IRA-FSOD: Instant-Response and Accurate Few-shot Object Detector

Junying Huang, Junhao Cao, Liang Lin, and Dongyu Zhang *
Sun Yat-sen University

***Abstract*—Aiming at recognizing and localizing objects of novel categories with just a few reference samples, few-shot object detection (FSOD) is quite a challenging task. Previous works rely heavily on the fine-tuning process to transfer their models to the novel categories. They are flawed in the real application since the fine-tuning process is time-consuming and it suffers from serious deterioration on the low-quality support set. Based on the observation, this paper proposes an instant-response and accurate few-shot object detector (IRA-FSOD) that can detect the objects from novel categories without fine-tuning. We carefully analyze the limitations of widely-used Faster R-CNN and transform it to IRA-FSOD. Specifically, we first propose a novel semi-supervised Region Proposal Network (SS-RPN) module and a switch classifier module to precisely recognize the potential foreground instances from novel categories without fine-tuning. Moreover, we introduce two explicit inference strategies into the localization module, including explicit localization score and semi-explicit box regression, to alleviate over-fitting towards the base categories. Extensive experiments demonstrates that the proposed IRA-FSOD not only accomplish few-shot object detection with the instant-response, but also reaches state-of-the-art performance under various FSOD protocols and settings.**

***Index Terms*—Deep learning, object detection, Few-Shot Object Detection, Few-Shot Learning, Instant-Response.**

I. Introduction

Deep-learning-based methods have achieved remarkable success in various computer vision tasks. However, the generalization ability of these methods towards the open domains is quite limited as the training data is scarce. This triggers active research on few-shot learning, which aims to develop models that can be generalized to the unseen categories with only a few data with annotations.

Specifically, when it comes to the field of few-shot object detection (FSOD), setbacks are frequently encountered due to the complexity of FSOD tasks. Most existing methods [3, 7, 15–17, 29, 38, 39, 45, 47, 51] require fine-tuning on the support set. It must be admitted that fine-tuning is an effective method to solve FSOD problems especially when the models are well pretrained on large dataset and the data in novel domains is limited. But the bottlenecks are quite apparent.

Bottleneck 1: the fine-tuning methods will lead to lots of preparation time during inference. As illustrated in Figure 1, the time spent on fine-tuning varies from 15 minutes to 6 hours. That means each time a brand-new detection task is given, these methods require a long preparation process before it works, which is unacceptable in real-world scenarios.

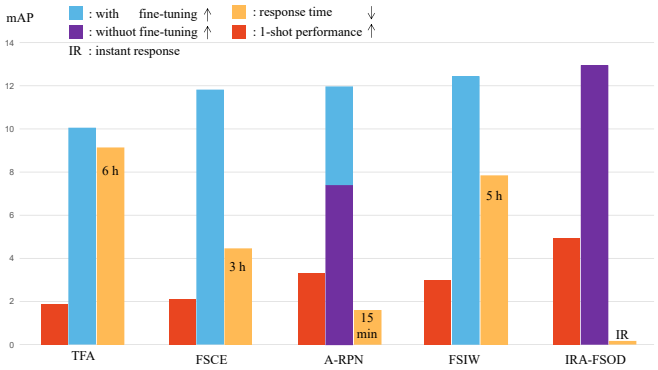


Fig. 1. Comprehensive comparison among different models on MS COCO dataset, including the performance under low-shot setting, performance with/without fine-tuning and response time for fine-tuning. TFA [38], FSIW [47], and FSCE [29] are invalidated before fine-tuning. A-RPN [7] can respond instantly the same as ours, but it performs poorly before fine-tuning. In summary, IRA-FSOD can achieved optimal performance while supporting instant-response (without fine-tuning).

Bottleneck 2: we expect models to dig out countless objects belonging to novel categories by only a few reference samples. But as illustrated in Figure 1, when it comes to 1-shot setting, methods that require fine-tuning seriously degrade. Intuitively speaking, the fine-tuning-based methods cause the model to over-fit the support set with quite limited samples.

Bottleneck 3: from the perspective of meta learning, we expect models to "learn to detect" rather than just "transfer" to the novel dataset. But we can't admit that models relying on fine-tuning have learnt to detect, even with the so-called promising results. Consequently, we hope that the models can be qualified for detection tasks towards novel categories, even without fine-tuning.

Driven by the aforementioned points of view strongly associated with fine-tuning, this paper proposes a novel Instant-Response and Accurate Few-shot Object Detector (IRA-FSOD), which is born out of vanilla Faster R-CNN [27] and is competent for FSOD tasks. We attempt to get rid of the cumbersome fine-tuning to accomplish the goal of "Instant-Response" (IR), that is, without the preparatory work such as fine-tuning, the model itself can directly and instantly detects objects from novel categories. To accomplish this goal, we carefully analyze the components of widely-used Faster R-CNN [27] and make improvements on them. On the whole, the region proposal network (RPN) module, the box classifier and the localization module will be modified.

Firstly, the training mode of the RPN is contradictory to

Corresponding author: Dongyu Zhang, Email: zhangdy27@mail.sysu.edu.cn

Fig. 2. The foreground instances in the training and testing phase are disjoint.

the few-shot learning setting. As shown in Figure 2, the foreground instances are disjoint in training phase and test phase, which means the potential foreground instances not belonging to base categories will be treated as background during the training process. We argue that the RPN module should focus on any potential foreground instances instead of only the annotated instances during the base training. To break the bottleneck, we train the RPN module by remarking the potential novel instances as unlabeled data and leveraging them in a semi-supervised paradigm. To the best of our knowledge, IRA-FSOD is the first work to solve such problem using the semi-supervised learning algorithm.

Secondly, most of the existing box classifiers are difficult to achieve satisfying performance without fine-tuning. In Figure 3, we compare three kinds of widely-used classifiers: (a) The multi-classifier [29, 38, 47] learns a hyper-plane for each base category in the feature space, which shows the best learnability but spends much time on fine-tuning, and it is invalidated on the novel category without fine-tuning; (b) The comparison-classifier [7, 36] learns a class-agnostic binary classifier in the joint feature space. It can directly recognize the objects from novel categories but still suffers from the bias of base categories since its parameters are trained on the base category data; (c) The distance-classifier [17, 28] is non-parameter and performs classification according to the nearest neighbor rule. It doesn't suffer from category bias, but it can't preserve the classification knowledge from training. Based on the observation, we propose a switch classifier module. It switches different classifiers during training and inference to build a box classifier, contributing both generalization and learnability to the IRA-FSOD. It can significantly improve the few-shot performance while supporting instant-response.

Thirdly, the localization module, composed of the localization score calculation and box regression, is supported by implicit fitting and lacks logical inference, which suffers from over-fitting to the distribution of the base categories. So we attempt to improve its generalization and accuracy by introducing explicit logical inferences into the localization module. For the localization score, we introduce the pixel-wise contrast into the box classifier to evaluate it, which can generate the confidence that has a high correlation with the localization result of given region proposal. For the box regression module, we propose a promoted box regressor to strengthen the logical relation between the region feature and its box regression. These logical inferences are category-agnostic, and thus can maintain generalization to novel categories.

Extensive experiments on two large and challenging few-shot detection benchmark datasets, i.e., MS COCO [19] and FSOD dataset [7], show that IRA-FSOD can reach the state-of-the-art FSOD performance while achieving instant-response.

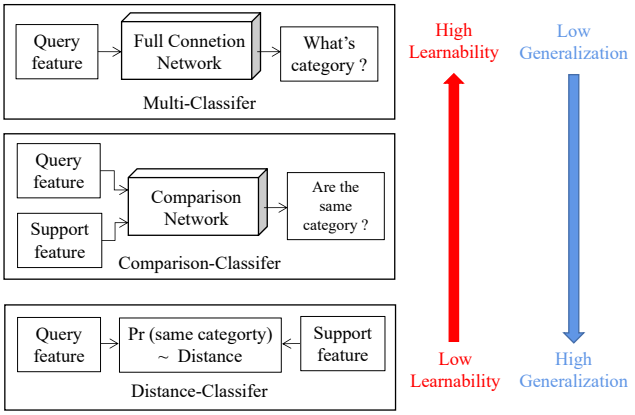


Fig. 3. Comparisons of the motivation, learnability, and generalization ability among three classifiers.

