


Article

Small-Sample Target Detection Across Domains Based on Supervision and Distillation

Fusheng Sun ^{1,2,3} , Jianli Jia ^{1,2,3}, Xie Han ^{1,2,3}, Liqun Kuang ^{1,2,3} and Huiyan Han ^{1,2,3,*}¹ Shanxi Key Laboratory of Machine Vision and Virtual Reality, Taiyuan 030051, China; fssun@nuc.edu.cn (F.S.)² School of Computer Science and Technology, North University of China, Taiyuan 030051, China³ Shanxi Province's Vision Information Processing and Intelligent Robot Engineering Research Center, Taiyuan 030051, China

* Correspondence: 20050537@nuc.edu.cn

Abstract: To address the issues of significant object discrepancies, low similarity, and image noise interference between source and target domains in object detection, we propose a supervised learning approach combined with knowledge distillation. Initially, student and teacher models are jointly trained through supervised and distillation-based approaches, iteratively refining the inter-model weights to mitigate the issue of model overfitting. Secondly, a combined convolutional module is integrated into the feature extraction network of the student model, to minimize redundant computational effort; an explicit visual center module is embedded within the feature pyramid network, to bolster feature representation; and a spatial grouping enhancement module is incorporated into the region proposal network, to mitigate the adverse effects of noise on the outcomes. Ultimately, the model undergoes a comprehensive optimization process that leverages the loss functions originating from both the supervised and knowledge distillation phases. The experimental results demonstrate that this strategy significantly boosts classification and identification accuracy on cross-domain datasets; when compared to the TFA (Task-agnostic Fine-tuning and Adapter), CD-FSOD (Cross-Domain Few-Shot Object Detection) and DeFRCN (Decoupled Faster R-CNN for Few-Shot Object Detection), with sample orders of magnitude 1 and 5, increased the detection accuracy by 1.67% and 1.87%, respectively.



Citation: Sun, F.; Jia, J.; Han, X.; Kuang, L.; Han, H. Small-Sample Target Detection Across Domains Based on Supervision and Distillation. *Electronics* **2024**, *13*, 4975. <https://doi.org/10.3390/electronics13244975>

Academic Editors: Guanghui Yue, Wei Zhou and Wenhan Yang

Received: 25 November 2024

Revised: 14 December 2024

Accepted: 16 December 2024

Published: 18 December 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: supervision; distillation; combination convolution; explicit visual center; spatial group-wise enhance

1. Introduction

Few-Shot Object Detection (FSOD) stands as a prominent area of inquiry in computer vision, aiming to detect objects from previously unseen categories with the aid of a minimal set of training instances. The existing FSOD algorithm encompasses two primary categories: single-domain FSOD and cross-domain FSOD.

Zhang et al. [1] introduced an unsupervised few-shot target detection framework designed for domain adaptation. Utilizing the PASCAL VOC [2], Clipart [3], and Comic [3] datasets as source domains, they evaluated the framework's performance on the Watercolor [3], Cityscapes [4], and FoggyCityscapes [5] datasets. This was achieved through an end-to-end adversarial learning approach. AcroFOD (An Adaptive Method for Cross-domain Few-shot Object Detection) [6] employs an adaptive optimization strategy to filter out data instances that are highly similar to the target sample from the existing dataset, cross-fuses the foreground and background information of the image, enhances the diversity of the augmented data, and conducts multi-level data enhancement. AsyFOD (An Asymmetric Adaptation Paradigm for Few-Shot Domain-Adaptive Object Detection) [7] achieves cross-domain small-sample target detection by leveraging the concept of domain adaptation. To address the issue of data imbalance between the source and target domains, it introduces an asymmetric adaptation paradigm. The source domain instance is segmented into target

similar instances and target different instances. The former serves to augment the target instance and adheres to the principle of asynchronous alignment, thereby mitigating premature overadaptation resulting from imbalanced data distribution. The region migration module, which employs an attention mechanism, accomplishes the distribution alignment between the source domain and the foreground domain within the target domain, and constructs a domain classifier for each category. CD-FSOD [8] introduces a cross-domain few-sample target detection benchmark, with MS COCO [9] as the source domain, which employs the ArTaxOr (Cross-Domain Few-Shot Object Detection) [10], UODD (Unseen Object Detection Dataset) [11], and DIOR (A Large-Scale Benchmark Dataset for Object Detection in Aerial Images) [12] datasets as the target domains for detection, leveraging a two-stage fine-tuning approach. It optimizes detection performance through the utilization of both student and teacher models.

