# Weakly Supervised Few-Shot Object Detection With DETR

**Chenbo Zhang**[1*], **Yinglu Zhang**[1*], **Lu Zhang**[1], **Jiajia Zhao**[2], **Jihong Guan**[3], **Shuigeng Zhou**[1†]
[1]Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, China
[2]Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing
Electro-Mechanical Engineering Institute, China
[3]Department of Computer Science & Technology, Tongji University, China
{cbzhang21,yingluzhang21}@m.fudan.edu.cn, {l_zhang19,sgzhou}@fudan.edu.cn
zhaojiajia1982@gamil.com, jhguan@tongji.edu.cn

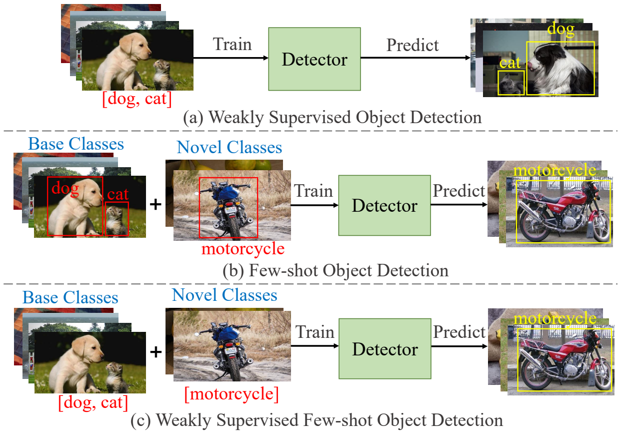
## Abstract

In recent years, Few-shot Object Detection (FSOD) has become an increasingly important research topic in computer vision. However, existing FSOD methods require strong annotations including category labels and bounding boxes, and their performance is heavily dependent on the quality of box annotations. However, acquiring strong annotations is both expensive and time-consuming. This inspires the study on weakly supervised FSOD (WS-FSOD in short), which realizes FSOD with only image-level annotations, i.e., category labels. In this paper, we propose a new and effective weakly supervised FSOD method named **WFS-DETR**. By a well-designed pretraining process, WFS-DETR first acquires general object localization and integrity judgment capabilities on large-scale pretraining data. Then, it introduces object integrity into multiple-instance learning to solve the common local optimum problem by comprehensively exploiting both semantic and visual information. Finally, with simple fine-tuning, it transfers the knowledge learned from the base classes to the novel classes, which enables accurate detection of novel objects. Benefiting from this "pretraining-refinement" mechanism, WSF-DETR can achieve good generalization on different datasets. Extensive experiments also show that the proposed method clearly outperforms the existing counterparts in the WS-FSOD task.

## Instruction

Object detection is a fundamental task in computer vision and has achieved great success in many practical scenarios. Currently, deep learning based techniques such as Faster R-CNN (Ren et al. 2015), YOLO (Redmon and Farhadi 2018), and DETR (Carion et al. 2020) have become mainstream. Typically, these methods rely on substantial amounts of well-annotated data to train models that can accurately recognize and localize the objects. Nevertheless, collecting and annotating such data is extremely expensive and time-consuming, which limits their applications.

In recent years, few-shot object detection (FSOD) has emerged as a promising direction, which aims to achieve effective object detection using only a small amount of annotated data of novel classes. However, to train the FSOD model (Chen et al. 2019), we still have to collect a large amount of strongly annotated training data for the base classes, including the category and the bounding box of each object of each target class in each training image, which incurs huge annotation cost. Moreover, the performance of FSOD models heavily relies on the quality of box annotations. But, due to the complexity of images and the diversity of object morphology, it is difficult to guarantee the quality of the box annotations, which inevitably impacts the performance of the models.



(a) Weakly Supervised Object Detection

(b) Few-shot Object Detection

(c) Weakly Supervised Few-shot Object Detection

Figure 1: A conceptual comparison among general weakly supervised object detection (WSOD), few-shot object detection (FSOD), and weakly supervised few-shot object detection (WS-FSOD). Here, on the left, red texts and boxes are annotations; And on the right, yellow texts and boxes are detected results.

To alleviate the annotation problem, recently a few works have tried to incorporate weakly supervised learning into FSOD (Karlinsky et al. 2021; Shaban et al. 2022; Gao et al. 2019). StarNet (Karlinsky et al. 2021) stands for the first weakly supervised FSOD (WS-FSDO) effort that utilizes a star model to perform non-parametric geometric matching between support and query images. However, its computational complexity will significantly increase when handling

[*]These authors contributed equally.
[†]Corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

high-resolution images, which is a critical issue in object detection. Furthermore, the prediction boxes of StarNet are directly generated by off-the-shelf CAM algorithms (Selvaraju et al. 2017), which makes it inherently face the discriminative region problem in weakly supervised object detection (WSOD). vMF-MIL (Shaban et al. 2022) and NOTE-RCNN (Gao et al. 2019) still need lots of strongly annotated data, so they do not strictly follow the WS-FSOD setting. Fig. 1 illustrates the differences among the three tasks: WSOD, FSOD, and WS-FSOD.

In this paper, we propose a novel WS-FSOD method called **WFS-DETR**, which strictly follows the settings of WS-FSOD, and the whole training process does not require any box annotations. Compared with fully supervised FSOD methods, our method is more practical for real-world scenarios. WFS-DETR adopts a *pretraining-refinement* mechanism. First, an initial object localization network is trained through pretraining. Data augmentation and knowledge distillation are employed to enable DETR to acquire general object localization and integrity judgment capabilities. Then, a progressive refinement network is developed for weakly supervised training and tuning, following the *Base-training + Fine-tuning* paradigm in FSOD.