Especially under the extremely-low-shot and class-incomplete setting, it promotes the current state-of-the-art by a large margin even without fine-tuning. **Beyond the satisfying result of getting rid of fine-tuning, IRA-FSOD motivates us to reflect on the inference mechanism that we need in few-shot learning.** In brief, our main contributions can be summarized as follows:

1) We propose IRA-FSOD to get detection of object instances from novel categories without fine-tuning.
2) We optimize the components in Faster R-CNN, including the box classifier, the RPN and the localization module. The optimized model equips both generalization and learnability to handle the open-world detection tasks.
3) By applying the improvements, IRA-FSOD successfully achieves state-of-the-art results in response time, precision, and recall.

II. Related Work

General Object Detection is a fundamental task in computer vision that has attracted lots of attention. Modern object detectors can be divided into two kinds: one-stage detectors and two-stage detectors. One-stage detectors directly predict categories and locations of objects, e.g., YOLO series [1, 24–26], SSD [21], etc. Two-stage detectors, pioneered by R-CNN [12], first generate class-agnostic region proposals, then further refine and classify the proposals [11, 13, 27]. These works heavily rely on a huge amount of annotated data and are invalidated on the data from unseen categories, thus they can not be directly used to solve the FSOD problem.

Few-shot learning aims to recognize novel classes with limited labeled data. Meta-learning methods [9, 22, 23, 30, 31, 48], also named as "learning to learn", are proposed to learn a meta-learner that can adapt to new tasks with a few labeled samples. Distance metric learning methods [18, 28, 34, 35, 41, 46, 55] focus on designing a distance formulation between the samples in an embedding space generated by deep neural networks. Popular metrics include cosine similarity [4, 32, 37], Euclidean distance [28] and graph distance [10].

Few-shot Object Detection is proposed to handle the object detection with only a few annotated samples. There are mainly two types of methods aiming to address the few-shot object

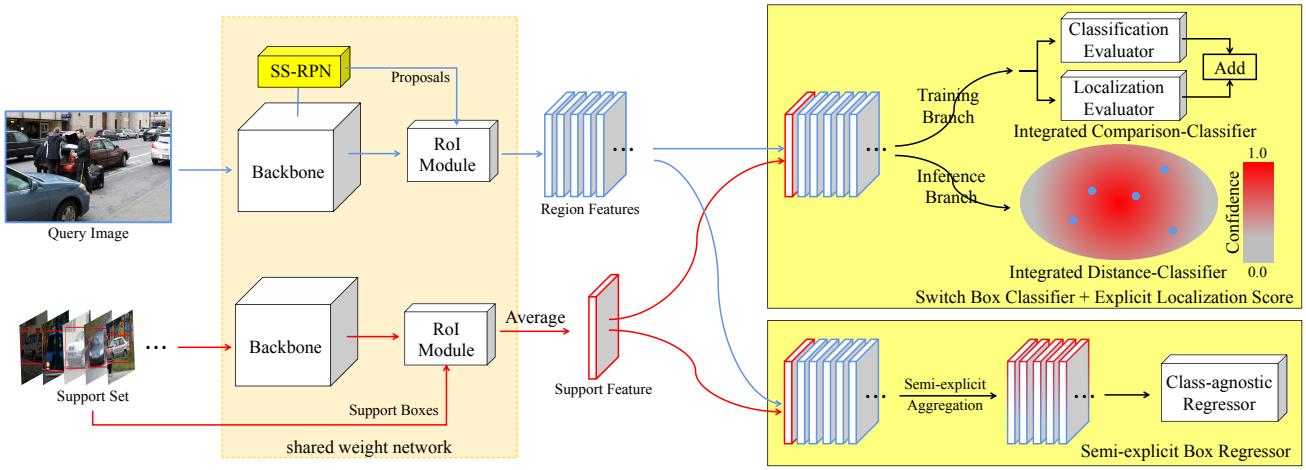


Fig. 4. IRA-FSOD: Given a query image and a support set, the weight-shared backbone first process them into feature maps. Then the RoI module extracts the region features from the region proposals predicted by the semi-supervised RPN, and extracts support features from the bounding boxes in the support set. Finally, the box classifier and box regressor predict whether the region proposal contains the object belonging to the support category and the offset between the region proposal and the ground truth bounding box by comparing the region and support features.

detection problem, i.e., meta-learning-based methods and fine-tuning-based methods:

Meta-learning-based methods attempt to build their few-shot detectors by employing various meta-learning techniques to extract the class-agnostic knowledge or transfer the knowledge from base categories to novel categories. Despite they are called "meta-learning-based", these methods still require a fine-tuning process. Otherwise, they are either invalidated [15, 16, 20, 39, 45, 47, 51] or lagging behind other methods [3, 7, 17, 33, 40, 43, 53, 54]. FSRW [16] extracts generic meta-features from base categories, then adjusts them using the re-weighting features for novel categories. Meta R-CNN [51] and FSIW [47] propose to use the re-weighting features over RoI features instead of the image feature. MetaDet [39] and GenDet [20] propose to estimate the new parameters in the detector for detecting novel category instances. RepMet [17] incorporates distance metric learning into few-shot detection to help classify the proposals. A-RPN [7] and Meta-RCNN [45] propose attention-RPN to generate the class-specific region proposal. A-RPN also proposes a multi-relation detector and a contrastive training strategy. DCNet [15] fully exploits local information to benefit the detection process and alleviates the scale variation problem by context-aware feature aggregation. OSWF [53] focuses on building a stronger connection between the novel and base category data. AirDet [40] proposes to extract the class-agnostic relation with the support images. QA-FewDet [33] applies GCNs to model the class-class, class-proposal, and proposal-proposal graph relationships.

Fine-tuning-based methods focus on improving the fine-tuning process on the support data to effectively transfer the category-specific model to the novel category. They are once suffered from poor performance, but recent works set the new state-of-the-art. TFA [38] simply fine-tunes the last layer of Faster R-CNN [27] but substantially improves the performance. MPSR [44] handles the scale variance issue by multi-scale positive sample refinement, but it needs a manual selection. MI-FSOD [42] focuses on making the model adapt to the unseen categories while avoiding forgetting to detection knowledge from base categories. FSCE [29] builds a strong baseline upon TFA [38] and boosts the performance by large margins. DeFRCN [52] decouples the gradients from the RPN and RCNN, which achieves impressive performance.

III. METHODOLOGY

A. Problem Definition

Given a base dataset $\mathcal{D}_b$ with annotated instances of the base (seen) category $\mathcal{C}_b$, the objective of few-shot object detection is to train a robust model on $\mathcal{D}_b$ which can be generalized on the novel dataset $\mathcal{D}_n$ with instances of the novel (unseen) category $\mathcal{C}_n$ ($\mathcal{C}_n \cap \mathcal{C}_b = \phi$). For each novel category, there is also a support set $\mathcal{S}$ with a few annotated instances. In the previous works, they are used in fine-tuning, but in our work they only need to be used in inference. In more detail, N-way K-shot object detection means $\mathcal{C}_n$ contains N categories and each support set contains K annotated instances (usually less than 10), i.e., $\mathcal{S}_c = \{(I_i, b_i), i = 1, ..., K\}$ where I and b denote the support image and the bounding box of the support instances. The novel dataset is also called the query set.

B. IRA-FSOD Framework

The overview of IRA-FSOD is shown in Figure 4, which is based on Faster R-CNN [27]. The general process is as follows: Given a query image and a support set, the goal is to detect the objects belonging to the support category in the query image. Firstly, the backbone extracts the feature maps of the query image and all support images. Then the RPN module predicts the region proposals in the query image. After that, the RoI module, including RoI-pooling and RoI-extractor, extracts the feature maps of region proposals and the feature maps of all the support instances. Finally, the box classifier and box regressor further predict the category and the box regression of the region proposals, by comparing the region features and the support feature. The support feature is the average of all support instance feature maps.