It can thus be observed that with single-domain FSOD and cross-domain FSOD, as two distinct settings of small-sample object detection, the former primarily concentrates on conducting efficient object detection within the same or similar data domains by utilizing a limited number of labeled samples. Its core challenge lies in how to, fully excavate and leverage prior knowledge, in the situation of scarce samples, through methods such as transfer learning and meta-learning, in order to train a detection model with favorable performance. The latter method further expands to scenarios where there exist significant domain disparities between the source and target domains. Besides encountering the small-sample issue, it also needs to address the additional challenges arising from the domain differences, such as small inter-class distances and ambiguous foreground–background boundaries. This demands that the algorithm not only possesses the ability to learn with a small number of samples, but also is capable of adapting to the visual styles and feature distributions in different domains. By introducing domain adaptation modules, using learnable instance features, and other strategies, the cross-domain generalization ability of the model can be enhanced. Hence, compared with single-domain FSOD, cross-domain FSOD is more complex and challenging in aspects such as algorithm design, model training, and evaluation standards.

Consequently, Cross-Domain Few-Shot Object Detection (FSOD) has become more prevalent; however, it continues to grapple with two significant challenges: (1) The disparity in data distribution and feature variations across different domains poses significant challenges for model transferability. Training models with limited samples often results in overfitting, which in turn diminishes detection performance in the target domain. Preserving the model's performance from the source domain necessitates the design of effective regularization techniques and data augmentation strategies to mitigate overfitting. To address this, we embrace the concepts of supervision and distillation, employing a two-stage fine-tuning approach to construct the network model. This method facilitates weight transitions between models using exponential moving averages, thereby alleviating the issue of model overfitting. (2) The cross-domain few-sample target detection method typically encounters experimental target domain datasets that are often limited to a specific field, leading to poor generalization performance of the model when faced with data from other fields. When the domain discrepancy between the source and target domain datasets is substantial, the original model struggles to generalize effectively to new samples. Enhancing the model's generalization ability in the target domain is a critical challenge. To address this, we propose incorporating a combined convolutional approach, an explicit visual center, and a spatial grouping enhancement module within the feature extraction module, multi-scale fusion module, and candidate region extraction module, respectively. These enhancements are designed to bolster feature representation and improve the model's generalization capabilities.

Given the above content, in the second part of the article, we focus on the deficiencies of the four commonly employed FSOD methods to achieve the detection of a new category target with a small number of training samples. In the third part, we specifically introduce a supervised and distillation-driven approach to realize small-sample target detection

across different domains. In the fourth part, we verify the outstanding performance of our proposed method for classification and identification on cross-domain datasets through numerous comparative and ablation experiments. In the concluding section, we summarize the work content of the entire text and discuss the direction of the work in the subsequent stage.

To summarize, our main contributions include the following:

- (1) We employed a concept grounded in supervision and distillation to construct our network model, utilizing a two-stage fine-tuning approach. Additionally, we implemented exponential moving average to facilitate weight transitions between models, thereby mitigating the issue of model overfitting.
- (2) We integrated combinatorial convolution, an explicit visual center, and a spatial grouping enhancement module within the feature extraction module, as well as a multi-scale fusion module and a candidate region extraction module. These additions were designed to enhance feature representation and bolster the model's generalization capabilities.
- (3) We subjected the proposed method to quantitative experiments and various ablation studies on the dataset with MS COCO [9] as the source domain and ArTaxOr [10], UODD [11], and DIOR [12] as the target domains. The experimental outcomes affirm the reasonableness and efficacy of the suggested approach.

2. Related Research

Given that limited training data are insufficient to fully support the algorithm in constructing dependable prediction models, FSOD methods typically depend on prior knowledge to address the challenges posed by data scarcity. We divided the FSOD method into four categories according to different models of prior knowledge: data enhancement-based methods, meta-learning-based methods, distance measure-based learning methods, and transfer-based learning methods.

