classes is larger than $n$ (mostly less than $10n$). In our experiment, we follow the imbalanced fine-tuning data set from Meta-DETR.

### 4.2 Training Strategy

We follow the same training strategy as Meta-DETR which uses ResNet-101 as the pre-trained backbone. Our DETR baseline is pre-trained on the base classes with no weights frozen. Then we fine-tune the model on few-shot novel and base classes, only freeze the ResNet-101 backbone. We run the training on 6 M40 GPUs with a batch-size of 1 for fine-tuning and 4 for the meta-learning paradigm. The position query is 900 for fine-tuning paradigm as in DINO [38], and 300 for meta-learning paradigm as in Meta-DETR [37]. The training epoch is 60 with an initial rate of 2e-4.

Table 3: Few-shot object detection performance on COCO dataset. Evaluation for novel classes AP and AP75 are listed. Our results have the highest scores above most previous works. **The highest AP for each column** is in black bold text, and the second highest scores are in regular text with underline. Sign † indicates the meta-learning paradigm. Sign ⋆ indicates the utilization of an imbalanced few-shot data set.

| Method \ Shot | Backbone | Novel AP 10 | Novel AP 30 | Novel AP75 10 | Novel AP75 30 |
|---|---|---|---|---|---|
| TFA w/ cos [32] | FRCN-R101 | 10.0 | 13.7 | 9.3 | 13.4 |
| FSCE [29] | FRCN-R101 | 11.9 | 16.4 | 10.5 | 16.2 |
| SVD [34] | FRCN-R101 | 12.0 | 16.0 | 10.4 | 15.3 |
| SRR-FSD [40] | FRCN-R101 | 11.3 | 14.7 | 9.8 | 13.5 |
| N-PME [22] | FRCN-R101 | 10.6 | 14.1 | 9.4 | 13.6 |
| FORD+BL [31] | FRCN-R101 | 11.2 | 14.8 | 10.2 | 13.9 |
| FSRC [26] | FRCN-R101 | 12.0 | 16.4 | 10.7 | 15.7 |
| ⋆ DETR baseline (Our Impl.) | DETR-R101 | 6.3 | 10.2 | 5.9 | 9.1 |
| ⋆ **DeDETR (Our)** | DETR-R101 | 10.6 | 14.3 | 10.2 | 14.1 |
| † FCT [11] | PVTv2-B2-Li | 17.1 | 21.4 | - | - |
| † Meta Faster R-CNN [9] | FRCN-R101 | 9.7 | 11.3 | 9.0 | 10.6 |
| ⋆† Meta-DETR [37] | DETR-R101 | 19.0 | 22.2 | 19.7 | 22.8 |
| ⋆† **Meta-DeDETR (Our)** | DETR-R101 | **23.2** | **26.3** | **20.6** | **23.1** |

### 4.3 Results on PASCAL VOC

We present our experiment results on PASCAL VOC, as shown in Tab. 2. We distinguish the methods based on fine-tuning and meta-learning. Also, we mark the evaluation scheme on balanced and imbalanced base-novel data sets.

For the meta-learning paradigm, we compared our method with previous SOTA, the results indicate that our method could outperform the previous works in most cases. For the fine-tuning paradigm, we not only report the results of our method but also report our implementation of the DETR baseline on FSOD. Our results could outperform the baseline by up to 10% in all cases.

Even though we could not beat the latest SOTA in the fine-tuning paradigm (based on DINO [38]), our result in the meta-learning paradigm (based on Meta-DETR [37]) could reach the SOTA. This is due to a relatively complex DETR baseline we rely on for the fine-tuning paradigm [38], which has more parameters and is easier to overfit (please see Appendix for model complexity and computation overhead). However, more importantly, we have achieved significant improvements in both paradigms. Specifically, we have improved the fine-tuning paradigm by 10% and the meta-learning paradigm by 5%. These improvements strongly demonstrate the generalization and robustness of our method. Surprisingly, the 10% improvement in the fine-tuning paradigm, even with the complex DINO baseline, suggests that our method is effective in reducing overfitting.

