CRTED: Few-Shot Object Detection via Correlation-RPN and Transformer Encoder-Decoder

Jinlong Chen , Kejian Xu * , Yi Ning , Lianyuan Jiang , Zhi Xu *

Posted Date: 9 April 2024

doi: 10.20944/preprints202404.0633.v1

Keywords: Few-shot object detection; Region proposal network; Transformer Encoder-Decoder; Training strategies

Article

CRTED: Few-Shot Object Detection via Correlation-RPN and Transformer Encoder-Decoder

Jinlong Chen [1], Kejian Xu [1,*], Yi Ning [2], Lianyuan Jiang [1] and Zhi Xu [1,*]

[1] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541000, Guangxi, China; 7259@guet.edu.cn; xkj897899031@gmail.com; jiangly198@gmail.com; xuzhi@guet.edu.cn

[2] School of Continuing Education, Guilin University of Electronic Technology, Guilin 541000, Guangxi, China; 13977316878@163.com

* Correspondence: xkj897899031@gmail.com; xuzhi@guet.edu.cn

Abstract: Few-shot object detection (FSOD) aims to address the challenge of requiring a substantial amount of annotations for training in conventional object detection, which is very labor-intensive. However, existing few-shot methods achieve high precision with the sacrifice of time-consuming for exhaustive fine-tuning, or take poor performance in novel-class adaptation. We presume the major reason is that the valuable correlation feature among different categories is insufficiently exploited, hindering the generalization of knowledge from base to novel categories for object detection. In this paper, we propose Few-Shot object detection via Correlation-RPN and Transformer Encoder-Decoder (CRTED), a novel training network to learn object-relevant features of inter-class correlation and intra-class compactness while suppressing object-agnostic features in the background with limited annotated samples. And we also introduce a 4-way tuple-contrast training strategy to positively activate the training progress of our object detector. Experiments over two few-shot benchmarks (Pascal VOC, MS-COCO) demonstrate that, our proposed CRTED without further fine-tuning can achieve comparable performance with current state-of-the-art fine-tuned works. The codes and pre-trained models will be released.

Keywords: few-shot object detection; region proposal network; transformer encoder-decoder; training strategies

1. Introduction

In recent years, [1–5] has seen remarkable advancements through the application of deep neural models and large-scale training. Nevertheless, conventional object detection techniques typically depend extensively on vast amounts of quantity and quality annotated data and necessitate extended training duration, which has sparked the recent pursuit of few-shot object detection (FSOD). The challenge in few-shot learning lies in the significant diversity of real-world objects and despite noteworthy advancements, existing methods [6–10] primarily focus on image classification, seldom delving into the complexities of few-shot object detection. This might be due to the non-trivial nature of transferring knowledge from few-shot classification to few-shot object detection.

The core difficulty in object detection with limited examples lies in pinpointing unseen objects against a cluttered background. This is essentially a general problem of locating objects from a few annotated examples in new categories. Potential bounding boxes often overlook unseen objects or generate numerous false detections in the background. We argue that it is due to the sub-optimal scoring of promising bounding boxes by a region proposal network (RPN), making it challenging to detect novel objects. This distinction underscores the inherent difference between few-shot classification and object detection. Additionally, recent efforts in few-shot object detection [11–13] necessitate fine-tuning, preventing their direct application to novel categories.

In this paper, we attempt to address the problem mentioned above in few-shot object detection. First, we propose a novel network structure named Correlation-RPN based on general RPN to

 doi:10.20944/preprints202404.0633.v1

activate model pay more attention to object-relevant regions and help learn the matching correlation between query and support image feature, for generalizing the knowledge learned from base classes to novel classes. Secondly, we integrally migrate the transformer encoder-decoder into our framework. With the new feature coding mechanism, we utilize decoder to get correlational metric of feature representation after feature extraction in the backbone of our network. Thus, we introduce a 4-way tuple-contrast training strategy to positively activate the training progress of our object detector.

The main contributions of this work include:

- We propose a novel correlation-aware region proposal network structure called Correlation-RPN, and migrate it to object detectors, improving detectors' capacity of object localization and generalization;
- We redesign a new feature coding mechanism and integrally migrate the encoder-decoder of transformer into our model to effectively learn support-query feature similarity representation;
- With our presented 4-way tuple-contrast training strategy, CRTED without further fine-tuning can achieve comparable performance with most of the representative methods in few-shot object detection.

2. Related Work

2.1. General Object Detection

Object detection remains a key topic in computer vision, particularly with the rise of deep learning. CNN-based methods, always pre-trained on vast datasets, have gained popularity. These methods split into two categories: proposal-based and proposal-free detectors. The RCNN series [14–16] falls into the former, relying on pre-trained CNNs to classify region proposals from selective search. SPP-Net [17] and Fast-RCNN [15] evolved from RCNN, extracting regional features via an RoI pooling layer from convolutional maps. Faster-RCNN [16] introduced a region proposal network (RPN) to enhance proposal quality. In contrast, YOLO [3,18–20] pioneered the proposal-free approach, using a single CNN for classification and bounding box prediction. Later works refined YOLO with default anchors for shape adjustment or multi-scale training. Proposal-free methods are simpler and faster but still rely heavily on annotated samples, limiting their performance in few-shot scenarios.

2.2. Few-Shot Object Detection

The challenging few-shot object detection (FSOD) problem aims to detect objects or novel classes at instance-level with limited annotations. Prior works on few-shot object detection can be mainly categorized into three paradigms: meta-learning, transfer-learning and metric-learning approaches. Meta-learning methods aim at devising a periodic and stage-wise meta-training paradigm to train a class-agnostic meta-model to help knowledge transfer from base classes to novel classes with few annotated labels, known as "Learning to learn". The Meta-RPN [21] and Meta faster-cnn [13] are proposed to generate class-relevant proposals with improving the instance alignment. Transfer-learning based methods, also known as the finetuning based methods, emphasize performing fine-tuning for few-shot object detection, which trained base classes and novel classes together, and fine-tuning the whole model, where the model is only trained on the base classes and then fine-tuned on a balanced set including both base classes and novel classes. MPSR [22] adopt manually defined positive sample refinement branch to mitigate the object scale scarcity issue for few-shot object detection. Specially, TFA [23] pre-trains a base detector from abundant samples on a base set and fine-tuning it for novel classes. Metric-learning approaches focus on learning good embedding spaces or appropriate metrics that facilitate downstream tasks, including cosine similarity [24], euclidean distance to class center, graph distances and so on. RepMet [25] achieves current SOTA results on few-shot object detection by simultaneously learning the parameters of backbone network, the embedding space, and the multimodal distribution of each training category within it in an end-to-

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0633.v1

end training manner. [21] exploits the similarity metric between the support set and query set in few-shot setup to detect novel objects and suppress false detection in the background.

2.3. Transformer Encoder-Decoder

The Transformer architecture, initially designed for machine translation, has been widely applied to various computer vision tasks. One notable example is DETRs, a representative class of object detectors that leverage the strengths of Transformers. These models employ a transformer encoder-decoder structure to understand and learn the relationships between the global image context and objects, using CNN features as input, and producing final predictions. As a variant of DETR, ViDT [26] introduced a pre-trained transformer to replace the CNN backbone, while maintaining a randomly initialized transformer neck. More recently, ViTDet [27] and MIMDet [28] have attempted to capitalize on the powerful network architecture pre-trained by MAE for object detection tasks. However, ViTDet only utilizes the pre-trained MAE encoder and discards the pre-trained decoder. In contrast, MIMDet retains the entire encoder-decoder for feature extraction, focusing on leveraging the reconstruction capabilities of the MAE decoder to mask input image patches, reducing additional inference costs. Unlike these approaches, imTED [29] employs a fully pre-trained transformer encoder-decoder that not only extracts features but also performs representation transformation, which offers a comprehensive utilization of the transformer's capabilities, enhancing the performance of object detection tasks.