2.1. Methods Based on Data Enhancement

The method described in [13–16], serving as an implementation of the preprocessing phase, is frequently employed to augment the dataset, and represents a prevalent approach to addressing the issue of limited samples. Nonetheless, data augmentation poses significant challenges when a limited number of samples are at hand. These techniques aim to enrich the training dataset by leveraging prior knowledge, thereby enabling the learning of more reliable hypotheses. However, these operations are highly task-specific: the same data augmentation strategy may yield meaningful samples in one task, yet in another scenario, it could result in a data distribution that deviates from the actual application context. Owing to this characteristic, contemporary research often employs transient data transformation techniques, which impede the reuse of data augmentation strategies across various tasks and datasets. Wu et al. [13] introduced the MPSR (Multi-Scale Positive Sample Refinement for Few-Shot Object Detection) algorithm, which employs data augmentation through the creation of multi-scale positive samples. A model was constructed utilizing Faster R-CNN (Faster Region-Convolutional Neural Network) [14]. The Feature Pyramid Network (FPN) [15] was employed to supplant the original backbone architecture, and a category-agnostic regression loss function was implemented within the detection head. Hua et al. [16] utilized the Feature Pyramid Network (FPN) to extract multi-scale features, enhance the Regional Proposal Network (RPN), and introduce a support set image branch to obtain the correlation between the support set and the query set images. However, these operations are highly task-specific: the same data augmentation strategy may yield meaningful samples in one task, yet in another scenario, it could result in a data distribution that deviates from the actual application context.

2.2. Meta-Learning-Based Methods

Meta-learning represents a pivotal avenue within the realm of machine learning [17–20], focusing on the acquisition of learning methodologies. Ideally, when presented with a distribution of learning tasks, a meta-learner should progressively become adept at extracting generalizable insights across these tasks, and fine-tune its algorithms to efficiently cater to such requirements. Kang et al. [17] introduced a novel few-shot target detection approach that leverages feature reweighting. This method harnesses the power of base classes, which are abundant in labeled data, to integrate a meta-feature learner and reweighted modules within a unified single-stage detector framework. The aim is to extract meta-features and create global vectors, thereby enhancing the model's generalization capabilities when encountering new classes. MetaDet (metadata detection) [18] introduces a meta-learning framework that leverages base class data, which is rich in annotation information, to generate meta-knowledge regarding model parameters. This approach facilitates the creation of new class detectors. Additionally, it incorporates a weight prediction meta-model to address the challenges of classifying and locating instances with limited samples, achieving a unified and coherent solution. Meta RCNN (Meta Region-based Convolutional Neural Networks) [19] extracts the class attention vector by inferring image information and employs a channel soft attention mechanism on the features of the ROI (Region of Interest) within the region of interest to reconstruct the head portion of the RCNN predictor. This enables the detection or segmentation of objects that align with the representation of a specific class of vectors. FSDetView (Few-Shot Detection View) [20] partitions the model into a query branch and a class data processing branch. The former operates on the query image, while the latter processes images with bounding box masks. It constructs a pre-training model using ample data, extracts distinguishing features, and subsequently directs the detection of novel object categories with minimal samples.

2.3. Learning Method Based on Distance Measurement

In the realm of computer vision, distance measurement learning stands as a widely adopted technique [21–24]. The core objective is to learn an appropriate mapping function that projects the original high-dimensional sample data into a low-dimensional feature space. In this space, the distance between similar samples is closely compressed, while the distinction between dissimilar samples is significantly enhanced. Schwartz et al. [21] introduced a classification approach and a single-sample target detection method known as RepMet (Representative-based Metric Learning). By generating a model representation with multiple pattern mixing for each class, the optimized backbone network parameters, the embedding space, and the distribution of training categories within that space are utilized for efficient classification and localization. Fan et al. [22] introduced a novel few-shot target detection approach that leverages an attention-based Region Proposal Network (RPN) and a multi-relationship detector. This method facilitates the learning of intricate relationships between query and support sets through the sharing of distinct weight branches. Consequently, the model can identify novel object classes without the need for re-training or fine-tuning. CME (Class Margin Equilibrium) [23], a small-sample target detection method grounded in class edge equilibrium, systematically optimizes the division of feature space and the reconstruction of new class features. Han et al. [24] introduced a query-adaptive few-shot target detection approach leveraging a heterogeneous graph convolutional network. This method employs two parallel branches and an attention-based Region Proposal Network (RPN) to learn from the query set and establish a matching relationship with the support sets. It constructs heterogeneous graphs for candidate bounding boxes and class nodes, thereby generating novel feature representations that enhance the model's predictive and classification capabilities.