Specifically, the purpose of our pretraining is to equip the detector with the ability to locate and determine the integrity of foreground objects. Instead of non-parametric proposal generation (e.g. random cropping or selective-search) as in previous works, we train an *Attention-based Localization Network* (ALN) to generate more accurate proposals. With the benefit of long-range modeling by vision transformer, ALN can effectively localize objects in images. Then, we utilize data augmentation to expand the diversity of proposals generated by ALN and jointly train ALN and DETR to distill knowledge to the detector. In the refinement phase, we jointly train DETR with a multiple-instance learning (MIL) structure and perform progressive refinement. Furthermore, we design the refinement strategy by incorporating both category confidence and object evidence. This enhancement provides more accurate supervision information for the refinement process and effectively addresses the problem of discriminative regions.

In summary, the contributions of this paper are as follows: (1) We propose a novel WS-FSOD method, which is the first WS-FSOD work based on DETR. Our method can precisely detect objects of novel classes using solely image-level label supervision. (2) We develop a pretraining-refinement mechanism to address the discriminative region problem in WS-FSOD. Our approach includes a **P**retraining-**D**istillation **L**ocalization **L**earning (**PDLL**) strategy to enhance the model's object localization and integrity judgment capabilities. Notably, **PDLL** stands as the first pretraining strategy tailored exclusively for WSOD. Additionally, we introduce a **D**ual-factor **D**riven **P**rogressive **R**efinement (**DDPR**) strategy, leveraging semantic and visual information to overcome local optimum challenges. (3) We conduct extensive experiments on benchmark datasets, which show that the proposed method significantly outperforms the state-of-the-art methods, validating its effectiveness.

# Related Work

## Weakly Supervised Object Detection

Weakly supervised object detection (WSOD) tries to train an object detector using images with only image-level category labels. Existing WSOD methods can be mainly divided into class activation map based (CAM) methods (Zhou et al. 2016) and multiple-instance learning (MIL) (Maron and Lozano-Pérez 1997) based methods. Recently, MIL-based methods (Bilen and Vedaldi 2016; Tang et al. 2017, 2018; Zeng et al. 2019) have gradually become mainstream. Although previous works have paid great efforts, there are some problems that still have no satisfactory solution, such as discriminative regions and missed detections. These problems may be caused by the "enumerate-select" paradigm for localization and rudimentary refinement strategies.

## Few-Shot Object Detection

Few-shot object detection (FSOD) aims to detect objects with only a few annotated instances. Currently, there are two main FSOD paradigms. Inspired by few-shot learning, most existing few-shot detection methods adopt the periodic *meta-learning paradigm* to transfer knowledge from base classes to novel classes (Kang et al. 2019; Yan et al. 2019; Yang and Renaud 2020; Fan et al. 2020; Hu et al. 2021; Zhang et al. 2021b,a). Recently, some approaches use a simple *fine-tuning paradigm* and demonstrate superior performance (Wang et al. 2020; Sun et al. 2021; Qiao et al. 2021; Fan et al. 2021; Wu et al. 2022; Pei et al. 2022).

## Weakly Supervised FSOD

All existing FSOD methods demand full annotations for both base and novel classes, making data collection labor-intensive and costly. To tackle this, weakly supervised few-shot object detection (WS-FSOD) is introduced, training models with images labeled only at the category level. Despite its increased difficulty, WS-FSOD proves to be more practical. Currently, limited related works exist in this area. However, as we point out in the "Introduction" section, these works (Karlinsky et al. 2021; Shaban et al. 2022; Gao et al. 2019) either cannot handle high-resolution images, or still require a large amount of strongly labeled data, which does not fully follow the WS-FSOD setting.

Different from existing works, our WFS-DETR strictly follows the WS-FSOD setting and is the first work built on Deformable DETR with an innovative *pretraining-refinement* mechanism where the model shows precise object localization capability, instead of relying on non-parametric methods for coarse search as in previous works.

# Method

## Problem Formulation

Given all the classes $\mathcal{C}$, which consist of base classes $\mathcal{C}_b$ and novel classes $\mathcal{C}_n$, we have $\mathcal{C}_b \cup \mathcal{C}_n = \mathcal{C}$ and $\mathcal{C}_b \cap \mathcal{C}_n = \varnothing$. Different from few-shot object detection (FSOD), all training images of both base classes in $\mathcal{C}_b$ and novel classes in $\mathcal{C}_n$ have only image-level category labels. In the whole training data $\mathcal{D}_{train}$, there are abundant base class data $\mathcal{D}_b$, but

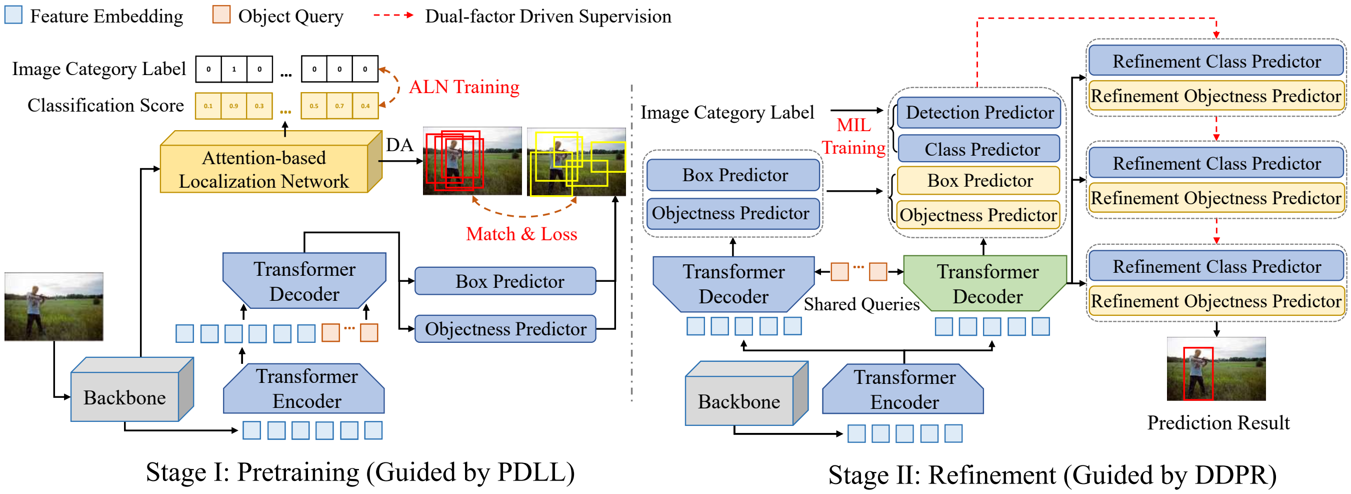


Figure 2: The Framework of WFS-DETR, where PDLL means Pretraining-distillation localization learning strategy and DDPR means Dual-factor driven progressive refinement strategy. The training process consists of a pretraining phase and a refinement phase. During pretraining, we first train an Attention-based Localization Network (ALN) and then distill its localization capability to the detector. In the refinement phase, we refine the predictions via a progressive structure containing $K$ refinement layers by comprehensively utilizing class confidence and object evidence.