3. Approach

In this section, we will walk through the whole architecture designs in our proposed CRTED step by step. The structure of CRTED is exhibited in Figure 1. First, before introducing CRTED, we consider few-shot object detection task it aims to achieve. We start with the preliminaries on the few-shot object detection setup that motivate our method. Then, we present our network architecture in detail for few-shot object detection. Finally, we describe the learning procedure of a CRTED.

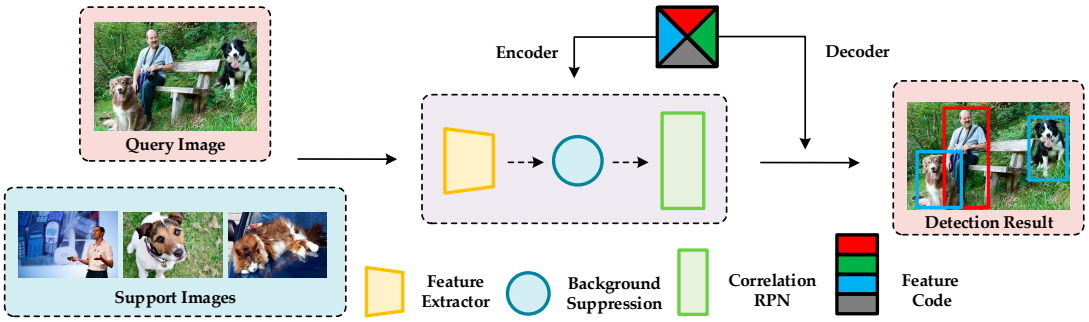


Figure 1. The whole framework of our proposed CRTED. Query/support feature from the weight-shared feature extractor will be conducted BS regularization and encoding, and then fed into Correlation-RPN for further extracted. Finally, the few-shot detection result is obtained by mapped to the query image with the decoder.

3.1. Preliminaries

Problem Definition In this work, we focus on processing the task of few-shot object detection (FSOD). Given two sets of classes, a base set C_{base} and a novel set C_{novel}, where $C_{base} \cap C_{novel} = \varnothing$. Defining two datasets, a base dataset D_{base} with sufficient annotated objects of C_{base} and a novel dataset D_{novel} with few annotated objects of C_{novel}. A few-shot object detector aims at classifying and localizing objects of $C_{base} \cup C_{novel}$ by learning from $D_{base} \cup D_{novel}$. In a task of N_n-way K-shot object detection with $N_n = |C_{novel}|$, there are exactly K annotated instances for each novel class in D_{novel}. The goal of this work is to train a model that can detect novel classes in C_{novel} by only providing K-shot labeled samples for C_{novel} and abundant images from C_{base}. Basically, images from C_{base} are split into support image set S_b containing support images s_c with a close-up of the target object, and query image set Q_b containing

 doi:10.20944/preprints202404.0633.v1

query images q_c which potentially contains objects belonging to the support class. Given all support images S_b, our model learns to detect objects in Q_b. For convenience, we denote C_{base}, C_{novel}, D_{base} and D_{novel} as C_b, C_n, D_b and D_n in the following sections.

Rethink Region-Based Object Detectors The majority of current few-shot object detection methods rely heavily on the Faster R-CNN framework [16], which leverages a region proposal network (RPN) to generate potentially relevant bounding boxes to facilitate subsequent detection tasks. The RPN plays a pivotal role, as it must not only distinguish between objects and the background but also filter out negative objects belonging to non-matching support categories. However, under the few-shot detection setting, where support image information is extremely limited, the RPN often struggles. It tends to indiscriminately focus on every potential object with a high objectness score, regardless of whether they belong to the support category or not. This behavior can hamper the generalization of knowledge from base classes to novel classes, and it also places a significant burden on the subsequent classification task of the detector, as it has to deal with a large number of irrelevant objects. Previous studies [1,3,4,16,21,30] have attempted to address this challenge by generating more accurate region proposals. Nevertheless, the issue persists, stemming from the inherent limitations of region-based object detection frameworks within the few-shot learning context. To truly address this challenge, it is essential to develop novel strategies that can effectively leverage the limited support image information, enhancing the discriminative capabilities of the RPN and ensuring that it focuses only on relevant objects, thus improving the overall performance of few-shot object detection systems.

Rethink Transformer-Based Detection Frameworks Transformer [31] emerged as a revolutionary self-attention-based building block specifically tailored for machine translation tasks. This architecture revolutionizes the way sequences are processed, updating each element by scanning through the entire sequence and subsequently aggregating information from it. Seeking to harness the immense potential of the Transformer, DETRs [1] introduced an innovative approach by integrating a transformer encoder-decoder architecture into an object detector. This integration enabled the system to tackle the intricate challenge of attending to multiple support classes within a single forward pass. Nevertheless, a notable issue persists: the vision transformers employed in DETRs were randomly initialized, limiting their capabilities to solely processing feature representations extracted by the backbone network. This constraint underscores the need for further advancements to fully unlock the potential of vision transformers in object detection tasks.

3.2. Architecture

Correlation-Aware Region Proposal Network In generic object detection, RPN is useful to provide region proposals and generate object-relevant anchors while suffer from performance drop when in few-shot object detection, since the low-quality region proposals for novel classes and the fatigue to capture inter-class correlation among different classes. Take inspiration of success of RPN-based FSOD framework [21], we propose a novel network structure based on general RPN which learns the matching correlation between the support set S_b and queries Q_b. Figure 2 shows the overall architecture of our proposed Correlation-RPN. The Correlation-RPN can make use of the support information to sensitively aware the similarities and difference between S_b and Q_b, which is able to provide high-quality region proposals with objects of target or novel classes, while relatively depressing proposals in background or in other categories.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0633.v1

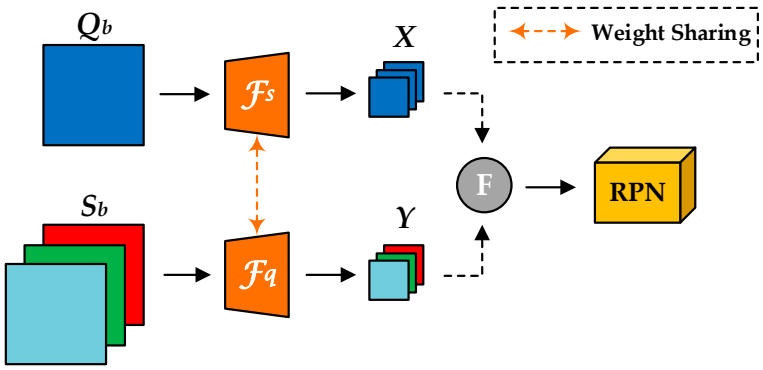


Figure 2. The overall view of Correlation-RPN design. S_b: support set; Q_b: queries; X: support feature map; Y: query feature map. $\mathcal{F}_s$ and $\mathcal{F}_q$ denote support-query feature extractor. F denotes the similarity of support-query feature map, which includes a 1 × 1 × C and a 3 × 3 × C vector. Then the correlation map computed is to be fed into RPN for generating proposals.