2.4. Methods Based on Transfer Learning

The backbone networks employed in a standard target detection framework typically undergo extensive pre-training, utilizing large datasets to minimize the performance gap

between detection and classification tasks. Furthermore, depth detectors are more susceptible to overfitting issues compared to depth classifiers. Therefore, beyond accurate category differentiation, it is essential to create distinct representations that maintain spatial coherence and robustness for each individual target instance, and address the intricate task of localization [25–27]. Chen et al. [25] introduced the LSTD (Low-Shot Transfer Detector) algorithm, which has officially sparked a surge in research on small-sample object detection. The algorithm adeptly merges the fundamental strengths of SSDs (Single-Shot MultiBox Detectors) [26] and Faster R-CNN within a unified framework, streamlining and enhancing the knowledge transfer process under scenarios of limited sample size. Wu et al. [27] introduced a universally applicable prototype-enhanced few-shot target detection approach. This method involves extracting common features from objects across all categories and subsequently augmenting these with universal features to enhance the representation of specific objects. This addresses the constraints associated with the straightforward use of class prototypes within FSOD tasks.

3. Methodology

In this paper, we introduce a supervised, distillation-driven approach to few-shot target detection that spans across different domains. By leveraging supervision and knowledge distillation, we achieve a synergistic enhancement between the student and teacher models, thereby elevating the detection capabilities of the model. A comprehensive overview of our method is depicted in Figure 1.