a small amount of novel class data $\mathcal{D}_n$ where each class has only dozen of or a few training images. $\mathcal{D}_b \cup \mathcal{D}_n = \mathcal{D}_{train}$ and $\mathcal{D}_b \cap \mathcal{D}_n = \varnothing$. $\mathcal{D}_b$ and $\mathcal{D}_n$ are used for base-training and fine-tuning, respectively. Additionally, box annotations are used only for evaluating the model in the testing phase.

## Framework

The framework of WFS-DETR is shown in Fig. 2, which is based on Deformable DETR. The training process is divided into two stages: the PDLL guided pretraining and the DDPR guided refinement. In pretraining, we use large-scale weakly labeled data (e.g. ImageNet (Deng et al. 2009)) to help the model obtain general object localization capability. Concretely, we first freeze the detector and train an *Attention-based Localization Network* (ALN). Then, we jointly train the localization network and the detector, distilling the localization capability learned by ALN to the detector. In refinement, we follow the common *Base-training* + *Fine-tuning* paradigm in FSOD. We progressively refine the detector by comprehensively leveraging the semantic distinction and foreground integrity of the predictions to obtain more accurate prediction results.

## Pretraining-Distillation Localization Learning

WSOD faces two major problems: inaccurate object localization and missed detection. The scarcity of training data in WS-FSOD makes these problems even worse. These problems are stemmed from poor object localization. Specifically, existing methods mostly follow the “enumerate-select” paradigm to locate objects, that is, first using non-parametric methods (e.g. selective-search or edge box) to generate initial proposal boxes and then selecting a part of these proposals as the object localization results. However, the proposals generated by non-parametric methods are usually of low quality and redundancy. Considering that the initial proposals are critical to WS-FSOD performance, the object localization strategy in previous works actually constrains their performance.

In general object detection, pretraining is widely used to improve the object localization performance of DETR detectors, but in WSOD, the absence of box annotations makes pre-training methods unsuitable (e.g. UP-DETR (Dai et al. 2021) and DetReg (Bar et al. 2022)). Without accurate box supervision, the localization performance of detectors is difficult to be optimized. Consequently, the performance of the WS-FSOD models trained with low-quality pseudo boxes and without any further optimizations will be unsatisfactory.

For effective WS-FSOD, we design a **P**retraining-**D**istillation **L**ocalization **L**earning strategy (**PDLL**), the first pretraining approach tailored exclusively for WSOD. Initially, we train an *Attention-based Localization Network* (ALN) using abundant pretraining data for initial object localization. Then, through *localization distillation learning*, we enhance and transfer the ALN’s localization capability to the detector, endowing the model with general object localization and integrity judgment capabilities.

**Attention-based localization network.** Some works (Gao et al. 2021; Xu et al. 2022; Gupta et al. 2022) have shown that the vision transformer is able to model the entire object well due to its long-range feature dependence, which is ideal for locating objects in the images accurately. However, these works rely on single-scale vision transformers like DeiT (Touvron et al. 2021). As a consequence, they are not directly applicable to object detection tasks, which demand a multi-scale feature-producing backbone for identifying objects of varying sizes. While the swin transformer is often used as a detector backbone, its window attention and sliding window mechanisms

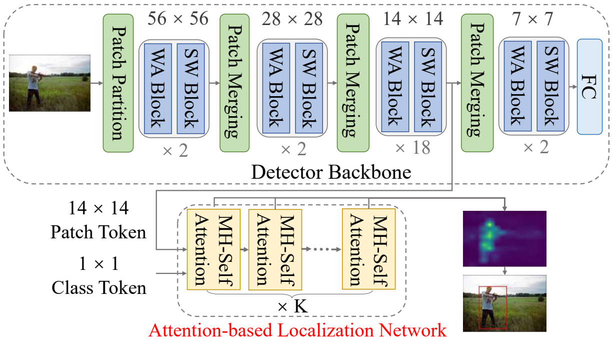


Figure 3: The structure of ALN, which consists of $K$ multi-head self-attention blocks, is inserted after the 3-th stage of the detector backbone (swin transformer).

fall short of capturing comprehensive global information interactions, impeding thorough object modeling. Motivated by prior research, we introduce the *Attention-based Localization Network* (ALN). Comprising multiple multi-head self-attention blocks, ALN serves as a plug-in module compatible with various multi-scale ViT backbones, such as the swin transformer. This integration produces refined *proposal boxes* crucial for effective pretraining and ultimately enhances object detection performance.