In particular, we first propose semi-supervised RPN (SS-RPN, Sec. III-C) and switch classifier (Sec. III-D) to adapt to the FSOD task without fine-tuning. Then, we present a novel

explicit localization module (Sec. III-E) to alleviate the over-fitting to base categories. As shown in Figure 4, we mark the corresponding position of these modules in yellow.

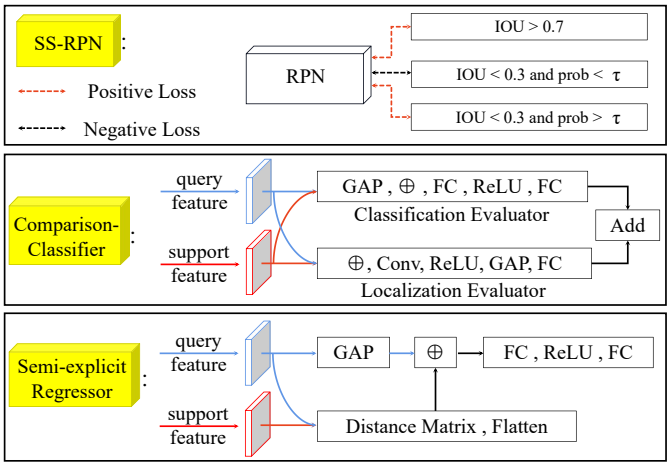


Fig. 5. The detail of components in the IRA-FSOD. The input of the comparison-classifier and semi-explicit box regressor are the feature map of a region proposal (blue) and the support feature map of a category (red). GAP, FC, and Conv mean the global average pooling, fully-connected layer, and convolutional layer. ⊕ means to concatenate the input features.

C. Semi-supervised RPN

Generating region proposals by the class-agnostic detector, such as RPN, is a crucial idea in two-stage detection models, but it has a fatal defect in FSOD. As shown in Figure 2, the potential foreground proposals not belonging to the base categories are easily regarded as background in the base training phase. Thus the RPN module in the FSOD framework is implicitly class-specific to the base categories, which is hard to capture the anchors related to novel categories.

To address the problem, we adopt a semi-supervised algorithm to train the RPN module, which gains satisfying performance without fine-tuning. Concretely, all the positive anchors are certain foreground instances. But the negative anchors actually consist of background and potential object instances from novel categories, so we remark them as unlabeled data. In more detail, as shown at the top of Figure 5, we first annotate the anchors whose Intersection over Union (IoU) with the ground truth (GT) bounding box is less than 0.3 as negative, and the anchors whose IoU is greater than 0.7 as positive, following the standard RPN training process. Then, we annotate the negative anchors with RPN prediction probabilities greater than threshold τ as the hard one-hot pseudo positive label and compute positive loss.

To calculate a balance loss, we keep the ratio of the positive anchors, the negative anchors, and the pseudo positive anchors as 1:1:1. The influence of different choices of threshold τ is shown in Sec. IV-F.

Discussion: In the object detection task, it is unrealistic to achieve both high recall and high precision. The RPN with semi-supervised training inevitably leads to more background proposals in the inference. However, we argue that this is worthy because it is possible to eliminate these background proposals by the following box classifier. On the contrary, the foreground anchors ignored by the RPN module are irreparable, which is an important reason why existing methods rely so much on fine-tuning. As shown in Table I, experiments also demonstrate that although this technique is simple, it significantly boost the object detection performance for the novel category without fine-tuning.

TABLE I
ABLATION EXPERIMENTAL RESULTS FOR SWITCH CLASSIFIER AND SEMI-SUPERVISED RPN ON MS COCO UNDER 10-SHOT SETTING.

Switch Classifier (Training)	Switch Classifier (Inference)	AP	AP_{50}	AP_{75}
Multi	Multi	invalidated		
Comparison	Comparison	4.71	8.79	4.41
Distance	Distance	5.94	15.64	2.82
Multi	Distance	6.89	15.28	5.44
Comparison	Distance	**8.69**	**17.38**	**7.78**
+ Semi-supervised RPN		**10.54**	**20.96**	**9.08**

D. Switch Classifier

As analyzed in Sec. I, we argue that there are drawbacks in three kinds of existing commonly used classifiers in the FSOD task. Therefore, we adopt a switch classifier module. Specifically, it switches different classifiers in training and inference to improve both the generalization to novel categories and the learnability for feature space. In addition, since the distance-classifier is non-parameter, it doesn't require re-training when replacing the trained classifier with the distance-classifier during inference, so we can benefit a lot from distance-classifier as well as not using fine-tuning. Table I shows the ablation study of different classifier combinations on the MS COCO dataset under the instant-response setting and 10-shot one-time FSOD evaluation protocol. For a fair comparison, both the multi-classifier and comparison-classifier are single connection layers, and the distance-classifier is the cosine distance between the region feature and the support feature.

As shown in Table I, the multi-classifier is invalidated on the novel category before re-training. The comparison-classifier performs worse than the distance-classifier since its parameters are still affected by the bias of the base categories. However, the distance-classifier can significantly benefit from models trained by the multi-classifier or comparison-classifier due to their learnability. Based on the results, the IRA-FSOD adopts the comparison-classifier to calculate the loss during training, which can better balance the learnability and generalization on the whole. During inference, it adopts the distance-classifier to calculate the confidence, as shown in Figure 4. The details of the two classifiers will be described in Sec. III-E.

E. Explicit Localization Module

In this section, we focus on the localization module in the framework, including the localization score to evaluate the localization accuracy and the box regression to predict the offset between the region proposal and the ground truth

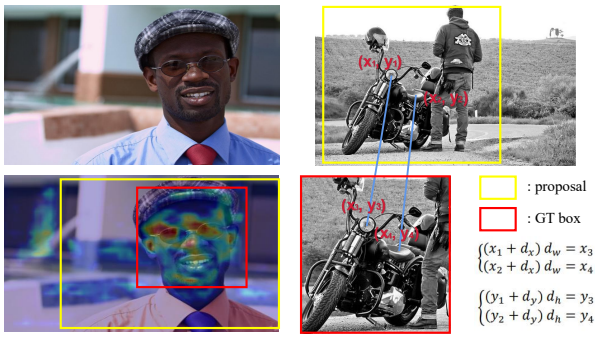
bounding box. The original R-CNN model implements this module by training the network to fit the localization result. We call this method implicit since it lacks logical reflection about the process of localization inference. The implicitly implemented localization module will suffer from over-fitting to the distribution of the base category, which is also one of the reasons why existing methods rely on fine-tuning. Therefore, we introduce class-agnostic logical inferences into the localization module, which can mitigate the over-fitting even without fine-tuning. We call it the explicit localization module. In the end, we also provide specific cases in the **Appendix** of the supplementary materials to intuitively illustrate the following proposed approaches.

Explicit Localization Score: R-CNN based model implicitly evaluates the localization score of the region proposal by the classification score from the box classifier. However, the object bounding box usually contains some low-confidence regions, such as the background and the ambiguous parts of the target object. In contrast, the classification score often only considers the high-confidence region. For example, the classification scores of the two region boxes in Figure 6 (a) are almost the same. Thus the classification score isn't correlated with the localization score, i.e., it cannot reflect the localization result. Although this irrelevancy can be alleviated by training on a large amount of annotated data, it can not be generalized to the novel category. Thus it is necessary to explore a training-independent alleviating method.

To tackle the problem, we integrate the pixel-wise contrast among feature maps into the box classification to explicitly evaluate the localization confidence of the region proposal. Specifically, the RoI module first extracts the feature maps of the region proposal and the support feature with the same shape. Then the core idea is to integrate the features comparison on each pixel of the feature map, which can compare the similarity between the instance distribution in the region proposal and the standard distribution in the support box. Obviously, the higher similarity between the instance distribution can often indicate better localization. As mentioned in Sec. III-D, IRA-FSOD adopts a switch classifier module, thus we design different integration methods for the comparison-classifier and the distance-classifier.