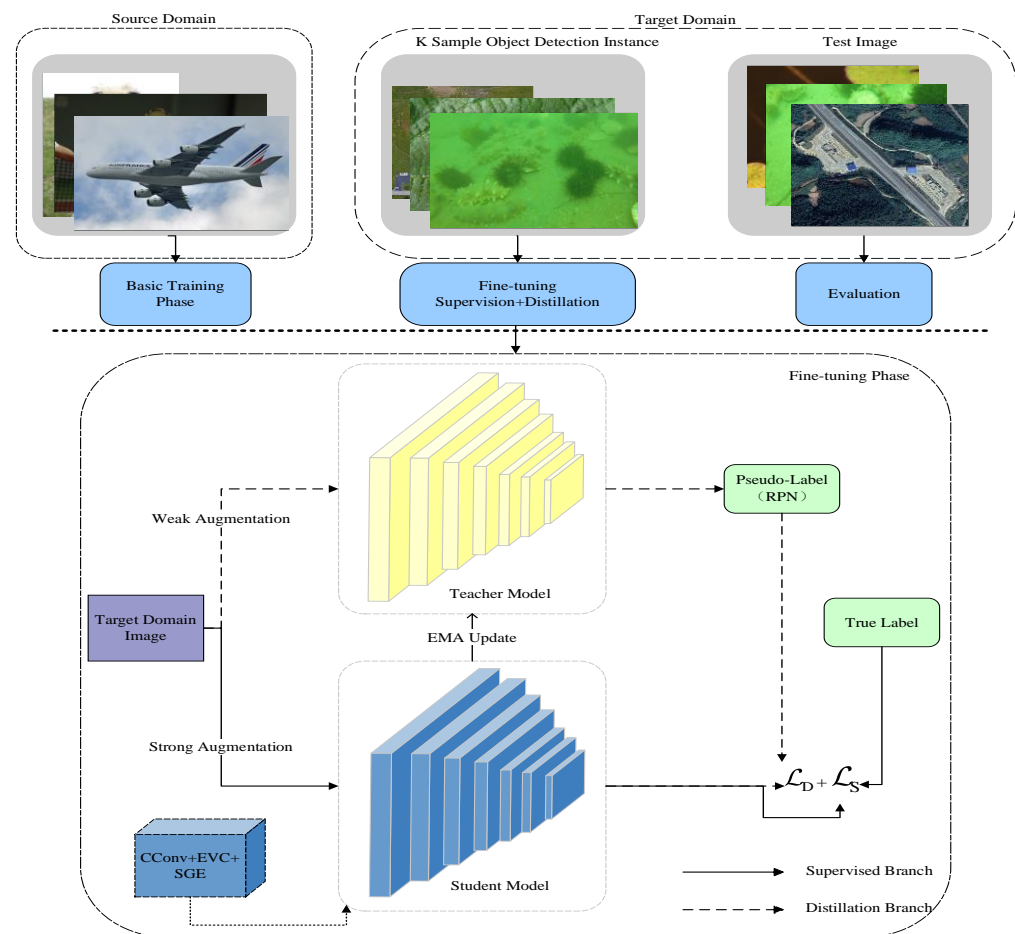


Figure 1. Overview of methods.

During the basic training phase, we utilize the rich source domain data to facilitate the pre-training of the model. Subsequently, during the fine-tuning phase, visual perturbations such as random color jitter, grayscale transformation, and Gaussian blur are introduced

to enhance the dataset and adapt the student model, while generating a set of enhanced images. These enhanced images not only improve the robustness of the model but also help in generalizing the learning process to unseen data. Using the enhanced images of the target domain, the teacher model is fed into the network. The teacher model generates pseudo-labels, and the prediction result (L_D) generated by the joint student model calculates the distillation loss function (L_S). Within the supervised branch, combined convolution (CConv), containing Partial Convolution and PointWise Convolution (PConv and PWConv) [28], is employed to diminish the redundancy of model calculations and reduce memory access during the feature extraction phase. During the multi-scale feature fusion process, Explicit Visual Center (EVC) [29] is employed to capture data dependencies and enhance the feature representation of the data's local key areas. In the process of generating corresponding region candidates, the Spatial Group-wise Enhance (SGE) method [30], which incorporates spatial grouping, is employed to optimize the weight distribution of feature groups. This approach aims to enhance feature representation and mitigate the impact of noise. Subsequently, the generated sample labels are utilized to compute the supervised loss function. The Exponential Moving Average (EMA) [31] is employed to facilitate the weight updates for both the teacher and student models. Ultimately, the model's performance is assessed using the provided target domain data.

3.1. Supervision and Distillation

During the fine-tuning phase, we initially transferred the foundational detector weights to the student model for model initialization. Subsequently, we employed the standard detection [8] supervision loss function to train the student model on k samples from each category's target instance. The training weights were then fed back to both the student and teacher models, enabling them to engage in a synergistic joint training process. The teacher model refined the student model through the distillation branch loss function, while the student model updated the teacher model's weights via Exponential Moving Average (EMA).

The overall loss function during the fine-tuning phase is computed as indicated in Equation (1). Here, L_S represents the loss function for the supervised branch, while L_D corresponds to the loss function for the distilled branch. The hyperparameter λ serves to adjust the relative weights of these two branches within the comprehensive optimization process.

$$L = L_S + \lambda L_D \quad (1)$$

The architecture of the student model is depicted in Figure 2. The PConv and PWConv module is integrated into the feature extraction network, enabling a deeper processing of the initial image features. This results in a more refined feature mapping that is fed into the FPN. Additionally, EVC modules are incorporated into the network to enhance the interaction and aggregation capabilities of features across different layers. After the fusion of the feature afferents with the FPN, the SGE module is employed to enhance the grouping capabilities of the RPN, thereby efficiently screening and optimizing candidate regions. Ultimately, the optimized features undergo decoding and classification through the ROI header network, enabling the prediction of the target object and its associated attribute information within the image. The supervised detection loss function is presented in Equation (2), wherein L_S represents the supervised branch loss function, and L_{cls} and L_{loc} denote the classification and regression loss functions, respectively. These are computed based on the prediction results x_i^s from the supervised branch and the actual sample labels.