As shown in Fig. 3, ALN consists of $K$ multi-head self-attention blocks and is inserted after the 3-th stage of the detector backbone (e.g. swin transformer). A learnable class token $t_c$ is fed into ALN to make information interaction with the patch tokens $t_{ns}$:

$$\begin{aligned} t^*_{c+n} &= Attn_{multi}([t_c, t_{ns}]W_Q, [t_c, t_{ns}]W_K, [t_c, t_{ns}]W_V) \\ &= A^*_{c+n}[t_c, t_{ns}]W_V \end{aligned} \quad (1)$$

where $Attn_{multi}$ is the standard multi-head self-attention (Vaswani et al. 2017). We take the last $N$ columns of $t^*_{c+n} \in \mathbb{R}^{D\times(1+N)}$ from the last block of ALN to obtain patch tokens $t^*_n \in \mathbb{R}^{D\times N}$, and then use a FC layer parameterized by $W_C = [w_1, ..., w_C], w_i \in \mathbb{R}^{D\times 1}$ to map $t^*_n$ to class-aware patch tokens. Finally, the class-aware patch tokens are reshaped to $t^*_{n'} \in \mathbb{R}^{C\times W\times H}$. By optimizing ALN with $\mathcal{L}_{ALN}$, we can assign class semantics to patch tokens:

$$\mathcal{L}_{ALN} = -log(\frac{exp[GAP(T(t^*_n)w_j)]}{\sum_i^C exp[GAP(T(t^*_n)w_i)]}) \quad (2)$$

where $T(\cdot)$ is the matrix transpose operation.

Then, we obtain the average attention map $\bar{A}^*_{c+n} \in \mathbb{R}^{(1+N)\times(1+N)}$ of $K$ blocks and get the class-agnostic attention vector by taking the last $N$ columns of the first row in $\bar{A}^*_{c+n}$ and reshape it to $\bar{A}^* \in \mathbb{R}^{1\times W\times H}$. The attention maps for different classes are generated by multiplying the class-agnostic $\bar{A}^*$ with the class-aware $t^*_{n'}$. Finally, the localization result $Result_{Loc}$ is obtained by applying threshold filtering and the minimum rectangular algorithm to the attention map:

$$Result_{Loc} = MMR(Thr(\bar{A}^* \otimes t^*_{n'})) \quad (3)$$

**Localization distillation learning.** As shown in Stage I of Fig. 2, for a proposal box $b_{ori} = [x1, y1, x2, y2]$ produced by ALN, we use box augmentation (Feng, Zhong, and Huang 2021) to generate a set of augmented proposal boxes around $b_{ori}$. By increasing the diversity of proposal boxes both in shape and position, the proposal boxes are able to cover the whole object as accurately as possible:

$$b_{aug} = [x_1 \pm \alpha_1 * w, y_1 \pm \alpha_2 * h, x_2 \pm \alpha_3 * w, y_2 \pm \alpha_4 * h] \quad (4)$$

where $w = x2 - x1$, $h = y2 - y1$. $\alpha$ is a random number obtained from [0, $\frac{1}{6}$]. This value range ensures that the IoU between the augmented proposal boxes and the original one is greater than 0.5 in all cases, effectively preventing the augmented proposal boxes from deviating from the objects. We define the augmented $M$ proposal boxes as $y = \{(b_i, o_i)\}_{i=1}^M$, where $b_i$ and $o_i$ are the box coordinates and the objectness score of the $i$-th proposal respectively. After augmentation, $y$ can cover different foreground objects well, and we use $y$ as supervision to distill the object localization ability from ALN to DETR. Specifically, first we match the prediction results of DETR $\hat{y} = \{(\hat{b}_i, \hat{o}_i)\}_{i=1}^M$ to $y$ (padded with no object $\varnothing$) via the Hungarian bipartite matching algorithm (Carion et al. 2020):

$$\hat{\sigma} = \underset{\sigma\in\mathfrak{S}_N}{\arg\min} \sum_i^N \mathcal{L}_{match}(y_i, \hat{y}_{\sigma(i)}) \quad (5)$$

where $\mathcal{L}_{match}$ follows DETR (Carion et al. 2020). Using the best matching sequence $\hat{\sigma}$, we calculate the distillation loss $L_{dis}$ as follows:

$$\mathcal{L}_{dis} = \sum_i^N \lambda_{obj}\mathcal{L}_{obj}(o_i, \hat{o}_{\hat{\sigma}(i)}) + 1_{\{o_i \neq \varnothing\}}\lambda_{box}\mathcal{L}_{box}(b_i, \hat{b}_{\hat{\sigma}(i)}) \quad (6)$$

where $\mathcal{L}_{obj}$ is implemented via binary cross entropy loss, and $\mathcal{L}_{box}$ is based on the $Smooth\ L1$ loss and the Distance-IoU loss (DIoU) (Zheng et al. 2020). Compared with GIoU loss (Rezatofighi et al. 2019) used in DETR (Carion et al. 2020), DIoU loss handle the common problem of discriminative regions in WSOD better, i.e., the prediction box occupies only a portion of the GT box.

## Dual-Factor Driven Progressive Refinement

Multiple-instance learning (MIL) paradigm is widely used in WSOD (Bilen and Vedaldi 2016), and some WSOD works try to apply progressive refinement structures with MIL to optimize the detection results layer-by-layer (Tang et al. 2017, 2018). However, these methods are usually trapped by the discriminative region problem because they focus merely on optimizing with classification scores while ignoring the integrity of the objects. To address this challenge, we propose a **D**ual-factor **D**riven **P**rogressive **R**efinement strategy for multiple instance learning (**DDPR**), which tackles this problem by both taking both class confidence and object evidence into account.

**Accurate supervision mining.** In order to alleviate the discriminative region problem, WSOD2 (Zeng et al. 2019) takes superpixel maps as object evidence into account to optimize object localization. However, there exist two main problems: (1) When being applied to different datasets, superpixel maps must be regenerated, which incurs high costs. (2) WSOD2 (Zeng et al. 2019) utilizes selective-search to generate initial proposal boxes, while the principle of selective-search is merging superpixels to generate proposal boxes, which is essentially the same as superpixel maps. As a consequence, it is difficult for superpixel maps to provide effective object evidence for the generation of accurate proposal boxes. As shown in Stage II of Fig. 2, we introduce a novel refinement strategy that does not incur additional cost and leverages the results of pretraining to obtain accurate supervision effectively.