Specifically, we compute the correlational metric between the feature map of S_b and Q_b in a depth-wise manner. The similarity map then is utilized to build the region proposal generation. In particular, we denote the support features of S_b as $X \in R^{H\times W\times C}$ and feature map of the query q_c as $Y \in R^{H\times W\times C}$, the similarity is defined as:

$$\boldsymbol{F}_{h,w,c} = \sum_{i,j} \alpha X_{i,j,c} \cdot \beta Y_{h+i-1,w+j-1,c}, \qquad i,j \in \{1,\dots,K\} \tag{1}$$

where $\boldsymbol{F}$ is the resultant correlation feature map and α, β are control coefficients to prevent overly favoring features of either side. Here the X is used as the kernel to slide on the query feature map [32,33] in a depth-wise cross correlation way [34]. Our work adopts the top architecture of the Attention RPN [21]. We empirically find that a kernel size of $K = 1$ performs well in our case, since we argue that global feature can provide a great object prior for objectness classification, consistent with [16]. In our case, the kernel is calculated by averaging on the support feature map X. The correlation map is processed simultaneously by a 1 × 1 convolution and a 3 × 3 convolution followed by the objectiveness branch and regression branch. The Correlation-RPN is trained jointly with the network and elaborated as in the section 4.3.

Feature Metric Matching Based on the idea of [1,29], we integrally migrate the transformer encoder-decoder as the pillars of correlational metric aggregation module into our object detector. The feature metric matching is accomplished in transformer encoder by multi-head attention mechanism. Specifically, given the support features of S_b, denoted as $F_s \in R^{H\times d}$, and the query image q_c, of which feature map is denoted as $F_q \in R^{H\times W\times d}$, the matching coefficients $\boldsymbol{M}$ can be obtained by:

$$\boldsymbol{M}(HW,d,c) = Match\left(F_s,F_q\right)$$

$$= Softmax\left(\frac{(F_s(d,c)\cdot S)\left(F_q(HW,d)\cdot S\right)^T}{\sqrt{d}}\right), \qquad c \in \{1,\dots,C\} \tag{2}$$

where HW is the feature spatial size, d is the feature dimensionality, C is the number of support categories, and S is a cosine similarity shared by F_s and F_q, to ensure they are embedded into the same linear feature projection space. To calculate cosine similarity as the correlational metric of each pair of feature representation of F_s and F_q, which is calculated via:

$$cos\left(f_s,f_q\right) = \frac{1}{C}\sum_{i}^{C}\left(\frac{f_s\cdot f_q}{|f_s|\cdot|f_q|}\right), \qquad i \in \{1,\dots,C\} \tag{3}$$

where f_s and f_q denote the single feature representation of F_s and F_q. Finally, for each q_c that may contain multiple complex instances, we ensure a fact that the choice of the m potential support cases for each of these instances is same. Therefore, the average correlation score of the m same

 doi:10.20944/preprints202404.0633.v1

potential support objects can be considered as the similarity or shared feature representation between q_c and the m potential S_b, with which we prefer the s_c containing the most similar support instance as the powerful support patch of s_c in training. And this process has been experimentally demonstrated to be helpful for our training. The effectiveness of support-query feature similarity metric mining, i.e., distinguishing support objects similar to the query, is discussed in Section 4.3.

Encoding Matching In order to achieve class-agnostic object prediction, we propose the utilization of a carefully crafted set of predefined task encodings, which serve as a bridge between the given support classes and the abstract task encodings space. By mapping the support classes to these encodings, we ensure that the final object predictions are constrained within the task encodings space, rather than being limited to predicting specific classes on the surface level. Drawing inspiration from the positional encodings employed in the Transformer architecture, we implement task encodings $\boldsymbol{T} \in R^{H\times d}$ utilizing sinusoidal functions. This allows us to capture both local and global patterns within the task encodings space, enhancing the representational power of our approach.

Furthermore, encoding matching and feature metric matching share the same matching coefficients. This ensures consistency across different matching processes and simplifies the overall pipeline. The matched encodings $\boldsymbol{Q}_E$ are simply obtained through a straightforward process, further streamlining the prediction framework:

$$\boldsymbol{Q}_E = \boldsymbol{M} \otimes \boldsymbol{T}, \tag{4}$$

where $\otimes$ denote sinusoidal functional multiplication. In essence, our approach offers a more flexible and generalizable framework for object prediction, enabling us to transcend the limitations of traditional class-specific prediction methods and move towards a more abstract and powerful representation of objects.

Modeling Background for Object Prediction Generally, under a few-shot object detection setup, background does not belong to any target classes and usually takes up a lot of space in a support or query image. Those images that objects only account for a small proportion and most of the area is complex background, which we also called hard samples, as shown in Figure 3. Taking the consideration of this reason, we propose a learnable prototype BG-P and a corresponding task encoding BG-E (fixed to zeros), for explicitly modeling the background class. This can significantly eliminate the matching ambiguity when query is very hard to match any of the given support classes. And we additionally introduce a background-suppression (BS) regularization as an auxiliary branch to help addressing this problem, which will be described in detail in the next section. The final output of the feature metric matching module can be obtained via the following equation:

$$\boldsymbol{Q}_F = \tau(\boldsymbol{M} \cdot \sigma(F_s), BS(F_q)), \tag{5}$$

where $\tau(\cdot)$ denotes Hadamard product, $BS(\cdot)$ denotes background-suppression operation and $\sigma(\cdot)$ denotes sigmoid function. By applying the matching coefficients $\boldsymbol{M}$, we filter out features not matched to S_b, producing a feature map $\boldsymbol{Q}_F$ that inhibits the negative impact of hard samples and highlights class-related objects from query set Q_b for each individual support class.