Table 4: Ablation study on our proposed modules based on the fine-tuning DETR baseline and meta-learning Meta-DETR baseline. The decoupling module (DeMod) module gets the highest gain, up to 5%; then the skip connections (SkipCon) module and the adaptive decoder (AdptDec) could get the second and third highest improvement respectively.

| Model | nAP50 | |
|---|---|---|
| | 1-shot | 5-shot |
| DETR baseline (Our Impl.) | 19.4 | 49.8 |
| DETR baseline+DeMod | 22.6 (+3.2) | 55.3 (+5.5) |
| DETR baseline+DeMod+SkipCon | 24.1 (+1.5) | 57.5 (+2.2) |
| DETR baseline+DeMod+SkipCon+AdptDec | **25.3** (+1.2) | **59.1** (+1.6) |
| Meta-DETR baseline (Our Impl.) | 40.3 | 58.9 |
| Meta-DETR baseline+DeMod | 43.6 (+3.3) | 63.4 (+4.5) |
| Meta-DETR baseline+DeMod+SkipCon | 44.3 (+0.7) | 64.3 (+0.9) |
| Meta-DETR baseline+DeMod+SkipCon+AdptDec | **44.9** (+0.6) | **65.1** (+0.8) |

### 4.4 Results on COCO

Our experimental results on COCO dataset are listed in Tab. 3. We evaluate our model on both fine-tuning and meta-learning paradigms, including AP and AP75 for the novel categories. We could observe that our method could get steady improvement on both fine-tuning and meta-learning networks, and we have reached the SOTA results.

### 4.5 Ablation Study

In this part, we mainly discuss the accuracy gain from each of our proposed three modules. The experiments are implemented on the PASCAL VOC 1-shot and 5-shot dataset based on the fine-tuning paradigm. As shown in Tab. 4, we accumulate our proposed modules on the DETR and Meta-DETR baseline. We want to highlight that our decoupling module provides the highest gain for the nAP50, while the skip connections module and the adaptive decoder module could achieve moderate improvement.

We assume that compared with the generalized class-agnostic feature extraction capability enhanced by skip connections and adaptive decoder, the decoupling module can focus more on the distinction between novel and old categories. As mentioned in the introduction, the misclassification of FSOD is to a large extent because it is easy to confuse some categories between the novel and old classes. However, our proposed decoupling module can effectively distinguish the feature embedding of the old and novel classes from the source by physically isolating them at the model weight level during training. Thus the maximum accuracy gain is achieved. This can be seen in more detailed experimental data. We take PASCAL VOC 5-shot split1 as an example and list the respective AP for each novel class, as shown in the Tab. 5, in which the improvement on 'bus' and 'motorbike' is prominent. This situation align with the analysis in FSRC [26]

Table 5: Class level comparison between our model and baseline.

| Model | Bird | Bus | Cow | Motorbike | Sofa |
|---|---|---|---|---|---|
| DETR baseline | 42.4 | 57.7 | 66.1 | 48.6 | 34.3 |
| DeDETR (our) | 48.9 (+6.5) | **69.4 (+11.7)** | 74.1 (+8.0) | **61.8 (+13.2)** | 41.1 (+6.8) |

Table 6: Effect of different $w$ strategies.

| $w$ | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
|---|---|---|---|---|---|---|
| Hard (train and eval) | 39.7 | 53.4 | 52.9 | 39.6 | 38.7 | 35.2 |
| Soft (eval) | 35.4 | 41.3 | 47.2 | **55.9** | 51.5 | 48.8 |
| Learnable | 40.2 | | | | | |

and FSCE [29] that 'bus' (novel) and 'train' (base) are easily confused, while 'motorbike' (novel) and 'bicycle' (base) are easily confused. Such improvements can be also seen in Fig. 5f to Fig. 5l. Thus, our proposed decoupling module is effective in improving the model's ability to recognize novel classes.

### 4.6 Effect of Hyper-parameters

**Different coefficient $w$ for decoupling module** are list in Tab. 6, the experiment is implemented base on PASCAL VOC 5-shot. We observe that the soft coefficient of $w$ could reach the highest nAP, while the hard and learnable coefficients of $w$ are weaker. The principal differences are: when a training batch contains both base and novel samples, the hard and learnable coefficients are unable to perceive the ratio of novel and base samples directly, and therefore hard to assign the proper gradient to the novel and base feature extractor respectively, which makes the model harder to converge. However, the soft coefficient directly assigns the loss energy according to the number of samples, which could achieve better convergence.

**Comparison between soft and learnable skip connections** are listed in Tab. 7. We implement this experiment on PASCAL VOC 1-shot and 5-shot, and we could observe that the gap between the soft and learnable skip connections is negligible. Therefore, to some extent, as the input of the $i^{th}$ decoder layer, the weighted combination of the last layer and corresponding $(6-i)^{th}$ layer of the encoder is sufficient. Even though the learnable full skip connections that utilize all of the encoder layers could achieve higher AP50, considering the newly introduced extra model parameters, such marginal improvement is not a desirable trade-off. Therefore, we recommend that future works use our explored soft skip connections, which are simple yet effective.