For an image $I$ having only image-level category label $Y = [y_1, ..., y_C]^T \in \mathbb{R}^{C\times1}$, where $y_c = 1$ or 0 indicates the presence or absence of an object class $c$, our basic decoder generates a set of proposal boxes $\hat{\mathcal{P}}^0 = \{(\hat{b}_i^0, \hat{o}_i^0, \hat{c}_i^0, \hat{d}_i^0)\}_{i=1}^N$, where $\hat{b}_i^0 \in \mathbb{R}^{1\times4}$, $\hat{o}_i^0 \in \mathbb{R}^{1\times1}$, $\hat{c}_i^0 \in \mathbb{R}^{1\times C}$ and $\hat{d}_i^0 \in \mathbb{R}^{1\times C}$ indicate the box coordinates, objectness score, classification score, and detection score respectively. As shown in Eq. (7), we can obtain the image-level classification result $\hat{Y} = [\hat{y}_1, ..., \hat{y}_C]^T \in \mathbb{R}^{C\times1}$ of image $I$ by aggregating classification scores $\{\hat{c}_i^0\}_{i=1}^N$ and detection scores $\{\hat{d}_i^0\}_{i=1}^N$ of $N$ proposal boxes together:

$$\hat{Y} = \textstyle\sum_{i=1}^N \left\{ \frac{e^{\hat{c}_i^0}}{\sum_{k=1}^C e^{\hat{c}_{ik}^0}} \odot \frac{e^{\hat{d}_i^0}}{\sum_{i=1}^N e^{\hat{d}_{ik}^0}} \right\} \tag{7}$$

As shown in Eq. (8), the basic MIL classifier is optimized by $\mathcal{L}_{mil}$:

$$\mathcal{L}_{mil} = \textstyle\sum_{c=1}^C \{y_c \log \hat{y}_c + (1 - y_c) \log(1 - \hat{y}_c)\} \tag{8}$$

In order to produce more accurate proposal boxes, we construct $K$ refinement layers to generate refined predictions $\hat{\mathcal{P}}^j = \{(\hat{b}_i^j, \hat{o}_i^j, \hat{c}_i^j)\}_{i=1,j=1}^{N,K}$ from the proposal boxes $\hat{\mathcal{P}}^0$ generated by the basic refinement decoder.

After pretraining, our detector is equipped with the class-agnostic localization capability and is able to produce objectness scores to evaluate the completeness of the predicted foreground object boxes. This process is of no additional computation cost, and the objectness score has excellent generalization. Therefore, we utilize the objectness score as object evidence and then consider the class confidence as well as the object evidence comprehensively to select more precise supervision proposals. In the $(k-1)$-th refinement layer, for proposal $\hat{p}_i^{k-1}$, we first calculate its selection score $\mathcal{S}_i^{k-1} = \hat{o}_i^{k-1} * \hat{c}_i^{k-1}$ by using the class confidence and the object evidence. And then all the proposal boxes will be sorted by the selection scores $\{\mathcal{S}_i^{k-1}\}_{i=1}^N$. After applying NMS (No Max Suppression) to the sorted proposals, we will obtain a set of supervision proposal boxes $\mathcal{P}_s^{k-1} = \{(b_i^{k-1}, o_i^{k-1}, c_i^{k-1})\}_{i=1}^{N_s}$.

Following that, we make a match between $\mathcal{P}_s^{k-1}$ and the proposals $\hat{\mathcal{P}}^k$ generated in the $k$-th refinement layer. For a proposal $\hat{p}_i^k$ in $\hat{\mathcal{P}}^k$, if there exists a set of proposal boxes $\{p_j^{k-1}\}_{j=1}^{n_s}$ in $\mathcal{P}_s^{k-1}$ where each proposal has a IoU score with $\hat{p}_i^k$ over the threshold $\phi$, then the proposal in $\{p_j^{k-1}\}_{j=1}^{n_s}$ has the highest IoU score with $\hat{p}_i^k$ will be selected as the supervision proposal for $\hat{p}_i^k$. All proposals in $\hat{\mathcal{P}}^k$ that successfully match the proposals in $\mathcal{P}_s^{k-1}$ compose the ROI proposal set $\hat{\mathcal{P}}_r^k = \{(\hat{b}_i^k, \hat{o}_i^k, \hat{c}_i^k)\}_{i=1}^{N_r}$. $\mathcal{P}_s^{k-1}$ acts as supervision to $\hat{\mathcal{P}}_r^k$, and the refinement loss $\mathcal{L}_{ref}^k$ is calculated between $\mathcal{P}_s^{k-1}$ and $\hat{\mathcal{P}}_r^k$ to refine the $k$-th class predictor and objectness predictor:

$$\begin{aligned} \mathcal{L}_{ref}^k = -\frac{1}{|N_r|} \sum_{i=1}^{N_r} &(c_j^{k-1} * o_j^{k-1})(CE(c_j^{k-1}, \hat{c}_i^k) \\ &+ BCE(o_j^{k-1}, \hat{o}_i^k)) \\ &where\, match(p_j^{k-1},\, \hat{p}_i^k) \;=\; 1. \end{aligned} \tag{9}$$

With the supervision of $\mathcal{P}_s^{k-1}$, the discriminative information captured by the small proposals will be delivered to the overlapping large proposals, and the object evidence of large proposals will, in turn, be passed to the small ones simultaneously. This bi-directional exchange of information effectively improves detection accuracy.

# Experiments

## Training and Inference Details

**Model training.** Our method follows the pretraining-refinement mechanism. In the pretraining phase, we utilize the pretraining dataset (e.g., ImageNet (Deng et al. 2009)) to train the ALN and then distill its object localization and integrity judgment capabilities into the detector. The learning target is formulated as:

$$\mathcal{L}_P = \lambda_P \mathcal{L}_{ALN} + (1 - \lambda_P)\mathcal{L}_{dis} \tag{10}$$

where $\lambda_P$ is the hyperparameter used to control the learning target. During the first half of pretraining, we set $\lambda_P$ to 1 to train the ALN alone. During the second half of pretraining, we set $\lambda_P$ to 0.5 to jointly train ALN and DETR, distilling the knowledge learned by ALN to DETR. After pretraining, our model is equipped with general object localization and integrity judgment capabilities and is able to be generalized to other datasets without repeating the pretraining.