The integration method for the comparison-classifier is shown in the middle of Figure 5, which integrates a pixel-wise contrast network as a localization evaluator. For the distance-classifier, it can integrate the pixel-wise contrast by the distance between the flattened feature maps to evaluate the localization. Concretely, given a region proposal x and a support set of category c, it firstly calculates the cosine distance between global feature vectors, and the cosine distance between flattened feature maps. Two kinds of distances evaluate the effect of classification and localization respectively. Then their weighted sum is adopted to integrate the two distances. Finally, the probability of x belonging to category c is predicted as:

$$Pr(c;x) = \sigma\left[(1-\alpha)\, D\left(f_x, f_c\right) + \alpha D\left(v_x, v_c\right)\right] \tag{1}$$



(a) Inconsistency between high confidence region and bounding box

(b) Two pairs of coordinates can provide a bounding box offset

Fig. 6. The motivations of explicit localization inferences. The yellow box and the red box in (a) mean the bounding box and the local high confidence region, respectively. (b) shows an explicit box regression mechanism.

$$D(x,y) = \frac{x^T y}{||x|| \cdot ||y||} \tag{2}$$

$$\sigma(x) = \frac{1}{1+e^{-\lambda x}} \tag{3}$$

where v and f represent the vectors obtained from the feature map by global average pooling and flatten function. D and σ mean the cosine distance and sharp sigmoid function. The selections of α and λ are shown in Sec. IV-F. Additionally, the operation of global average pooling (vector v) can gather and mix the information from all locations in a united image, that is, it can extract global semantic information. The flattened feature maps contain abundant local contextual information, which contributes to the localization. The combination of two kinds of distance is actually similar to the multi-level image information. They are complementary to each other and refine the classification accuracy. On the whole, the original features and flattened features are jointly considered.

Semi-explicit Box Regression: General object detectors [24, 27] often implicitly fit the mapping between the features and the box regression by a network. These mappings are dependent on the base category, thus the trained regressor is hard to generalize to the novel category. To tackle the problem, we propose a semi-explicit box regressor by introducing the category-agnostic logical relation into the regression mapping. It utilizes an explicit regression mechanism: any two pairs of coordinates between the region proposal and the GT box can provide two regression equations equivalent to a correct box regression, as shown in Figure 6 (b). Despite the equivalence, this explicit regression is invalidated since the GT box is unavailable during inference. Therefore, we extract sufficient possible coordinate pairs from the comparison between the region proposal and the support box, and then predict the box regression by these coordinate pairs.

Specifically, given the feature map of a region proposal x and the average support feature map of category c (e.g., $F_x, F_c \in R^{d\times r\times r}$), we first reshape them as lists of feature vectors (e.g., $\hat{F}_x, \hat{F}_c \in R^{r^2\times d}$), then compute the distance matrix M between two lists, where

$$M_{i,j} = D(\hat{F}_x^i, \hat{F}_c^j). \tag{4}$$

Then, we flatten the distance matrix to a distance vector $d_M \in R^{r^4}$ and concatenate it with the region proposal feature. Finally, we feed the concatenated feature into a lightweight network to predict the box regression, as shown at the bottom of Figure 5.

Discussion: In d_M, each index represents a coordinate pair between two feature maps, and the corresponding value indicates the confidence score of the coordinate pair. However, these confidence scores may not be accurate due to the difference between the support instance and the GT instance. Explicitly calculating box regression by these coordinate pairs will suffer from serious errors by inaccurate scores. Thus we still predict the box regression by feeding all confidence scores into a neural network to implicitly synthesize all the equations. This regression method is between the implicit regression in general object detectors and the explicit regression by the regression equation, so we call it semi-explicit.

F. Training Details

Inspired by A-RPN [7], we train our model by the 2-way 10-shot contrastive training strategy. For each training image as a query image, we first randomly select a positive category c_1 that appears in the image and a negative category c_2 that doesn't appear in the image ($c_1, c_2 \in \mathcal{C}_b$) and then collect their support sets ($\mathcal{S}_{c_1}$ and $\mathcal{S}_{c_2}$) from $\mathcal{D}_b$, both containing ten object instances. After the forward propogation process described in III-B, we train the box comparison-classifier by the positive loss that matches the same category and the negative loss that distinguishes the different categories. For the box regressor, we only calculate the box regression loss of the region proposal belonging to c_1. The training of the semi-supervised RPN is the same as Faster R-CNN [27] except for the pseudo-label described above. The final loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{rpn} + \mathcal{L}_{cls} + \mathcal{L}_{reg} \tag{5}$$

where $\mathcal{L}_{rpn}$ consists of the classification loss and regression loss of proposals, $\mathcal{L}_{cls}$ is the binary cross-entropy loss for box classification, $\mathcal{L}_{reg}$ is the smoothed L_1 loss for box regression.

IV. Experiments

A. Experimental Setup

Dataset: In this paper, we conduct experiments on two large and challenging few-shot detection benchmark datasets, MS COCO [19] and FSOD dataset [7], which contains 800K objects belonging to 80 categories and 182K objects belonging to 1000 categories respectively. For MS COCO, we set the 20 categories belonging to PASCAL VOC [6] as the novel categories and the remaining 60 categories as the base categories, following the existing works [7, 16, 29, 38]. We use the *train2017* with only annotations of base categories for training and evaluate the detection result of novel categories on the *val2017*. FSOD dataset is specially designed for few-shot object detection, whose training set and test set only contain disjoint 800 base categories and 200 novel categories.

Implement Details: We use the commonly used ResNet-50 [14] as our backbone. Following the existing works [7, 29, 38], the backbone is pre-trained on ImageNet [5]. The most of network architectures and the hyper-parameters remain the same as Faster R-CNN [27] except for the proposed box classifier and box regressor. In addition, we halve the number of sampled anchors in RPN and proposals in the RoI head used for loss calculation, from (512, 256) for the positive and negative anchors to (128, 128, 128) for the positive, the negative, and the pseudo positive anchors during training. Our model is trained by the SGD optimizer on 3 RTX 2080Ti GPUs with a batch size of 9 (3 query images per GPU) for 120,000 iterations. The learning rate is initialized as 0.003 with the weight decay of factor 0.1 at 80,000th and 110,000th iteration.

Evaluation protocol: We conduct experiments on the one-time FSOD protocol proposed in [16] and the meta-testing protocol commonly used in the few-shot learning [36]. Given the support sets of all novel categories, the one-time FSOD protocol directly evaluates the performance of detecting these novel categories on the complete test set. The meta-testing protocol requires evaluating the average performance of the detector under numerous random episodes. Each episode randomly collects a novel category subset and consists of the corresponding support set and query set.

B. One-time FSOD evaluation

MS COCO result: In Table II, we compare our IRA-FSOD with the previous state-of-the-art methods under the 10-shot setting. For a fair comparison, we also report the backbone used in the models and the time required for the tuning process. As shown in the table, IRA-FSOD achieves new state-of-the-art results in most settings. **Under the instant-response setting,** it outperforms the latest method [40] on all AP metrics and achieves similar performance against it on AR metrics. **Even compared with the fine-tuning-based methods,** the proposed IRA-FSOD can also outperform all of them except DeFRCN[52]. Despite lagging behind DeFRCN in AP, IRA-FSOD still significantly outperforms it in efficiency since IRA-FSOD saves more than an hour of fine-tuning time. Plus, IRA-FSOD also outperforms DeFRCN in terms of AR metrics. It's worth emphasizing that IRA-FSOD achieves both high precision and recall, which is rare in the existing methods. For example, A-RPN [7] and AirDet [40] are competitive to our model on the recall, but their precisions (AP) are less than 66% of ours; DCNet [15] is competitive to our model on the precision, but its recall (AR_{10}) is only 80% of ours. **To sum up,** IRA-FSOD achieves overall leadership in response time, precision, and recall.