 doi:10.20944/preprints202404.0633.v1

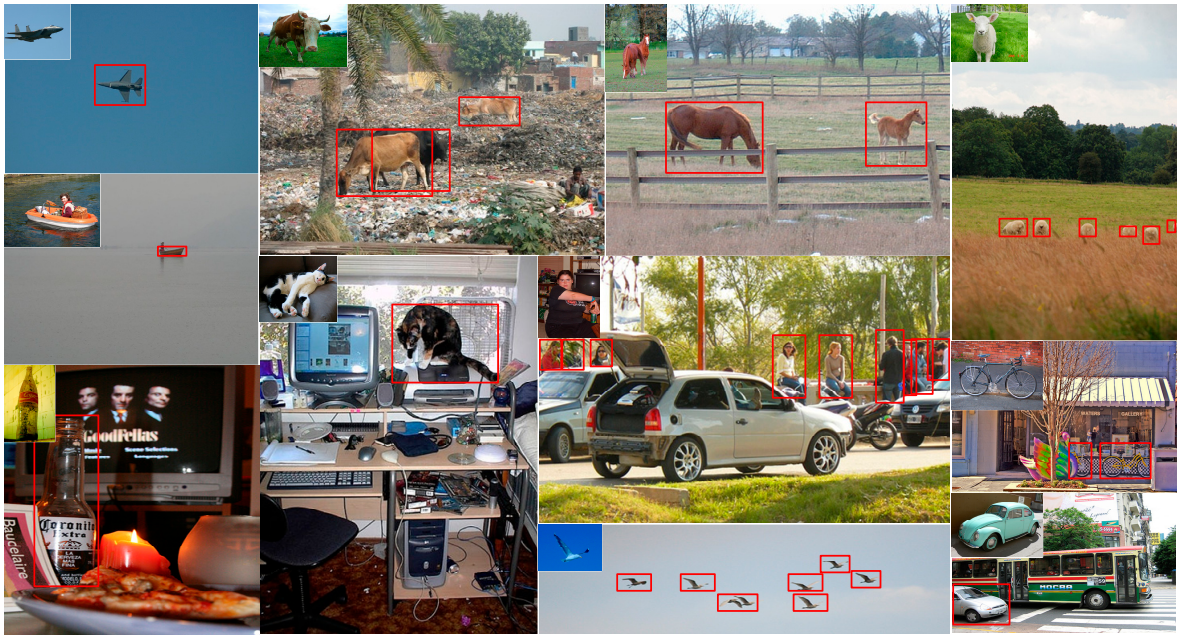


Figure 3. There are some hard samples selected from Pascal VOC test set and the 1-shot detection results of our method.

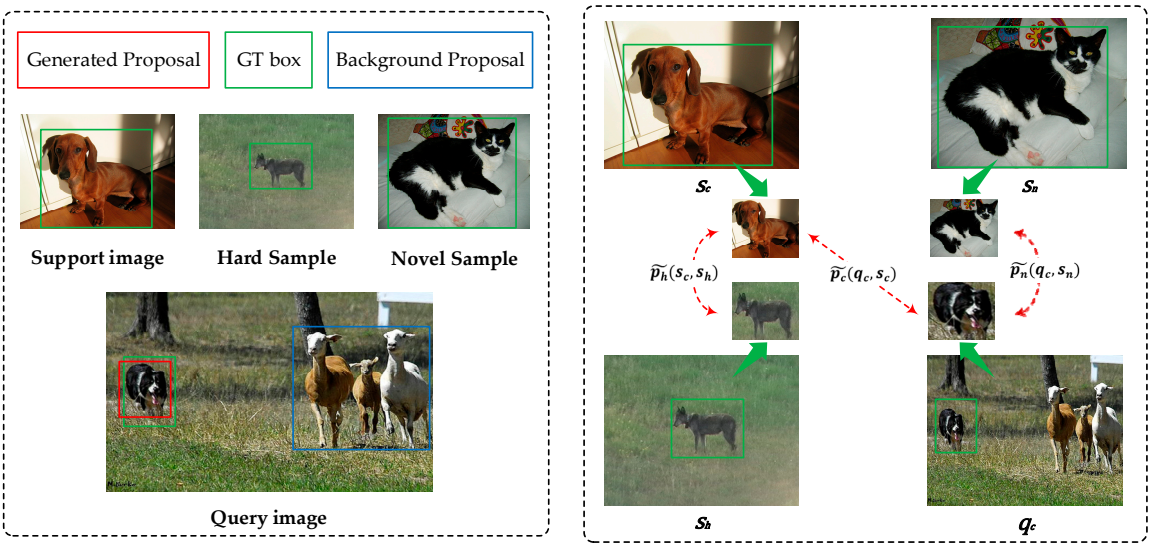
3.3. Training Procedure

Two-Stage Training strategy Our training procedure consists of two stages: the base-class training stage on samples from C_b ($C_{train} = C_b$), followed by K-shot few-shot fine-tuning stage on a balanced set of samples from both C_b and C_n ($C_{train} = C_b \cup C_n$). More precisely, in the second stage where only K labeled samples are individually available for each class in C_n, K samples are randomly selected for each class in C_b to balance the training iterations between C_b and C_n.

Generally, a naive training strategy is matching the objects of same class by constructing a training pair $\widetilde{p_c}(q_c, s_c)$ where the q_c and s_c are both in the same c-th class. However, a powerful model should not only can perform query-support feature similarity mining but also allow capturing the inter-class correlation among different categories. For this reason, according to the different matching results in Figure 4, we present a novel 4-way tuple-contrast training strategy to match the same category while distinguishing different categories. We randomly choose a query image q_c, a support image s_c and a hard sample s_h containing the same c-th category object and one other support image s_n containing a different n-th category object, to construct the training pair $\widetilde{p_t}(q_c, s_c, s_h, s_n)$, where c ≠ n. In the training pair $\widetilde{p_t}(q_c, s_c, s_h, s_n)$, only the objects of c-th category in the q_c are needed and annotated as foreground, while all other objects are neglected and treated as background.

During training, our model learns to match every proposal generated by the Correlation-RPN in the q_c with the object of s_c. Thus, the model needs to not only match the same category objects from $\widetilde{p_c}(q_c, s_c)$ and $\widetilde{p_h}(s_c, s_h)$, but also distinguish objects in different categories from $\widetilde{p_n}(q_c, s_n)$. Nevertheless, there are a massive amount of background proposals, especially with s_h, which usually dominate the training. Taking the consideration of this reason, we adjust these training pairs $\tilde{p}$ to balance the ratio of proposals between queries and supports. The ratio of $\tilde{p}$ is kept as 2:1:1 for $\widetilde{p_c}(q_c, s_c)$, $\widetilde{p_h}(s_c, s_h)$ and $\widetilde{p_n}(q_c, s_n)$. According to their matching scores, we pick all N $\widetilde{p_c}(q_c, s_c)$ and select top $2N$ $\widetilde{p_h}(s_c, s_h)$ and top N $\widetilde{p_n}(q_c, s_n)$ respectively and calculate the matching loss on the selected training pairs.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0633.v1



(**a**) 4-way Tuple-Contrast Training Triplet (**b**) Training pairs

Figure 4. The 4-way tuple-contrast training triplet and different training pairs. s_c: positive support image; s_h: hard sample; s_n: support image of novel class. The s_c and s_n both have the same class with the ground truth in the q_c. The training pair $\widetilde{p_t}(q_c, s_c, s_h, s_n)$ consists of the query-support training pair $\widetilde{p_c}(q_c, s_c)$, the support-hard training pair $\widetilde{p_h}(s_c, s_h)$ and the query-novel training pair $\widetilde{p_n}(q_c, s_n)$.

Detection Loss Function During the training process, we use the multi-task loss on each sampled proposal. It is worth mentioning that we choose the different loss function during optimizing the network in two stages. In the first stage, for each bounding box B, we predict a 2D classification vector $p \in [0,1]$ to represent the probability for target object and background respectively. Inspired by [35–37], concretely, for a mini-batch of N RoI box features $\{r_{x,y}^i, u_{x,y}^i, t_{x,y}^i\}_i^N$, our first stage loss $\boldsymbol{L}_{first}$ is defined as follows with considerations tailored for detection:

$$\boldsymbol{L}_{first} = \frac{1}{N}\sum_{i}^{N} f\left(u_{x,y}^i, p_i\right) \cdot \boldsymbol{L}_{r_{x,y}^i} \tag{6}$$

$$\boldsymbol{L}_{r_{x,y}^i} = \frac{1}{N_{t_i} - 1}\sum_{j=1, j\neq i}^{N} \vec{\mathrm{S}}\{t_i = t_j\} \cdot \log \frac{\exp\left(\widetilde{r_i} \cdot \widetilde{r_j}/\vartheta\right)}{\sum_{k=1}^{N} \vec{\mathrm{S}}\{k \neq i\} \cdot \exp\left(\widetilde{r_i} \cdot \widetilde{r_k}/\vartheta\right)}, \tag{7}$$

where x, y denotes locations and ϑ is the hyper-parameter temperature as in [36]. The $r_{x,y}^i$ refers to encoded RoI feature of detector head for i-th region proposal generated by Correlation-RPN, $u_{x,y}^i$ denotes the IOU score of $r_{x,y}^i$ with matched ground truth bounding box B^*, and $t_{x,y}^i$ denotes the truth annotation.