In the refinement phase, we refine our model on the training dataset. The learning target is formulated as follows:

$$\mathcal{L}_R = \mathcal{L}_{mil} + \lambda_1 \textstyle\sum_{k=1}^K \mathcal{L}_{ref}^k + \lambda_2 \mathcal{L}_{box} \tag{11}$$

where $\lambda_1$,$\lambda_2$ are the hyperparameters used to balance the loss function, and we set $\lambda_1$=1,$\lambda_2$=10. Other hyperparameters of the refinement structure are set following OICR (Tang et al. 2017) (e.g., K=3).

**Model inference.** We use the mean of the classification scores output by $K$ refinement class predictors as the class confidence result and the outputs of the box predictor as the box prediction results.

## Experimental Setting

**Existing benchmarks.** Following the only previous work StarNet (Karlinsky et al. 2021) in WS-FSOD, we takes three benchmark datasets for evaluation: **ImageNetLoc-FS**,

| **Method** | **Dataset** | **1-shot** | | **5-shot** | |
|---|---|---|---|---|---|
| | | $AP30$ | $AP50$ | $AP30$ | $AP50$ |
| MetaOpt+GC (Lee et al. 2019) | **ImageNetLoc-FS** | 32.4 | 13.8 | 51.9 | 22.1 |
| MetaOpt+SS (Lee et al. 2019) | | 16.1 | 4.9 | 27.4 | 10.2 |
| PCL (Tang et al. 2018) | | 25.4 | 9.2 | 37.5 | 11.3 |
| CAN (Hou et al. 2019) | | 23.2 | 10.3 | 38.2 | 12.7 |
| WSOD2 (Zeng et al. 2019) | | 25.8 | 10.9 | 39.8 | 12.8 |
| StarNet (Karlinsky et al. 2021) | | 50.0 | 26.4 | 63.6 | 34.9 |
| **WFS-DETR (ours)** | | **58.4** | **45.3** | **68.3** | **52.8** |
| MetaOpt+GC (Lee et al. 2019) | **CUB-200** | 53.3 | 12.0 | 72.8 | 14.4 |
| MetaOpt+SS (Lee et al. 2019) | | 19.4 | 6.0 | 26.2 | 6.4 |
| PCL (Tang et al. 2018) | | 29.1 | 11.4 | 41.1 | 14.7 |
| CAN (Hou et al. 2019) | | 60.7 | 19.3 | 74.8 | 26.0 |
| WSOD2 (Zeng et al. 2019) | | 47.8 | 16.2 | 54.6 | 18.7 |
| StarNet (Karlinsky et al. 2021) | | 77.1 | 27.2 | 86.1 | 32.7 |
| **WFS-DETR (ours)** | | **84.4** | **42.6** | **92.5** | **53.4** |
| TFA(fully-supervised upper bound) | **PASCAL VOC** | - | 31.4 | - | 46.8 |
| StarNet (Karlinsky et al. 2021) | | 34.1 | 16.0 | 52.9 | 23.0 |
| **WFS-DETR (ours)** | (average over 5-way sets) | 36.2 | 23.2 | 55.3 | 31.5 |

Table 1: Comparison with SOTAs on ImageNetLoc-FS, CUB-200 and PASCAL VOC. GC = GradCAM (Selvaraju et al. 2017), SS = Selective-Search (Uijlings et al. 2013).

**CUB-200** and **PASCAL VOC**. For ImageNetLoc-FS (Eli et al. 2019), we divide the total 331 classes into three sets: 101 base classes for base-training, 214 novel classes for fine-tuning and evaluation, and 16 classes for validation. For CUB (Wah et al. 2011), we split the 200 classes into three sets: 100 base classes for base-training, 50 novel classes for fine-tuning and evaluation, and 50 classes for validation. For PASCAL VOC (Everingham et al. 2010) we divide the total 20 classes into two sets: 15 base classes for base-training, 5 novel classes for fine-tuning and evaluation. In the base-training phase, all base data are used for training. In the fine-tuning phase, we follow the “N-way K-shot” training paradigm in FSOD.

**Implementation details.** We conduct experiments on the Deformable DETR (Zhu et al. 2020) detector with swin transformer-s (Liu et al. 2021) as backbone. Please refer to our supplementary material for more details.

## Comparison with Existing Methods

We compare our method with the WS-FSOD SOTA StarNet (Karlinsky et al. 2021) and other WSOD SOTAs both on ImageNetLoc-FS (Eli et al. 2019), CUB-200 (Wah et al. 2011) and PASCAL VOC (Everingham et al. 2010). The box annotations in the dataset are only used for evaluation. To mitigate the influence of randomness, we conduct numerous 5-way-1/5 shot tests and compute the average as the final result. Please refer to our supplementary material for more training details.

As shown in Tab. 1, our WFS-DETR method outperforms all the compared methods in any shot and all metrics, demonstrating our method’s effectiveness and superiority in WS-FSOD. On ImageNetLoc-FS, for 1-shot, our method surpasses StarNet (Karlinsky et al. 2021) by 8.4 $\%$ and 18.9 $\%$ in terms of $AP30$ and $AP50$. With the growth of the number of shots, our method always keeps advantageous. For 5-shot, our method outperforms StarNet (Karlinsky et al. 2021) by 4.7 $\%$ and 17.9 $\%$ in $AP30$ and $AP50$, respectively. Specifically, our method performs significantly better than the previous SOTA method (Karlinsky et al. 2021) in terms of the $AP50$, which requires higher localization accuracy. Experiments indicate that our method is able to accurately detect the entire object rather than the parts, effectively tackling the most challenging discriminative region problem in WS-FSOD.

We also compare our method with TFA (Wang et al. 2020), a representative method in fully supervised FSOD, on PASCAL VOC. With only image-level category labels, our approach approaches the upper performance bound set by fully supervised TFA, outperforming StarNet.