Then, we conduct further comparison experiments under different shot settings ($K \in \{1, 2, 3, 5, 10\}$). For a fair comparison, we evaluate all the methods over ten random runs. In each run, all the methods adopt the same support set. The support sets are generated from TFA [38]. As shown in Table III, IRA-FSOD is superior to most of the fine-tuning-based models and outperforms the no-fine-tuning models [7, 33, 40] under most settings, especially under the high-shot settings. Although [7, 33, 40] can also outperform other methods under the low-shot settings, they become significantly behind the method with fine-tuning as the shot number increases. On

TABLE II
FEW-SHOT DETECTION RESULTS FOR 20 NOVEL CLASSES ON COCO DATASET. "IR" MEANS THE MODEL IS INSTANT-RESPONSE, I.E., WITHOUT TUNING PROCESS. RED/BLUE INDICATE THE SOTA/SECOND BEST. $^+$ MEANS THE RESULT IS ESTIMATED BY THE DESCRIPTION IN THEIR PAPER."FT" MEANS FINE-TUNING. NOTE THAT OUR METHOD IS ONLY COMPARED WITH OTHER INSTANT-RESPONSE BASELINES, FOLLOWING THE ARRANGEMENT OF [54].

Model	Backbone	Average Precision						Average Recall						Tuning time
		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_1	AR_{10}	AR_{100}	AR_S	AR_M	AR_L	
MI-FSOD (SS) [42]	Res-101	4.8	7.7	5.0	1.1	3.2	7.2	15.7	29.1	32.5	5.1	21.8	41.7	25^+ min
MI-FSOD (MS) [42]	Res-101	6.7	12.0	6.6	1.7	3.8	10.2	17.7	32.6	34.3	8.1	25.3	42.4	25^+ min
Meta R-CNN [51]	Res-50	8.7	19.1	6.6	2.3	7.7	14.0	12.6	17.8	17.9	7.8	15.6	27.2	5^+ h
MPSR [44]	FPN-101	9.8	17.9	9.7	3.3	9.2	16.1	15.7	21.2	21.2	4.6	19.6	34.3	40^+ min
TFA [38]	FPN-101	9.8	19.7	8.9	2.8	9.2	16.1	14.5	18.6	18.6	5.3	14.8	33.1	16 h
FSCE [29]	FPN-101	11.9	22.3	11.6	2.9	11.1	17.6	17.0	26.6	26.5	6.7	26.3	42.3	3 h
A-RPN+FT[7]	Res-50	12.0	22.4	11.8	2.9	12.2	20.7	18.8	26.4	26.4	3.6	23.6	45.6	15 min
FSIW [47]	Res-50	12.5	27.3	9.8	2.5	13.8	19.9	20.0	25.5	25.7	7.5	27.6	38.9	5 h
DCNet [15]	Res-101	12.8	23.4	11.2	4.3	13.8	21.0	18.1	26.7	25.6	7.9	24.5	36.7	40^+ min
AirDet+FT [40]	Res-101	13.0	23.9	12.4	4.5	15.2	22.8	20.5	33.7	34.4	9.6	36.4	55.0	-
DeFRCN [52]	Res-101	16.8	30.8	15.6	6.4	17.1	25.7	19.2	29.1	-	11.1	29.6	43.9	1.2 h
A-RPN [7]	Res-50	7.3	13.2	7.1	**4.4**	8.7	10.7	17.5	32.3	33.2	10.0	34.7	50.4	**IR**
AirDet [40]	Res-101	**8.7**	**15.3**	**8.8**	4.3	**9.7**	**14.8**	**19.1**	**33.8**	**34.8**	**13.0**	**37.4**	**52.9**	**IR**
IRA-FSOD	Res-50	**13.1**	**24.5**	**12.3**	**5.9**	**16.2**	**22.0**	**19.1**	**33.5**	**35.6**	**11.4**	**39.4**	**54.9**	**IR**

TABLE III
FEW-SHOT DETECTION RESULTS ON COCO DATASET UNDER DIFFERENT SHOT SETTINGS. * MEANS THE RESULT IS RE-IMPLEMENTED. RED/BLUE INDICATE THE SOTA/SECOND BEST OF WHETHER TO USE FINETUNING. THE RESULTS ARE AVERAGED OVER TEN RANDOM RUNS. NOTE THAT RESULTS OF FS-DETR [54] ARE JUST LISTED IN *italic* TYPE AND THEY AREN'T USED TO COMPARE WITH OURS SINCE THE BACKBONE IS DIFFERENT.

Model	Backbone	1-shot			2-shot			3-shot			5-shot			10-shot			Fine-Tune
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	
A-RPN+FT*[7]	Res-50	-	-	-	4.8	9.2	3.9	5.9	11.6	5.7	8.1	15.5	7.4	10.9	20.5	9.4	✓
TFA [38]	FPN-101	1.9	3.8	1.7	3.9	7.8	3.6	5.1	9.9	4.8	7.0	13.3	6.5	9.1	17.1	8.8	✓
FSIW [47]	Res-50	3.2	8.9	1.4	4.9	13.3	2.3	6.7	18.6	2.9	8.1	20.1	4.4	10.7	25.6	6.5	✓
FSCE* [29]	FPN-101	2.0	4.9	1.3	4.2	9.5	3.4	5.7	12.0	4.7	7.6	15.6	6.5	11.2	22.3	9.8	✓
AirDet+FT [40]	Res-101	6.1	11.4	6.0	8.7	16.2	8.4	9.9	19.4	9.1	10.8	20.8	10.3	13.0	23.9	12.4	✓
QA-FewDet[33]	Res-101	4.9	10.3	4.4	7.6	16.1	6.2	8.4	18.0	7.3	9.7	20.3	8.6	11.6	23.9	9.8	✓
DeFRCN [52]	Res-101	4.8	-	-	8.5	-	-	10.7	-	-	13.6	-	-	16.8	-	-	✓
FS-DETR[54]	*DETR+Res50*	*7.0*	*13.6*	*7.5*	*8.9*	*17.5*	*9.0*	*9.8*	*18.5*	*9.8*	*10.7*	*20.5*	*10.8*	*11.1*	*21.6*	*11.0*	✗
A-RPN* [7]	Res-50	3.6	7.2	3.2	5.1	9.7	4.7	5.6	10.7	5.2	6.3	11.9	5.9	6.7	12.5	5.8	✗
AirDet [40]	Res-101	**5.9**	**10.5**	**5.9**	6.6	12.0	**6.3**	7.0	12.9	6.7	7.7	14.3	7.3	8.7	15.3	8.8	✗
AirDet [40]	Res-50	4.6	9.6	4.0	5.6	10.8	5.2	6.4	12.9	5.8	7.4	13.8	7.1	-	-	-	✗
QA-FewDet[33]	Res-101	**5.1**	**10.5**	**4.5**	**7.8**	**16.4**	**6.6**	**8.6**	**17.7**	**7.5**	**9.5**	**19.3**	**8.5**	**10.2**	**20.4**	**9.0**	✗
IRA-FSOD	Res-50	**5.1**	**10.8**	4.3	**7.7**	**15.7**	**6.6**	**8.9**	**17.6**	**7.9**	**10.5**	**20.8**	**9.1**	**12.0**	**23.5**	**10.8**	✗

the contrary, IRA-FSOD can always stay ahead (except for DeFRCN [52]), demonstrating its generalized effectiveness under various few-shot settings.

FSOD dataset result: Similar to MS COCO, we evaluate all the methods over ten random runs, and all the methods adopt the same support set in each run. The support sets are generated from the code of TFA [38]. The average results under the 1/3/5 shot setting are shown in Table IV. As shown in the table, IRA-FSOD achieves state-of-the-art results on 1/3 shots and comparable results on 5-shots with instant-response, demonstrating the strong generalization to the various novel categories. It should be noted that the FSOD dataset has 200 novel classes, i.e., the support set in the 5-shot setting has 1000 object instances. It's usually hard to obtain so many instances in the practice scenario. Therefore, the detection performance under the extremely-low-shot setting is more important, such as 1-shot and 3-shot circumstances. In addition, IRA-FSOD also performs both high precision and high recall on the FSOD dataset, the same as the results on MS COCO, indicating that it is not accidental on a particular dataset and is equipped with persuasive generalization towards different domains.

TABLE IV
FEW-SHOT DETECTION RESULTS FOR 200 NOVEL CLASSES ON FSOD DATASET. "TIME" MEANS THE TUNING TIME. "IR" MEANS THE MODEL IS INSTANT-RESPONSE, I.E., WITHOUT TUNING PROCESS. RED/BLUE INDICATE THE SOTA/SECOND BEST. ALL RESULTS ARE RE-IMPLEMENTED AND AVERAGED OVER TEN RANDOM RUNS.