The second stage output also includes the vector $p \in [0,1]$ for distinguishing between background and target object classes. Different from the first stage, following the parameterization in [14], we present a new regression vector $t = (t_x, t_y, t_w, t_h)$, to specify a scale-invariant translation and height/width shift of log-space relative to a region proposal. In the second stage, we adopt binary cross entropy (BCE) loss for classification and smooth L1 loss for regression. In combination:

$$\boldsymbol{L}_{second} = \frac{\lambda_{cls}}{N}\sum_{i}^{N} \boldsymbol{BCE}(c_i^*, p_i) + \frac{1}{N}\sum_{i}^{N} \vec{\mathrm{S}}\{c_i^* == 1\} \cdot \boldsymbol{L}_{smooth}\left(t_i - u_{x,y}^i\right), \tag{8}$$

where c^* refers to class label for target object and λ_{cls} denotes a balancing factor, which we empirically set to 2.

Our total loss function is the combination of the first and second stage loss:

$$\boldsymbol{L}_{total} = \boldsymbol{L}_{first} + \rho \cdot \boldsymbol{L}_{second}, \tag{9}$$

 doi:10.20944/preprints202404.0633.v1

where ρ = 3 is a balancing factor for second stage loss.

Background-Suppression (BS) Regularization In our proposed structure of CRTED, feature metric matching is developed with the encoder architecture design in transformer by multi-head attention mechanism. It is sure that this design can moderate the training stress for objects with various sizes, but it may still disturb the detector performance when for localization in the scenario of hard samples, especially when in the few-shot condition. For this reason, we propose a novel background-suppression (BS) regularization, by utilizing object knowledge in the domain of ground-truth bounding boxes for each training pair $\widetilde{p_c}(q_c, s_c)$. Specifically, for the q_c in $\widetilde{p_c}(q_c, s_c)$, we first obtain the middle-level F_q of target domain generating from Correlation-RPN. Then, we adopt a masking method that enables the ground-truth labels of target objects in the image s_c to be mapped to the convolutional cube. Consequently, we can identify the feature regions corresponding to background, namely $\boldsymbol{R}_{BS}$. To minimize the adverse effects of background disturbances, we choose L2 regularization to penalize the activation of $\boldsymbol{R}_{BS}$:

$$\boldsymbol{L}_{BS} = BS(F_q) = ||\boldsymbol{R}_{BS}||_2 \tag{10}$$

With this $\boldsymbol{L}_{BS}$, CRTED can depress regions of indifference while pay more attention to where we interest, which is especially important for training in few-shot learning. More details and visualization results of the experiment are shown in Sec. 4.3.

Proposal Consistency Control One of the differences of image classification and object detection is that the former extract semantics from the entire image while the classification signals for the latter come from region proposals. We adopt a settled IoU threshold $\boldsymbol{T}_{iou}$ to assure the consistency of proposals, with the consideration that low IoU proposals may result in excessive deviation to the center of regressed objects, therefore might include irrelevant semantic information. In the following formula, $f(\cdot)$ is responsible for controlling the consistency of proposals, defined with proposal consistency threshold φ:

$$\boldsymbol{T}_{iou} = f(u_{x,y}^i) = \frac{1}{N}\sum_i^N \vec{S}\{u_i \geq \varphi\} \cdot r(u_{x,y}^i), \tag{11}$$

where $r(\cdot)$ can re-weight for object proposals with different level of IoU scores. We experimentally find that φ = 0.75 is a good cutoff point which the detector head can be trained according to most centered object recommendations.

4. Experiments

We mainly perform extensive experiments in both few-shot object detection (FSOD) benchmark datasets PASCAL VOC and MS COCO to assess the effectiveness of our proposed CRTED.

4.1. Few-Shot Object Detection Benchmarks

Pascal VOC [38] consists of images with object annotations of 20 categories where the categories split for C_b and C_n are 15 and 5 separately. We use train set $D_b \cup D_n$ from Pascal VOC 07+12 trainval sets for training, where D_n is randomly sampled from previously unseen novel classes with *K*-shot in {1, 2, 3, 5, 10}. Following the existing works [12,39,40], we consider the same three random partitions of base / novel classes and samplings introduced. Each split is referred as: Novel Split set 1: {"bottle", "aeroplane", "sofa", "cow", "horse" / others}; Novel Split set 2: {"bus", "horse", "motorbike", "cow", "sofa" / others}; Novel Split set 3: {"boat", "cat", "aeroplane", "sheep", "sofa" / others}. For fair comparison, in each partition, we use the same sampled novel instances, and report AP50 for the detection precision for C_b (bAP50) and C_n (nAP50) on Pascal VOC 07 test set. Results are averaged over 10 randomly sampled support datasets.

MS COCO [41] is a large-scale and more challenging object detection dataset, which consists of 80 categories where $|C_b|$ = 60, $|C_n|$ = 20 and C_n are common to Pascal VOC. The train set $D_b \cup D_n$ are from MS COCO 2017 train set, and we perform evaluations on 5K images from COCO 2017 val dataset, which the number of shots is set to 1, 2, 3, 5, 10 and 30. The COCO-style detection precision

 doi:10.20944/preprints202404.0633.v1

of $C_b \cup C_n$ (AP), C_b (bAP) and C_n (nAP) are reported. Results are averaged over 5 randomly sampled support datasets.

4.2. Implementation Details

We follow the training pipeline of [21], and the basic deep architecture of our CRTED is trained end-to-end on 4 V100 GPUs parallel with optimizer Adam and a batch size of 8 for each $\widetilde{p_t}(q_c, s_c, s_h, s_n)$. During training, we find that more training iterations may result to the model over-fit to the training set $D_b \cup D_n$ and damage performance. The learning rate is thus experimentally set to 0.01 during the first training stage and gradually decayed to 0.0002 for later 500 iterations, which can lead to a better converge point. We first perform on MS COCO 2017 training dataset and ensure only simple class object appearing in images for each $\widetilde{p_t}(q_c, s_c, s_h, s_n)$. During few-shot training, for one q_c belonging to $C_b \cup C_n$ in $\widetilde{p_t}(q_c, s_c, s_h, s_n)$, we provide 40 support images, containing 30 belonging to C_b and 10 belonging to C_n, termed as 4-way 10-shot contrastive training.

4.3. Ablation Studies

Evaluation of Correlation-RPN We take sufficient experiments to assess our Correlation-RPN on different training strategies. To evaluate the proposal quality, we first compare the precision and recall on top 50 proposals of the regular RPN, attention RPN and our proposed Correlation-RPN at 0.5 IoU threshold. In addition, we also add the average best overlap ratio (ABO) across ground truth bounding box B^* as one of our evaluation metrics. As demonstrated in Table 1, our model with Correlation-RPN reveals better performance than the other two counterparts under the same training pairs and *K*-shot, producing performance improvement on all the evaluation, which indicate that our proposed RPN architecture can generate more object-relevant proposals to benefit the total detection prediction.