## Ablation Study

Here we conduct ablations on ImageNetLoc-FS (Eli et al. 2019). To minimize the impact of randomness, we take the average of numerous experiments as the final result.

**Effect of pretraining strategy.** In this study, we investigate the impact of pretraining strategy on model performance. As shown in Tab. 2, we compare the effects of different pretraining strategies on the model’s performance. When the model is pretrained by pseudo boxes generated by non-parametric methods such as random crop, edge box (Zitnick and Dollár 2014), and selective-search (Uijlings et al. 2013), its performance is poor. This could be attributed to the poor quality of the boxes generated by non-parametric methods. The boxes generated bring excessive background noise to the localization pretraining, hindering the detection performance. By utilizing ALN to generate pseudo boxes for

| Pretraining Strategy | 1-shot |  | 5-shot |  |
|---|---|---|---|---|
|  | $AP30$ | $AP50$ | $AP30$ | $AP50$ |
| pretrained $with$ RC | 14.3 | 5.1 | 22.6 | 6.9 |
| pretrained $with$ EB | 21.2 | 8.3 | 30.4 | 9.7 |
| pretrained $with$ SS | 19.6 | 7.4 | 31.2 | 10.2 |
| pretrained $with$ ALN | **58.4** | **45.3** | **68.3** | **52.8** |

Table 2: Ablation study on pretraining strategy. Here, 'RC','EB','SS', and 'ALN' refer to random crop, edge box, selective-search and the attention-based localization network, respectively.

| CC | OE | 1-shot |  | 5-shot |  |
|---|---|---|---|---|---|
|  |  | $AP30$ | $AP50$ | $AP30$ | $AP50$ |
|  |  | 47.3 | 28.9 | 62.4 | 37.2 |
| ✓ |  | 53.4 | 40.2 | 64.7 | 49.6 |
|  | ✓ | 51.2 | 38.3 | 62.8 | 46.1 |
| ✓ | ✓ | **58.4** | **45.3** | **68.3** | **52.8** |

Table 3: Ablation study on refinement strategy. Here, 'CC'and 'OE' mean class confidence and object evidence.

pretraining, the performance is improved significantly. This demonstrates the importance of high-quality pretraining.

**Effect of pretraining proportion.** Here, we examine how the size of the pretraining dataset impacts model performance, considering two factors: the number of images and the diversity of categories. As illustrated in Fig. 4, we divide the pretraining dataset into different groups based on image count and category variety.

From an image standpoint, different groups share the same total categories but differ in image quantities. Regarding categories, different groups have varying category numbers while maintaining uniform images per category. Notably, when the pretraining dataset reaches $40\%$ of the total, performance has been close to the maximum. When a subset $C_s$ of categories $C$ is pretrained with all its images, performance increases with $\frac{C_s}{C}$ until $80\%$, when performance peaks. This highlights the greater impact of training categories over the number of images on performance and encourages efficient pretraining with a subset covering all original $C$ categories.

**Effect of DDPR**. As shown in Tab. 3, we explore the effect of the two factors: class confidence $CC$ and object evidence $OE$. The detector trained without refinement acts as the baseline. The results presented in Tab. 3 show that the improvements brought by individually using the class confidence or object evidence for refinement are close. While by comprehensively considering the optimization of classification and localization, the cooperation of the two factors is able to make effective refinement and significantly improve the model performance.

**Visualization results.** Examples of the detection results are shown in Fig.5. Compared with other methods (Tang et al. 2017; Karlinsky et al. 2021), WFS-DETR locates objects more accurately and solves the problem of discriminative regions, which proves the effectiveness of our method.

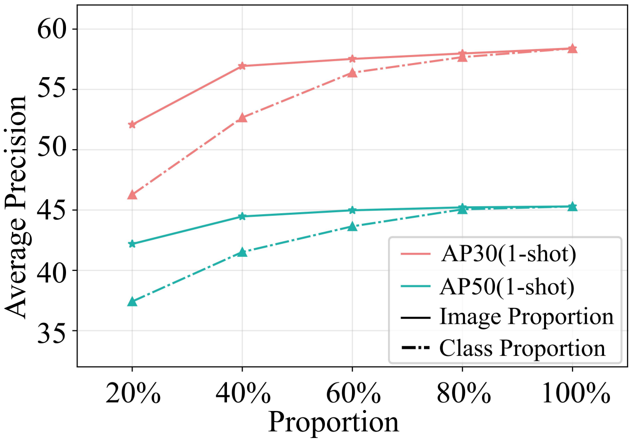


Figure 4: Ablation study on the size of the pretraining dataset. To strictly follow the few-shot setting, we remove all categories contained in the training dataset.

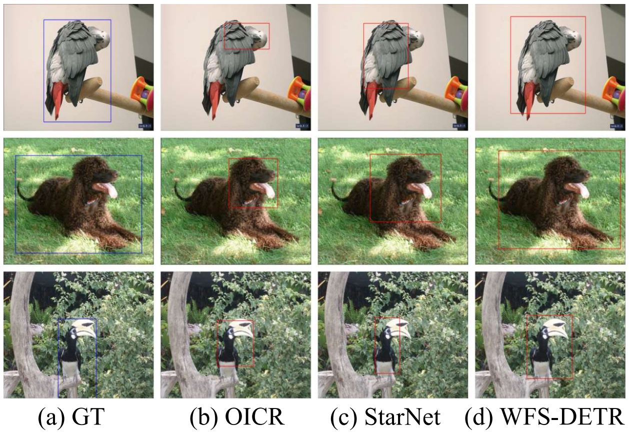

(a) GT (b) OICR (c) StarNet (d) WFS-DETR

Figure 5: Comparison of WS-FSOD results. (a) ground truth. (b) OICR. (c) StarNet. (d) WFS-DETR. The GT boxes are in blue, and the prediction boxes are in red.