Shot	Method	Backbone	AP	AP_{50}	AP_{75}	AR_{10}	Time
1	A-RPN [7]	Res-50	**9.79**	**16.00**	**10.25**	**40.33**	**IR**
	TFA [38]	FPN-101	7.43	12.07	7.79	14.42	2 h
	FSCE [29]	FPN-101	7.21	12.54	6.92	15.20	1.5 h
	IRA-FSOD	Res-50	**10.66**	**18.41**	**10.65**	**38.47**	**IR**
3	A-RPN [7]	Res-50	14.94	23.68	**15.91**	**50.01**	**IR**
	TFA [38]	FPN-101	13.21	21.10	14.16	24.87	3 h
	FSCE [29]	FPN-101	**14.96**	**25.85**	14.73	29.21	3 h
	IRA-FSOD	Res-50	**16.33**	**27.59**	**16.62**	**48.91**	**IR**
5	A-RPN [7]	Res-50	17.28	27.13	18.49	**52.30**	**IR**
	TFA [38]	FPN-101	15.85	24.89	17.31	27.74	3.5 h
	FSCE [29]	FPN-101	**19.58**	**33.42**	**19.79**	36.10	3.5 h
	IRA-FSOD	Res-50	**19.24**	**32.08**	**19.75**	**52.69**	**IR**

TABLE V
META-TESTING EVALUATION WITH 95% CONFIDENCE INTERVAL ON THE MS-COCO DATASET UNDER 5-WAY AND 1,000 EPISODES SETTING. SPE MEANS SECONDS-PER-EPISODE.

K	Method	AP	AP_{50}	AP_{75}	SPE
5	DANA [3]	$12.60_{\pm 0.29}$	$25.90_{\pm 0.44}$	$11.30_{\pm 0.35}$	10
	A-RPN [7]	$14.27_{\pm 0.27}$	$26.61_{\pm 0.45}$	$13.58_{\pm 0.31}$	9
	IRA-FSOD	$\mathbf{16.59}_{\pm 0.28}$	$\mathbf{31.97}_{\pm 0.47}$	$\mathbf{15.12}_{\pm 0.33}$	**5**
10	A-RPN [7]	$15.12_{\pm 0.29}$	$27.74_{\pm 0.47}$	$14.61_{\pm 0.32}$	10
	IRA-FSOD	$\mathbf{17.74}_{\pm 0.29}$	$\mathbf{33.59}_{\pm 0.47}$	$\mathbf{16.56}_{\pm 0.34}$	**5**

TABLE VI
META-TESTING EVALUATION WITH 95% CONFIDENCE INTERVAL ON THE MS-COCO DATASET UNDER 10-WAY AND 1,000 EPISODES SETTING. SPE MEANS SECONDS-PER-EPISODE.

K	Method	AP	AP_{50}	AP_{75}	SPE
5	A-RPN [7]	$11.32_{\pm 0.19}$	$20.84_{\pm 0.28}$	$10.87_{\pm 0.20}$	28
	IRA-FSOD	$\mathbf{14.23}_{\pm 0.16}$	$\mathbf{27.39}_{\pm 0.26}$	$\mathbf{13.11}_{\pm 0.18}$	**12**
10	A-RPN [7]	$12.11_{\pm 0.15}$	$22.05_{\pm 0.27}$	$11.78_{\pm 0.19}$	30
	IRA-FSOD	$\mathbf{15.52}_{\pm 0.14}$	$\mathbf{29.41}_{\pm 0.26}$	$\mathbf{14.55}_{\pm 0.16}$	**13**

C. Meta-testing protocol

In this section, we perform the meta-testing protocol on the MS COCO dataset. For an N-way K-shot few-shot object detection, we collect 1,000 episodes and evaluate the average object detection performance with 95% confidence interval. Each episode consists of an N-way K-shot support set and a query set containing ten images for each category. Since the evaluation is performed on each episode independently, including the fine-tuning process and the inference process, the models with fine-tuning require unacceptable time. For example, DCNet [15] requires more than a month to perform the whole meta-testing. Therefore, we only compare our IRA-FSOD with the models that support instant-response [3, 7].

Table V and VI report the average results with the 95 % confidence interval and the detection time (seconds-per-episode) under different few-shot settings, including $K \in \{5, 10\}$ and $N \in \{5, 10\}$. As shown in the table, our IRA-FSOD achieves a significant lead in both performance and efficiency. Specifically, it outperforms A-RPN by 2.3%-3.4% AP, 5.4%-7.3% AP_{50}, and 1.5%-2.8% AP_{75}. In addition, IRA-FSOD runs only half as long as A-RPN, since A-RPN introduces many complex processes, such as generating the class-specific proposals for each category and integrating multiple relation modules. In contrast, the approaches in IRA-FSOD are simpler but more effective.

TABLE VII
ABLATION FOR KEY COMPONENTS PROPOSED IN THIS PAPER: RESULTS FROM ON THE COCO DATASET UNDER THE 10-SHOT SETTING.

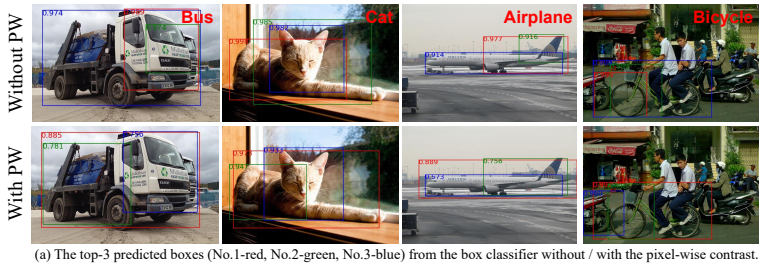
Ablation			AP	AP_{50}	AP_{75}
Faster R-CNN (baseline)	+ Multi-classifier		invalidated		
	+ Comparison-classifier		4.71	8.79	4.41
	+ Distance-classifier		5.94	15.64	2.82
+ Dyn-cls:	(Multi	+ Distance)	6.89	15.28	5.44
	(Comparison	+ Distance)	**8.69**	**17.38**	**7.78**
+ SS-RPN			10.54	20.96	9.08
+ SE-Reg			10.64	20.59	9.58
+ Cls-PW			10.67	20.92	9.47
+ SS-RPN	+ Cls-PW		11.65	22.52	10.31
+ SS-RPN	+ SE-Reg		11.82	22.92	10.60
+ Cls-PW	+ SE-Reg		11.94	22.21	11.19
+ SS-RPN	+ Cls-PW	+ SE-Reg	**13.05**	**24.50**	**12.33**

D. Quantitative Ablation Studies

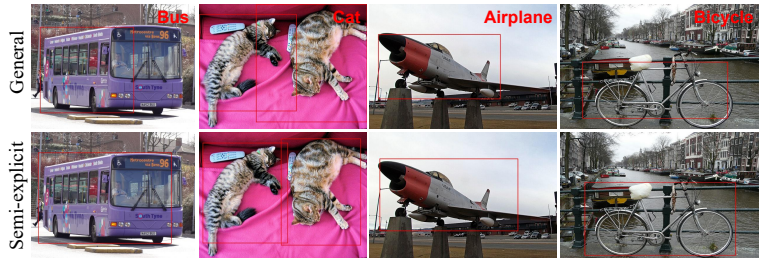
In this section, we evaluate the effects of the core components in IRA-FSOD. All ablation studies are conducted on the COCO dataset under the 10-shot setting and one-time FSOD evaluation protocol. IRA-FSOD is built on top of Faster R-CNN [12], which is designed for general object detection. Thus we adopt it as the baseline and design the ablation experiments in two stages in Table VII.

In the first stage, we evaluate the effect of the proposed switch classifier module, which is a essential strategy to transform the general object detector into a few-shot object detector with instant-response. Without the switch classifier module, the model suffers from low performance or even invalidation due to low learnability or low generalization of the classifier. Just by introducing the switch classifier module into the Faster R-CNN, it is already comparable with some methods requiring fine-tuning [39, 45, 51].