### 4.7 Detection Results

We conduct the inference on the PASCAL VOC test set, as illustrated in Fig. 5 with the confidential threshold set as 0.3 for all images. Fig. 5a to Fig. 5c display

Table 7: Comparison to the soft and learnable skip connections.

| | Soft | Learnable |
|---|---|---|
| 1-shot nAP | 24.0 | 24.1 (+0.1) |
| 5-shot nAP | 57.3 | 57.5 (+0.2) |

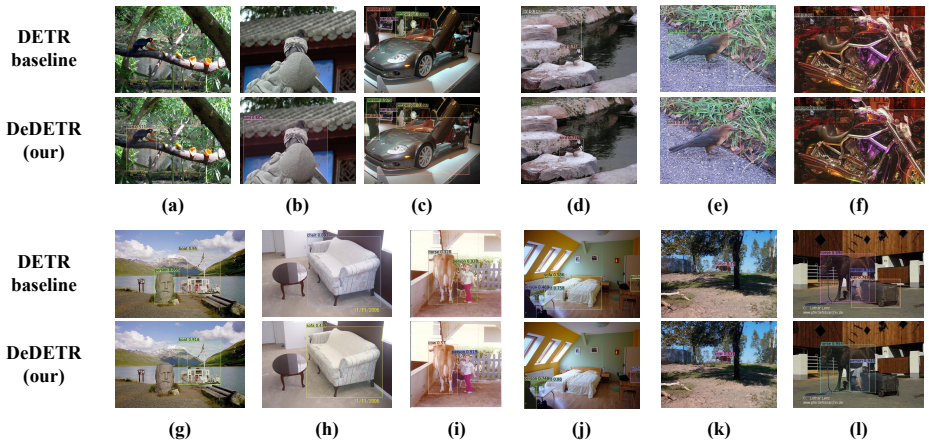



Fig. 5: The visualization of our detection results comparing to the DETR baseline on PASCAL VOC test set.

our improvements concerning missing detection. Fig. 5d and Fig. 5e demonstrate our improvement in terms of incorrect box regression. Additionally, our progress on misclassification is demonstrated in Fig. 5f through Fig. 5l.

## 5 Conclusion

To further improve the accuracy of few-shot object detection, we propose improvements targeting sample imbalance and feature propagation. To counter model bias towards pre-existing knowledge, our decoupling module demonstrates that the weight separation strategy effectively reduces bias from data-abundant classes. For an improved model structure, we also introduce simplified soft skip connections between the encoder and decoder, offering competitive performance against dense skip connections. Additionally, we propose the effective utilization of each decoder layer by adaptively fusing intermediary decoder layers for output generation. After testing on popular datasets like PASCAL VOC and MS COCO, our model consistently outperformed its counterparts with a performance boost of 5-10%.

**Acknowledgments.** There are no acknowledgments for this study.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References


1. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv: Computer Vision and Pattern Recognition (2020)
2. Bulat, A., Guerrero, R., Martinez, B., Tzimiropoulos, G.: Fs-detr: Few-shot detection transformer with prompting and without re-training. In: ICCV. pp. 11793–11802 (October 2023)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 213–229. Springer International Publishing, Cham (2020)
4. Chen, F., Zhang, H., Hu, K., Huang, Y.K., Zhu, C., Savvides, M.: Enhanced training of query-based object detection via selective query recollection. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23756–23765 (2023). https://doi.org/10.1109/CVPR52729.2023.02275
5. Fan, Z., Ma, Y., Li, Z., Sun, J.: Generalized few-shot object detection without forgetting. In: CVPR. pp. 4525–4534 (2021)
6. Fu, Y., Wang, Y., Pan, Y., Huai, L., Qiu, X., Shangguan, Z., Liu, T., Kong, L., Fu, Y., Van Gool, L., et al.: Cross-domain few-shot object detection via enhanced open-set object detector. arXiv preprint arXiv:2402.03094 (2024)
7. Guo, X., Yang, H., Wei, M., Ye, X., Zhang, Y.: Few-shot object detection via class encoding and multi-target decoding. IET Cyber-Systems and Robotics **5**(2), e12088 (2023)
8. Han, G., Chen, L., Ma, J., Huang, S., Chellappa, R., Chang, S.F.: Multi-modal few-shot object detection with meta-learning-based cross-modal prompting (2023)
9. Han, G., Huang, S., Ma, J., He, Y., Chang, S.F.: Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 780–789 (2022)
10. Han, G., Lim, S.N.: Few-shot object detection with foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28608–28618 (June 2024)
11. Han, G., Ma, J., Huang, S., Chen, L., Chang, S.F.: Few-shot object detection with fully cross-transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5311–5320 (2022)
12. Hu, H., Bai, S., Li, A., Cui, J., Wang, L.: Dense relation distillation with context-aware aggregation for few-shot object detection. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10180–10189 (2021)
13. Kaul, P., Xie, W., Zisserman, A.: Label, verify, correct: A simple few shot object detection method. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14217–14227 (2022)
14. Köhler, M., Eisenbach, M., Gross, H.M.: Few-shot object detection: A comprehensive survey. IEEE Transactions on Neural Networks and Learning Systems (2023)
15. Lai, Z., Sun, H., Tian, R., Ding, N., Wu, Z., Wang, Y.: Rethinking skip connections in encoder-decoder networks for monocular depth estimation. ArXiv **abs/2208.13441** (2022)
16. Li, A., Li, Z.: Transformation invariant few-shot object detection. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3093–3101 (2021)
17. Li, B., Yang, B., Liu, C., Liu, F., Ji, R., Ye, Q.: Beyond max-margin: Class margin equilibrium for few-shot object detection. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7359–7368 (2021)

18. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: CVPR. pp. 13609–13617 (2022)
19. Li, Y., Zhu, H., Cheng, Y., Wang, W., Teo, C.S., Xiang, C., Vadakkepat, P., Lee, T.H.: Few-shot object detection via classification refinement and distractor retreatment. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15390–15398 (2021)
20. 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. In: ICLR (2022)
21. Liu, T., Zhang, L., Wang, Y., Guan, J., Fu, Y., Zhao, J., Zhou, S.: Recent few-shot object detection algorithms: A survey with performance comparison. ACM Transactions on Intelligent Systems and Technology **14**(4), 1–36 (2023)
22. Liu, W., Wang, C., Yu, S., Tao, C., Wang, J., Wu, J.: Novel instance mining with pseudo-margin evaluation for few-shot object detection. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 2250–2254 (2022)
23. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2532–2541 (2019). https://doi.org/10.1109/CVPR.2019.00264
24. Qiao, L., Zhao, Y., Li, Z., Qiu, X., Wu, J., Zhang, C.: Defrcn: Decoupled faster r-cnn for few-shot object detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8661–8670 (2021)
25. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. pp. 91–99 (2015)
26. Shangguan, Z., Huai, L., Liu, T., Jiang, X.: Few-shot object detection with refined contrastive learning. In: 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI). pp. 991–996 (2023)
27. Shangguan, Z., Rostami, M.: Identification of novel classes for improving few-shot object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3356–3366 (2023)
28. Shangguan, Z., Rostami, M.: Improved region proposal network for enhanced few-shot object detection. Neural Networks p. 106699 (2024)
29. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: Fsce: Few-shot object detection via contrastive proposal encoding. In: CVPR. pp. 7352–7362 (June 2021)
30. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
31. Vu, A.K.N., Nguyen, N.D., Nguyen, K.D., Nguyen, V.T., Ngo, T.D., Do, T.T., Nguyen, T.V.: Few-shot object detection via baby learning. Image and Vision Computing **120**, 104398 (2022). https://doi.org/https://doi.org/10.1016/j.imavis.2022.104398
32. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly simple few-shot object detection. ICML (2020)
33. Wu, A., Han, Y., Zhu, L., Yang, Y.: Universal-prototype enhancing for few-shot object detection. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 9547–9556. IEEE (2021)

34. WU, A., Zhao, S., Deng, C., Liu, W.: Generalized and discriminative few-shot object detection via svd-dictionary enhancement. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 6353–6364. Curran Associates, Inc. (2021)
35. Wu, J., Liu, S., Huang, D., Wang, Y.: Multi-scale positive sample refinement for few-shot object detection. In: European Conference on Computer Vision (2020)
36. Xu, J., Le, H., Samaras, D.: Generating features with increased crop-related diversity for few-shot object detection. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19713–19722 (2023). https://doi.org/10.1109/CVPR52729.2023.01888
37. Zhang, G., Luo, Z., Cui, K., Lu, S., Xing, E.P.: Meta-detr: Image-level few-shot detection with inter-class correlation exploitation. PAMI **45**(11), 12832–12843 (2023)
38. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: The Eleventh International Conference on Learning Representations (2023)
39. Zhang, S., Wang, L., Murray, N., Koniusz, P.: Kernelized few-shot object detection with efficient integral aggregation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19185–19194 (2022)
40. Zhu, C., Chen, F., Ahmed, U., Shen, Z., Savvides, M.: Semantic relation reasoning for shot-stable few-shot object detection. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8778–8787 (2021). https://doi.org/10.1109/CVPR46437.2021.00867
41. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable {detr}: Deformable transformers for end-to-end object detection. In: ICLR (2021)