## Conclusion

In summary, we propose WFS-DETR, the first WS-FSOD work based on DETR, which leverages a pretraining-refinement mechanism to address the problem of discriminative regions. We enhance the detector's robustness in object localization and integrity judgment using the vision transformer (ViT) with pretraining and knowledge distillation and refine the model progressively by integrating object integrity into the multiple-instance learning (MIL) structure. Experiments on WS-FSOD benchmark datasets show that WFS-DETR achieves state-of-the-art performance, demonstrating the effectiveness of our approach.

## Acknowledgements

Jihong Guan was partially supported by National Key Research and Development Program of China (grant No.2021YFC3300300), and the National Natural Science Foundation of China under Grant U1936205.

## References

Bar, A.; Wang, X.; Kantorov, V.; Reed, C. J.; Herzig, R.; Chechik, G.; Rohrbach, A.; Darrell, T.; and Globerson, A. 2022. Detreg: Unsupervised pretraining with region priors for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14605–14615.

Bilen, H.; and Vedaldi, A. 2016. Weakly supervised deep detection networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2846–2854.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 213–229. Springer.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C. F.; and Huang, J.-B. 2019. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*.

Dai, Z.; Cai, B.; Lin, Y.; and Chen, J. 2021. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1601–1610.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Eli, S.; Leonid, K.; Joseph, S.; Sivan, H.; Mattias, M.; Sharathchandra, a. R. S. F., Pankanti; Abhishek, K.; Raja, G.; and Alexander M., B. 2019. RepMet: Representative-based metric learning for classification and one-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5197–5206.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Fan, Q.; Zhuo, W.; Tang, C.-K.; and Tai, Y.-W. 2020. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4013– 4022.

Fan, Z.; Ma, Y.; Li, Z.; and Sun, J. 2021. Generalized few-shot object detection without forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4527–4536.

Feng, C.; Zhong, Y.; and Huang, W. 2021. Exploring classification equilibrium in long-tailed object detection. In *Proceedings of the IEEE/CVF International conference on computer vision*, 3417–3426.

Gao, J.; Wang, J.; Dai, S.; Li, L.-J.; and Nevatia, R. 2019. Note-rcnn: Noise tolerant ensemble rcnn for semi-supervised object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9508–9517.

Gao, W.; Wan, F.; Pan, X.; Peng, Z.; Tian, Q.; Han, Z.; Zhou, B.; and Ye, Q. 2021. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2886–2895.

Gupta, S.; Lakhotia, S.; Rawat, A.; and Tallamraju, R. 2022. Vitol: Vision transformer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4101–4110.

Hou, R.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2019. Cross Attention Network for Few-shot Classification. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems*, 4005–4016.

Hu, H.; Bai, S.; Li, A.; Cui, J.; and Wang, L. 2021. Dense Relation Distillation With Context-Aware Aggregation for Few-Shot Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10185–10194.

Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; and Trevor, D. 2019. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE International Conference on Computer Vision*, 8420–8429.

Karlinsky, L.; Shtok, J.; Alfassy, A.; Lichtenstein, M.; Harary, S.; Schwartz, E.; Doveh, S.; Sattigeri, P.; Feris, R.; Bronstein, A.; et al. 2021. Starnet: towards weakly supervised few-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1743–1753.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10657–10665.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Maron, O.; and Lozano-Pérez, T. 1997. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10.

Pei, W.; Wu, S.; Mei, D.; Chen, F.; Tian, J.; and Lu, G. 2022. Few-Shot Object Detection by Knowledge Distillation Using Bag-of-Visual-Words Representations. In *Proceedings of the European Conference on Computer Vision*, 283–299.

Qiao, L.; Zhao, Y.; Li, Z.; Qiu, X.; Wu, J.; and Zhang, C. 2021. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 8661–8670.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. 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*, 658–666.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shaban, A.; Rahimi, A.; Ajanthan, T.; Boots, B.; and Hartley, R. 2022. Few-shot Weakly-Supervised Object Detection via Directional Statistics. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3920–3929.

Sun, B.; Li, B.; Cai, S.; Yuan, Y.; and Zhang, C. 2021. Fsce: few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7352–7362.

Tang, P.; Wang, X.; Bai, S.; Shen, W.; Bai, X.; Liu, W.; and Yuille, A. 2018. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 42(1): 176–191.

Tang, P.; Wang, X.; Bai, X.; and Liu, W. 2017. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2843–2851.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Uijlings, J. R.; Van De Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective search for object recognition. *International journal of computer vision*, 104: 154–171.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset 1–15.

Wang, X.; Huang, T. E.; Joseph, G.; Trevor, D.; and Yu, F. 2020. Frustratingly simple few-shot object detection. In *Proceedings of the International Conference on Machine Learning*, 9919–9928.

Wu, S.; Pei, W.; Mei, D.; Chen, F.; Tian, J.; and Lu, G. 2022. Multi-faceted Distillation of Base-Novel Commonality for Few-Shot Object Detection. In *Proceedings of the European Conference on Computer Vision*, 578–594.

Xu, J.; Hou, J.; Zhang, Y.; Feng, R.; Zhao, R.-W.; Zhang, T.; Lu, X.; and Gao, S. 2022. Cream: Weakly supervised object localization via class re-activation mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9437–9446.

Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; and Lin, L. 2019. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 9577–9586.

Yang, X.; and Renaud, M. 2020. Few-shot object detection and viewpoint estimation for objects in the wild. In *Proceedings of the European Conference on Computer Vision*.

Zeng, Z.; Liu, B.; Fu, J.; Chao, H.; and Zhang, L. 2019. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8292–8300.

Zhang, G.; Luo, Z.; Cui, K.; Lu, S.; and Xing, E. P. 2021a. Meta-DETR: Image-Level Few-Shot Detection with Inter-Class Correlation Exploitation. arXiv:2103.11731.

Zhang, L.; Zhou, S.; Guan, J.; and Zhang, J. 2021b. Accurate few-shot object detection with support-query mutual guidance and hybrid loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14424–14432.

Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; and Ren, D. 2020. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12993–13000.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Zitnick, C. L.; and Dollár, P. 2014. Edge boxes: Locating object proposals from edges. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 391–405. Springer.