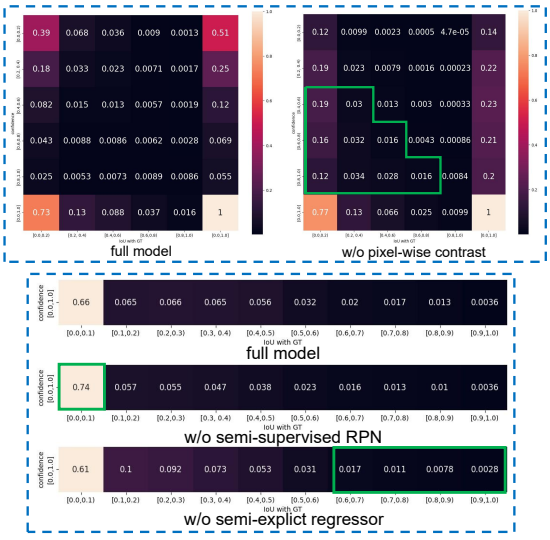
In the second stage, we evaluate the different combinations of three proposed boosted modules, including the semi-supervised RPN (SS-RPN), the box classifier with the pixel-wise contrast (Cls-PW), and the semi-explicit box regressor (SE-Reg). As shown in Table VII, their improvements for the model performance are different and are all in line with our expectations. Concretely, (a) The semi-supervised RPN mainly achieves the performance improvement on the AP_{50} metric



(a) The top-3 predicted boxes (No.1-red, No.2-green, No.3-blue) from the box classifier without / with the pixel-wise contrast.



(b) The localization result from the semi-explicit box regressor (bottom) and the general box regressor (top).



(c) The heat maps of the statistical results for all prediction boxes from different models

Fig. 7. The qualitative ablation result: (a) and (b) are some detection results from different ablated models. (c) is heat maps of the statistical results for all prediction boxes, in which the vertical axis means the confidence range of the prediction boxes and the horizontal axis means the max IoU range with the ground truth of the predicted category. We mark the noteworthy parts with green borders.

(+1.4%-3.6%), indicating that it successfully captured more potential region proposals belonging to the novel categories; (b) The semi-explicit box regressor can significantly improve the result on the AP_{75} metric (+1.6%-2.0%), which shows that it can generate more accurate high-quality boxes by improving the localization accuracy; (c) The pixel-wise contrast in the box classifier mainly improves the confidence ranking of all the predicted boxes, thus can achieve significant improvement on all evaluation metrics, e.g., AP_{50} (+1.4%-3.6%) and AP_{75} (+1.3%-1.7%). In summary, the proposed components breakthrough the limitations of fine-tuning and they gather into a powerful inference mechanism which is suitable and robust to few-shot object detection.

E. Qualitative Ablation Studies

In this section, we provide the qualitative ablation studies of the three proposed boosted modules. In Figure 7, we compare the detection result between the full model and the ablated model: the left sub-figure shows the detection examples from different ablated models; the right-top sub-figure statistics the relevance between the predicted confidence and the localization accuracy (i.e., the IOU between ground truth); the right-bottom sub-figure statistics the overall distribution of the localization accuracy. The specific analysis for the three boosted modules is as follows:

Semi-supervised RPN: As shown in Figure 7 (c), the detection result from the model without the semi-supervised RPN module contains significantly more irrelevant prediction boxes (i.e., $IoU < 0.1$), indicating the semi-supervised RPN module can capture more potential foreground proposals.

Pixel-wise contrast: As shown in Figure 7 (a), after introducing the classifier with the pixel contrast in our algorithm, the confidence is basically correlated with the localization accuracy (i.e., the IoU between ground truth). Without the pixel contrast, the confidence of the predicted box is irrelevant to localization, and the prediction with the highest confidence usually only encloses the local position of the object. Figure 7 (c) also shows that the detection result from the model without the pixel-wise contrast produces a large number of high-confidence but poorly localized detection results.

Semi-explicit regressor: As shown in Figure 7 (b), the model with the semi-explicit box regressor can significantly improve the localization accuracy and generate higher-quality predicted boxes. Figure 7 (c) also shows that the proportion of high-quality boxes (i.e., $IoU \geq 0.7$) in the detection result from the model with the semi-explicit regressor is greatly increased.

F. Hyper-parameter Studies:

In this section, we study the effect and selection of the hyper-parameters in IRA-FSOD, including the threshold τ in the semi-supervised RPN, as well as the balance weight α and the scaling factor λ in the distance-classifier. For each hyper-parameter, we first select a candidate set by observation and then evaluate their performance on the COCO dataset under the 10-shot setting and one-time FSOD evaluation protocol. The performances at different values of them in Table VIII.

The specific analysis is as follows:

TABLE VIII
HYPER-PARAMETER STUDIES OF τ AND α IN THE IRA-FSOD: RESULTS FROM ON THE COCO DATASET UNDER THE 10-SHOT SETTING.

τ	AP	AP_{50}	AP_{75}	α	AP	AP_{50}	AP_{75}
-	11.94	22.21	11.19	0.0	11.82	22.92	10.60
0.75	12.38	22.98	11.73	1/3	12.96	24.39	12.16
0.50	12.66	23.97	11.79	1/2	**13.05**	**24.50**	**12.33**
0.25	**13.05**	**24.50**	**12.33**	2/3	12.92	24.20	12.18
0.10	12.87	24.05	12.03	1.0	12.01	22.24	11.49

Threshold τ: Traditional Semi-supervised learning algorithms [2, 49, 50] usually requires high thresholds to reduce the incorrect pseudo labels. However, in the two-stage detector,

TABLE IX
THE STATISTICS RESULTS OF THE REGION PROPOSALS. "FP" AND "TP" MEAN FALSE POSITIVE AND TRUE POSITIVE RESPECTIVELY. WE DEFINE BACKGROUND PROPOSALS AND FOREGROUND PROPOSALS AS ONES WHOSE IOU WITH GROUND TRUTH BOX IS LESS THAN 0.1 AND LARGER THAN 0.5, RESPECTIVELY.

threshold	FP	filtered FP	saved TP
w/o	133421	21173789	423011
0.75	133470	21179057	428241
0.50	133754	21180742	428696
0.25	133888	21181604	433442
0.10	141201	21637617	417326

we expect RPN to capture all the potential objects as much as possible and then eliminate the incorrect proposals by the box classifier. Therefore, the model performs better when τ is lower and reaches the optimal performance at $\tau = 0.25$.

Moreover, we also analyze the positive and negative impacts of different thresholds. As shown in Table IX, when the threshold is larger than 0.1, the actual number of incorrect proposals (false positives) only marginally increases as the threshold declines. Despite the marginal increment, thanks to the contribution of the box classifier, more FP are filtered and more TP are kept, which indicates that the negative impact of increasing FP can be suppressed and the positive contribution of a relatively low threshold can be kept as well. It should be explained that each region proposal will be calculated once for each category in the box classifier, so the numbers of filtered FP and saved TP are much larger than the one of FP. Only when the threshold decreases to 0.1 does the number of FP begin to rise uncontrollably.

Balance weight α: Compared with $\alpha = 0.0$ and $\alpha = 1.0$, the performances with other values are improved significantly, indicating that the evaluation of localization and classification are both valuable. It reaches the optimal integration performance at $\alpha = \frac{1}{2}$.

Scaling factor λ: In the IRA-FSOD, the performance is not affected by the scaling factor since it doesn't affect the confidence ranking of the predicted box. Thus we empirically choose $\lambda = 20$ to adjust the sharpness of the prediction distribution.

V. CONCLUSION

In the field of few-shot object detection, the existing methods tend to transfer their model to the task by employing a fine-tuning process, resulting in many application drawbacks. To tackle the problem, this paper studies in-depth how to get rid of fine-tuning while maintaining the FSOD performance. Through careful study of each module in a general object detector (i.e., Faster R-CNN), this paper builds an instant-response and accurate few-shot object detector (IRA-FSOD) that can accurately detect the object of novel categories without fine-tuning. To more solidly validate the proposed analysis, we deliberately avoid introducing excessive extra-complexity when designing the improved components. Despite its simplicity, IRA-FSOD can reach state-of-the-art performance in both efficiency, precision and recall. It is noteworthy that our works are built on only Faster R-CNN without other prior methods, so all the approaches are easily compatible with the existing FSOD methods. We hope our studies can inspire future works to explore more powerful few-shot object detectors.