Table 1. Ablation studies on proposed Correlation-RPN and other counterparts.

Method			Precision	Recall	AP	ABO
Regular-RPN	Attention-RPN	Correlation-RPN				
√			0.7923	0.8804	54.5	0.7127
	√		0.8345	0.9130	56.9	0.7282
		√	**0.8509**	**0.9214**	**57.1**	**0.7335**

Specially, the visualization comparison of the attention from superficial layers, between our Correlation-RPN and other two counterparts, are also clearly offered in Figure 5. The results confirm that our Correlation-RPN owns the better ability to pay attention to target domain and provide more high-quality proposals. Especially when dealing with challenging samples, particularly objects that are significantly occluded, our method demonstrates an impressive ability to achieve a high level of confidence in accurate recognition, which ensures reliable and robust performance even in situations where traditional methods might struggle.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0633.v1

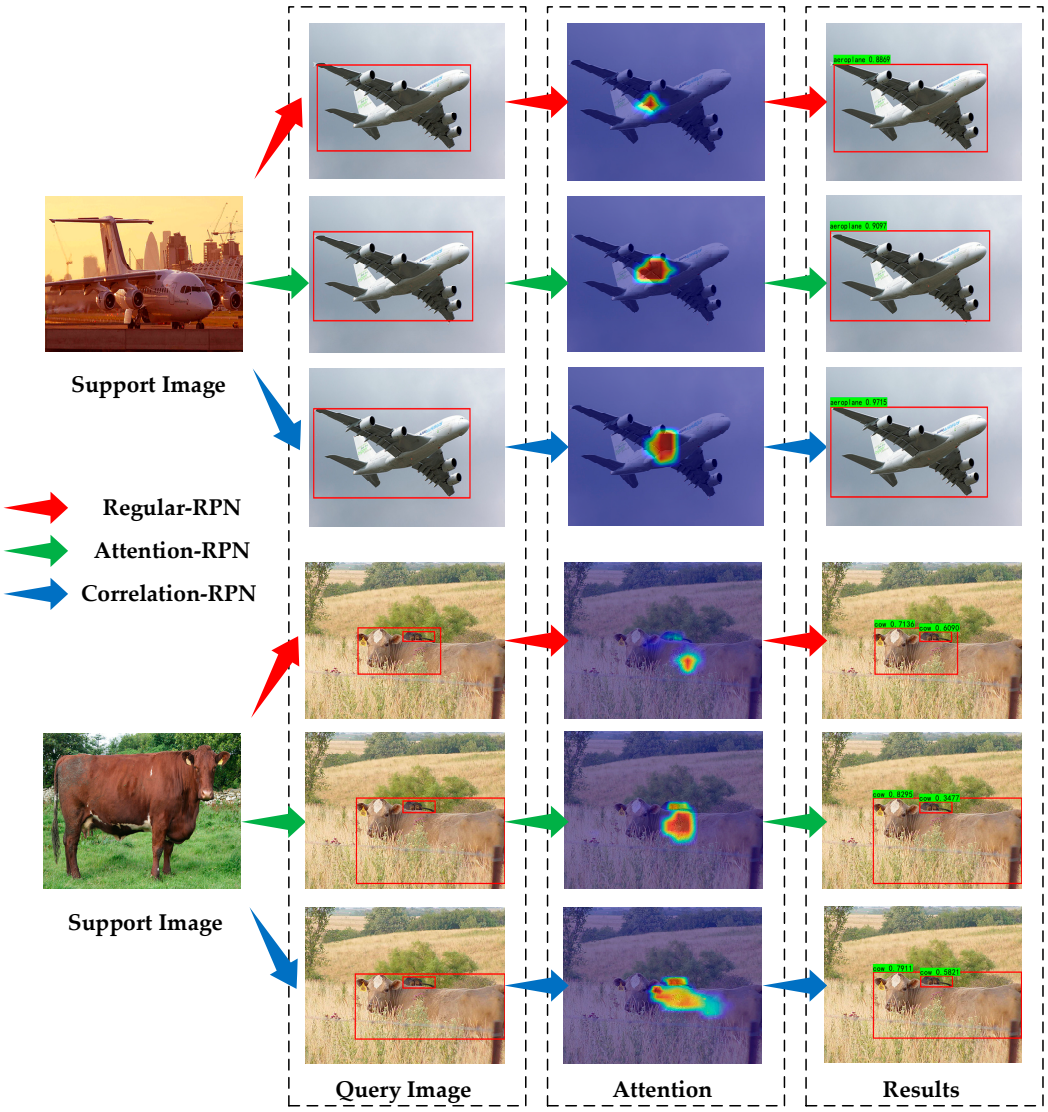


Figure 5. Visualization comparison between models with Correlation-RPN and other RPN architecture on one-shot object detection on Pascal VOC class novel split 1. Correlation-RPN can focus on the most representative region, resulting in more precise proposals and regression box.

Analysis of Matching Procedure for CRTED Figure 6 obviously shows the feature visualization of different object classes learned with and without the matching procedure under the same constraint. As demonstrated, with matching procedure introduced to learn inter-class correlation and capture intra-class compactness, different classes are better separated from each other, which helps to reduce model misclassification and boost generalization ability among similar categories. Specially, Table 2 clearly verify the effectiveness of our proposed matching procedure. When we adopt our method into the model, regardless of the number of support classes, CRTED can still boost detection performance of novel classes under the 1-shot setup, which indicates the capacity of our designed matching procedure in support-query feature similarity metric mining and the strong power of Transformer Encoder-Decoder. It is worth nothing that when there are multiple support classes, MP is able to exploit inter-class similarity of support-query image to obtain improvement of few-shot detection performance, especially about 2.8% mAP under 5-shot setting.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0633.v1

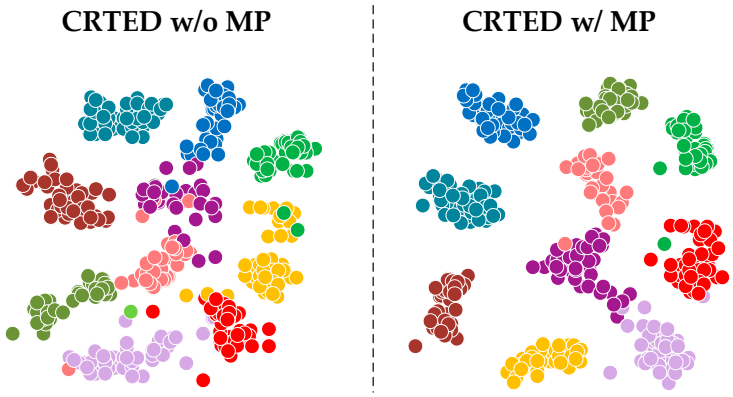


Figure 6. The t-SNE visualization of object classification in the feature space with and without our proposed matching procedure MP on novel split set 2 of Pascal VOC.

Table 2. Ablation studies to validate the effectiveness of our presented MP. C means the number of support classes.

Method	MP	C	Novel mAP (IoU = 0.5)				
			1	2	3	5	10
CRTED		1	28.2	43.3	51.6	54.0	60.3
	√	1	**31.3**	**45.2**	**53.1**	**56.8**	**63.0**
		5	33.7	46.5	52.4	57.1	61.8
	√	5	**37.3**	**51.1**	**54.5**	**58.2**	**63.3**

Impact of Background-Suppression (BS) Regularization For the purpose of enhancing few-shot detection, we mainly evaluate whether our proposed BS regularization method can boost transfer learning for CRTED. As shown in Figure 7 that our background-suppression (BS) regularization can effectively help CRTED to reduce the background disturbances. And it is obviously shown in Table 3, our proposed regularization method can significantly improve the performance, when the support images s_c in training set $\tilde{p}_t(q_c, s_c, s_h, s_n)$ is scarce in the few-shot domain, especially in one-shot. Additionally, it is worth mentioning that we try to pick objects from as many categories as possible to verify the effect, which show that BS is generally robust to different categories.

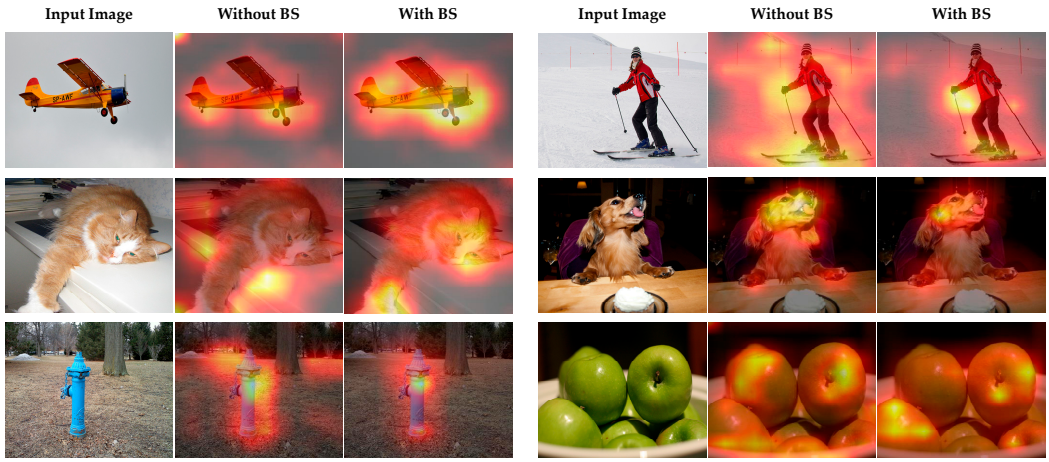


Figure 7. Feature heatmap of Background-Suppression (BS) regularization. Samples of different categories are selected from COCO val set. BS can effectively weaken negative influence from background disturbances, and then activate CRTED to pay attention to object-relevant regions.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0633.v1

Table 3. Regularized transfer learning for CRTED of AP50 on the Pascal VOC dataset on $C_b \cup C_n$. BL: baseline. BS: background-suppression (BS) regularization. The mAP results show that, our proposed BS regularization method can significantly boost the baseline CRTED, when in few-shot object detection.

Shots for Split 1	1	2	3	5	10
CRTED $_{BL}$	68.7	69.4	70.8	73.6	75.5
CRTED $_{BL+BS}$	**69.8**	**70.2**	**72.0**	**75.4**	**76.5**
Shots for Split 2	1	2	3	5	10
CRTED $_{BL}$	65.5	66.8	69.9	71.3	73.3
CRTED $_{BL+BS}$	**67.7**	**68.0**	**71.3**	**71.8**	**73.7**
Shots for Split 3	1	2	3	5	10
CRTED $_{BL}$	67.7	68.7	71.4	72.7	74.8
CRTED $_{BL+BS}$	**68.6**	**70.4**	**72.7**	**73.9**	**75.1**

Ablation of Training CRTED Refer to Table 4. We train our network with different training strategies and obtain 1.3% AP_{50} improvement at the $\widetilde{p_t}$ 10-shot training strategy, comparing with the $\widetilde{p_c}$ 10-shot training strategy. It is straightforward that the model performs better at training strategy with $\widetilde{p_t}$ than with $\widetilde{p_c}$, which shows the importance of training pairs $\widetilde{p_n}$ and $\widetilde{p_h}$, included in $\widetilde{p_t}$. With adding $\widetilde{p_n}(q_c, s_n)$, even the object in q_c belonging to unseen class, the model can learn inter-class correlation and intra-class compactness from novel classes to enhance generalization ability. And model can improve robustness with $\widetilde{p_h}(s_c, s_h)$, especially q_c is a hard sample. It is clear that with larger *K*-shot training, we achieve better performance which also indicates a certain number of support images is beneficial to few-shot learning. We think that controlling the number of s_h and s_n to 1 can suffice in training the model for distinguishing different classes. And from Figure 8, our proposed training strategy can positively activate the training progress of our detector. Our full model thus adopts the $\widetilde{p_t}$ 10-shot training strategy.

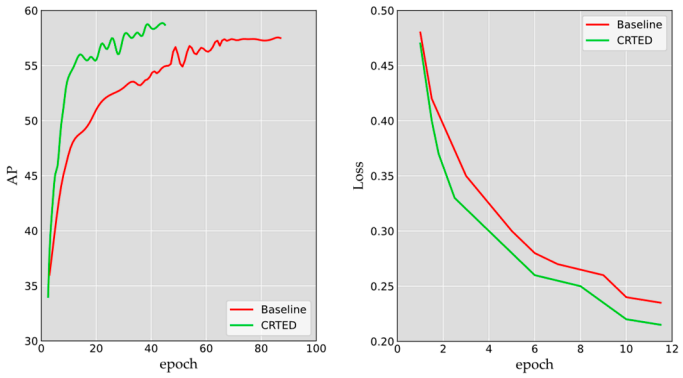


Figure 8. Comparison of performance gains (left) and object classification loss (right).

Table 4. Experimental results for different training strategies.

Training strategy	AP	AP_{50}	AP_{75}
$\widetilde{p_c}$ 1-shot	48.5	55.9	41.1
$\widetilde{p_t}$ 1-shot	**48.7**	**56.2**	**41.2**
$\widetilde{p_c}$ 5-shot	57.7	68.1	47.3
$\widetilde{p_t}$ 5-shot	**58.1**	**68.4**	**47.7**
$\widetilde{p_c}$ 10-shot	58.8	69.2	48.4
$\widetilde{p_t}$ 10-shot	**59.7**	**70.5**	**48.8**

 doi:10.20944/preprints202404.0633.v1

4.4. Comparison with State-of-the-Arts

Comparisons between our approaches and state-of-the-art few-shot object detectors on Pascal VOC and COCO are shown in Table 5 and Table 6. Following the default novel split setting of data with different *K*-shot in previous researches, our proposed CRTED without further fine-tuning can set comparable performance with fine-tuned methods or even new SOTA results for few-shot object detection. Our CRTED outperforms A-RPN without fine-tuning by 2.6% on AP metrics, which demonstrates the strong generalization ability of our detector, especially in few-shot scenario. Specially, in terms of AP over different classes, our CRTED obtains the comparable performance for several cases on Pascal VOC and performs better than most of the representative fine-tuned models with 1-shot on COCO val dataset, boosting about ~0.3% to ~3.8% improvement.

Table 5. Performance comparison of nAP_{50} on Pascal VOC. Red and green fonts denote the best and second-best performance, respectively. With fine-tuning, CRTED achieves a new SOTA performance on 1-shot setting, demonstrating its strong generalization capability.