VI. ACKNOWLEDGE

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61876224, 61836012, and GuangDong Province Key Laboratory of Information Security Technolog.

REFERENCES

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).

[2] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin and Colin Raffel. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in neural information processing systems* (2020).

[3] Tung-I Chen, Yueh-Cheng Liu, Hung-Ting Su, Yu-Cheng Chang, Yu-Hsiang Lin, Jia-Fong Yeh, Wen-Chin Chen, and Winston Hsu. 2021. Dual-Awareness Attention for Few-Shot Object Detection. *IEEE Transactions on Multimedia* (2021). https://doi.org/10.1109/TMM.2021.3125195

[4] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. 2019. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232* (2019).

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

[6] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. In *International Journal of Computer Vision*. 303–338.

[7] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. 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.

[8] Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. 2021. Generalized few-shot object detection without forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4527–4536.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*. PMLR, 1126–1135.

[10] Victor Garcia and Joan Bruna. 2018. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*.

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

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

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

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

[15] Hanzhe Hu, Shuai Bai, Aoxue Li, Jinshi Cui, and Liwei Wang. 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.

[16] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. 2019. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8420–8429.

[17] Leonid Karlinsky, Joseph Shtok, Sivan Harary, Eli Schwartz, Amit Aides, Rogerio Feris, Raja Giryes, and Alex M Bronstein. 2019. Repmet: Representative-based metric learning for classification and few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5197–5206.

[18] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV* (eccv ed.). European Conference on Computer Vision. https://www.microsoft.com/en-us/research/publication/microsoft-coco-common-objects-in-context/

[20] Liyang Liu, Bochao Wang, Zhanghui Kuang, Jing-Hao Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. 2021. Gendet: Meta learning to generate detectors from few shots. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[22] Alex Nichol and John Schulman. 2018. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999* 2, 3 (2018), 4.

[23] Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning. (2016).

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

[25] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7263–7271.

[26] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

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

[28] Jake Snell, Kevin Swersky, and Zemel Richard. 2017. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*. 4077–4087.

[29] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. 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.

[30] Zhang Lei, Zuo Liyun, Du Yingjun, Zhen Xiantong. Learning to Adapt With Memory for Probabilistic Few-Shot Learning. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT 2021)*.

[31] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Schiele Bernt. 2019. Optimization as a model for few-shot learning. (2019), 403–412.

[32] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation by joint identification-verification. *arXiv preprint arXiv:1406.4773* (2014).

[33] Yi Sun, Guangxing Han, Yicheng He, Shiyuan Huang, Jiawei Ma, Shih-Fu Chang. Query adaptive few-shot object detection with heterogeneous graph convolutional networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021).

[34] Chi Ziqiu, Wang Zhe, Yang Mengping, Li Dongdong, Du Wenli. Learning to Capture the Query Distribution for Few-Shot Learning. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT 2022)*.

[35] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1199–1208.

[36] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*. 3630–3638.

[37] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5265–5274.

[38] Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. 2020. Frustratingly Simple Few-Shot Object Detection. In *International Conference on Machine Learning*. PMLR, 9919–9928.

[39] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. 2019. Meta-learning to detect rare objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9925–9934.

[40] Bowen Li, Chen Wang, Pranay Reddy, Seungchan Kim, Sebastian Scherer. Airdet: Few-shot detection without fine-tuning for autonomous exploration. *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*.

[41] Zeyuan Wang, Yifan Zhao, Jia Li, and Yonghong Tian. 2020. Cooperative Bi-Path Metric for Few-Shot Learning. In *Proceedings of the 28th ACM International Conference on Multimedia*. New York, NY, USA, 1524–1532.

[42] Cheng Meng, Wang Hanli, Yu Long. Meta-Learning-Based Incremental Few-Shot Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT 2022)*.

[43] Zhu Pengkai, Wang Hanxiao, Saligrama Venkatesh. Zero Shot Detection. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT 2022)*.

[44] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. 2020. Multi-Scale Positive Sample Refinement for Few-Shot Object Detection. In *European Conference on Computer Vision*. Springer, 456–472.

[45] Xiongwei Wu, Doyen Sahoo, and Steven Hoi. 2020. Meta-RCNN: Meta Learning for Few-Shot Object Detection. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1679–1687.

[46] Zhang Jing, Zhang Xinzhou, Wang Zhe. Task Encoding With Distribution Calibration for Few-Shot Learning. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT 2022)*.

[47] Yang Xiao and Renaud Marlet. 2020. Few-Shot Object Detection and Viewpoint Estimation for Objects in the Wild. In *European Conference on Computer Vision (ECCV)*.

[48] Jiliang Yan, Deming Zhai, Junjun Jiang, and Xianming Liu. 2021. Target-Guided Adaptive Base Class Reweighting for Few-Shot Learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 5335–5343. https://doi.org/10.1145/3474085.3475656

[49] Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A. & Raffel, C. MixMatch: A Holistic Approach to Semi-Supervised Learning. *ArXiv Preprint ArXiv:1905.02249*. (2019)

[50] Xie, Q., Dai, Z., Hovy, E., Luong, M. & Le, Q. Unsupervised Data Augmentation for Consistency Training. *ArXiv Preprint ArXiv:1904.12848*. (2019)

[51] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. 2019. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9577–9586.

[52] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu*, Jianan Wu, and Chi Zhang. 2021. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8681–8690.

[53] Xiang Li, Lin Zhang, Yau Pun Chen, Yu-Wing Tai, Chi-Keung Tang. One-Shot Object Detection without Fine-Tuning. *arXiv preprint arXiv:2005.03819, 2020*.

[54] Adrian Bulat, Ricardo Guerrero, Brais Martinez, Georgios Tzimiropoulos. 2022. FS-DETR: Few-Shot DEtection TRansformer with prompting and without re-training. In *arXiv preprint arXiv:2210.04845*.

[55] Jinhai Yang, Hua Yang, and Lin Chen. 2021. Towards Cross-Granularity Few-Shot Learning: Coarse-to-Fine Pseudo-Labeling with Visual-Semantic Meta-Embedding. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3005–3014.

Junying Huang received the B.Eng degree from the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China, in 2020. He is currently a postgraduate student in the Human Cyber Physical Intelligence Integration Lab, School of Computer Science and Engineering, Sun Yat-sen University. His research interests include computer vision and machine learning, with a focus on few-shot learning.

Junhao Cao received the B.Eng degree from the School of Aeronautics and Astronautics, Sun Yat-sen University, Guangzhou, China, in 2022. He is currently a postgraduate student in the Human Cyber Physical Intelligence Integration Lab, School of Computer Science and Engineering, Sun Yat-sen University. His research interests include semi-supervised learning.

Liang Lin is a full professor of Computer Science in Sun Yat-sen University and CEO of DarkerMatter AI. He worked as the Executive Director of the SenseTime Group from 2016 to 2018, leading the R&D teams in developing cutting-edge, deliverable solutions in computer vision, data analysis and mining, and intelligent robotic systems. He has authored or co-authored more than 200 papers in leading academic journals and conferences. He is an associate editor of IEEE Trans. Human-Machine Systems and IET Computer Vision, and he served as the area/session chair for numerous conferences such as CVPR, ICME, ICCV. He was the recipient of Annual Best Paper Award by Pattern Recognition (Elsevier) in 2018, Dimond Award for best paper in IEEE ICME in 2017, ACM NPAR Best Paper Runners-Up Award in 2010, Google Faculty Award in 2012, award for the best student paper in IEEE ICME in 2014, and Hong Kong Scholars Award in 2014. He is a Fellow of IET.

Dongyu Zhang received the Ph. D. from Harbin Institute of Technology in 2010. He is an associate professor of School of Computer and Engineering of Sun Yat-sen University. His research interests include computer vision and machine learning.