Method	Fine-tune	nAP_{50} (Avg. on splits for each shot)				
		1	2	3	5	10
FSRW [12]	√	16.6	17.5	25.0	34.9	42.6
Meta R-CNN [13]	√	11.2	15.3	20.5	29.8	37.0
TFA_{fc} [23]	√	27.6	30.6	39.8	46.6	48.7
TFA_{cos} [23]	√	31.4	32.6	40.5	46.8	48.3
FSDetView [39]	√	26.9	20.4	29.9	31.6	37.7
A-RPN [21]	×	18.1	22.6	24.0	25.0	-
AirDet [42]	×	21.3	26.8	28.6	29.8	-
DiGeo [43]	√	31.6	36.1	45.8	51.2	55.1
CRTED (Ours)	×	20.7	25.6	28.6	30.0	34.4
	√	**31.8**	**32.8**	34.0	45.0	**48.9**

Table 6. Performance comparison of AP with k-shot on COCO validation dataset. Red and green fonts denote the best and second-best performance, respectively. CRTED achieves comparable performance on baseline without fine-tuning and outperforms most of the representative methods with fine-tuning, which indicates its strong power.

Method	Venue	Fine-tune	Shots					
			1	2	3	5	10	30
FSRW [12]	ICCV 2019	√	-	-	-	-	5.6	9.2
Meta R-CNN [13]	ICCV 2019	√	-	-	-	-	8.7	12.4
TFA_{fc} [23]	ICML 2020	√	2.8	4.1	6.3	7.9	9.1	-
TFA_{cos} [23]	ICML 2020	√	3.1	4.2	6.1	7.6	9.1	12.1
FSDetView [39]	ECCV 2020	√	2.2	3.4	5.2	8.2	12.5	-
MPSR [22]	ECCV 2020	√	3.3	5.4	5.7	7.2	9.8	-
A-RPN [21]	CVPR 2020	×	4.3	4.7	5.3	6.1	7.4	-
W. Zhang et al. [44]	CVPR 2021	√	4.4	5.6	7.2	-	-	-
FSCE [45]	CVPR 2021	√	-	-	-	-	11.1	15.3
FADI [46]	NIPS 2021	√	5.7	7.0	8.6	10.1	12.2	-
AirDet [42]	ECCV 2022	×	6.0	6.6	7.0	7.8	8.7	12.1
CRTED	**Ours**	×	**5.8**	6.2	**7.2**	7.4	8.6	**12.4**

5. Conclusion

This paper presents a novel few-shot object detection model, CRTED, which introduces a newly proposed object-relevant region-based modules Correlation-RPN and the powerful structure of

 doi:10.20944/preprints202404.0633.v1

Transformer Encoder-Decoder. Our model without fine-tuning has been trained and validated on Pascal VOC and COCO datasets, and extensive qualitative experimental results have been given. Specifically, with proposed matching procedure, BS regularization and novel 4-way tuple-contrast training strategy, CRTED can perform comparably or even better than those detectors with exhaustively fine-tuning in the same evaluation. We hope that this work can lead to good inspiration for further works in few-shot object detection.

Author Contributions: Conceptualization, K.X.; methodology, K.X.; validation, K.X.; formal analysis, K.X., L.J., Z.X., Y.N. and J.C.; investigation, K.X.; resources, K.X.; data curation, K.X.; writing—original draft preparation, K.X., L.J. and Z.X.; writing—review and editing, K.X., L.J., Y.N. and Z.X.; visualization, K.X.; supervision, Z.X., Y.N. and J.C.; project administration, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Guangxi Science and Technology Development Project (AB23026135; AB21220011), the Guilin Science and Technology Plan Project (20210220).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In European conference on computer vision, pages 213–229. Springer, 2020
2. Guangxing Han, Xuan Zhang, and Chongrong Li. Semi-supervised dff: Decoupling detection and feature flow for video object detectors. In Proceedings of the 26th ACM international conference on Multimedia, pages 1811–1819, 2018.
3. Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.
4. Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14454–14463, 2021.
5. Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605, 2022.
6. Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In NeurIPS, 2017.
7. Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In ICLR, 2017.
8. Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In ICML, 2017.
9. Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. Memory matching networks for one-shot image recognition. In CVPR, 2018.
10. Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In CVPR, 2018.
11. Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. Lstd: A low-shot transfer detector for object detection. In AAAI, 2018.
12. Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In ICCV, 2019.
13. Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In ICCV, 2019.
14. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
15. R. Girshick. Fast R-CNN. In ICCV, 2015.

16. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6), 1137–1149 (2016)
17. He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.
18. Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.
19. glenn jocher et al. yolov5. https://github.com/ ultralytics/yolov5, 2021.
20. Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 7464-7475.
21. Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In CVPR, 2020.
22. Wu, J., Liu, S., Huang, D., Wang, Y.: Multi-Scale Positive Sample Refinement for Few-Shot Object Detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 456–472 (2020)
23. Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. Frustratingly simple few-shot object detection. In International Conference on Machine Learning, pages 9919–9928. PMLR, 2020.
24. Wojke N, Bewley A. Deep cosine metric learning for person re-identification[C]//2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018: 748-756.
25. Karlinsky L, Shtok J, Harary S, et al. Repmet: Representative-based metric learning for classification and few-shot object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5197-5206.
26. Song H, Sun D, Chun S, et al. Vidt: An efficient and effective fully transformer-based object detector[J]. arXiv preprint arXiv:2110.03921, 2021.
27. Li Y, Mao H, Girshick R, et al. Exploring plain vision transformer backbones for object detection[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 280-296.
28. Fang Y, Yang S, Wang S, et al. Unleashing vanilla vision transformer with masked image modeling for object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 6244-6253.
29. Liu F, Zhang X, Peng Z, et al. Integrally migrating pre-trained transformer encoder-decoders for visual object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 6825-6834.
30. Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159, 2020.
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
32. Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In ECCV, 2016.
33. Erika Lu, Weidi Xie, and Andrew Zisserman. Class-agnostic counting. In ACCV, 2018.
34. Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In CVPR, 2019.
35. Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In Advances in neural information processing systems, pages 1988–1996, 2014.
36. Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. Advances in Neural Information Processing Systems, 33, 2020.
37. Oriol Vinyals Aaron van den Oord, Yazhe Li. Representation learning with contrastive predictive coding. Advances in Neural Information Processing Systems, 31, 2018.
38. Everingham, M., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV (2010)
39. Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In European Conference on Computer Vision (ECCV), 2020.

 doi:10.20944/preprints202404.0633.v1

40. Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-Learning to Detect Rare Objects. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 9924–9933, Seoul, Korea (South), October 2019. IEEE.
41. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Doll'ar, P.: Microsoft coco: Common objects in context. In: ECCV (2014)
42. Li B, Wang C, Reddy P, et al. Airdet: Few-shot detection without fine-tuning for autonomous exploration[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 427-444.
43. Ma J, Niu Y, Xu J, et al. Digeo: Discriminative geometry-aware learning for generalized few-shot object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 3208-3218.
44. Zhang, W., Wang, Y.X.: Hallucination Improves Few-Shot Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13008–13017 (2021)
45. Sun B, Li B, Cai S, et al. Fsce: Few-shot object detection via contrastive proposal encoding[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 7352-7362.
46. Cao Y, Wang J, Jin Y, et al. Few-shot object detection via association and discrimination[J]. Advances in neural information processing systems, 2021, 34: 16570-16581.