$$L_S = \sum_i L_{cls}(x_i^s, y_i^s) + L_{loc}(x_i^s, y_i^s) \quad (2)$$

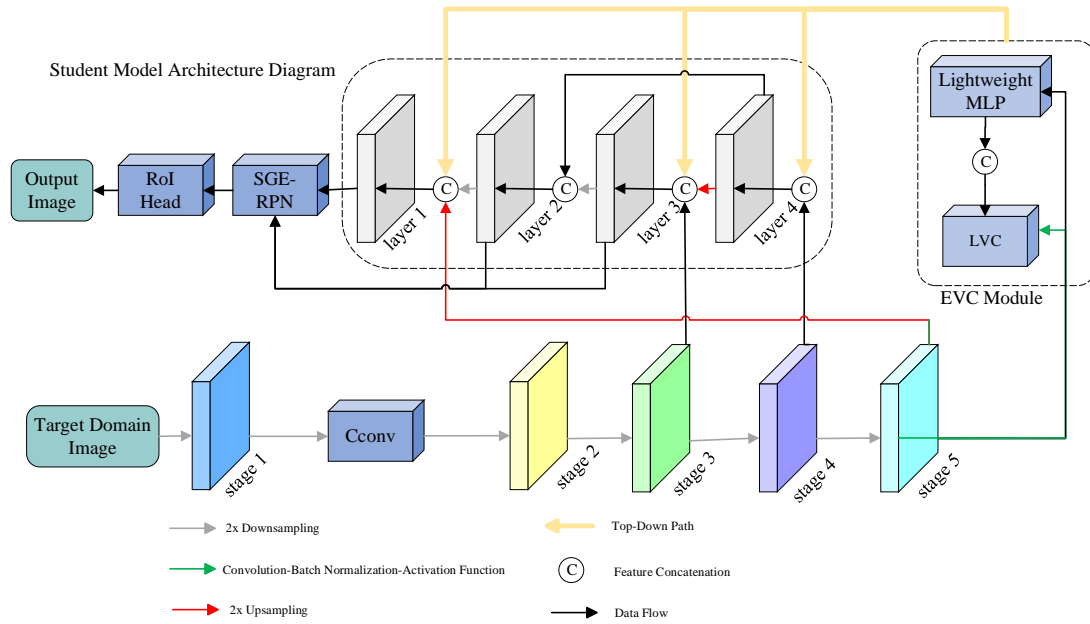


Figure 2. Student model architecture diagram.

Within the distillation field, we performed the enhancement of the target sample images by applying stochastic level flipping and cropping techniques. The essence of this phase is to produce high-quality pseudo-labels for unlabeled images from the target domain. The enhanced images were fed into the teacher model, which then generated a set of candidate boxes via the Region Proposal Network (RPN). Utilizing the foreground scores of the resulting candidate boxes, the one with the highest threshold was chosen as the pseudo-boundary box, denoted as p_i^d . The loss function during the distillation phase was computed utilizing the prediction outcomes x_i^d from the student model and the pseudo-labels furnished by the teacher model. As illustrated in Equation (3), L_D represents the overall loss function for the distillation branch, while L_{cls} and L_{loc} denote the classification and regression loss functions, respectively.

$$L_D = \sum_i L_{cls}(x_i^d, p_i^d) + L_{loc}(x_i^d, p_d^s) \quad (3)$$

The teacher model's weight updates were carried out using the Exponential Moving Average (EMA) technique, which significantly mitigates the issue of overfitting in optimization algorithms such as Adam [17], Batch Normalization (BN) [18], and Momentum Contrast (MoCo) [19]. Following the generation of pseudo-labels by the teacher model, the student model's trainable weights W_s were updated via the backpropagation process. As illustrated in Equation (4), γ denotes the learning rate. The student model refines the teacher model's weights in accordance with the EMA (Exponential Moving Average) mechanism. The refinement process is detailed in Equation (5). Here, α signifies the hyperparameter that governs the rate at which the moving average combines the old and new weights.

$$W_s \leftarrow W_s + \gamma \frac{\partial L}{\partial W_s} \quad (4)$$

$$W_t \leftarrow \alpha W_t + (1 - \alpha) W_s \quad (5)$$

3.2. Combined Convolution

Feature extraction across various channels frequently results in redundant model computations and elevated memory consumption. In this paper, we present a straightforward combinatorial convolution within the backbone, aimed at reducing the model's computa-

tional complexity and memory traffic. The combined convolution module integrates partial convolution (PConv) and pointwise convolution (PWConv) to achieve efficient fusion of feature channel information.

PConv employs a filter that selectively operates on a subset of the input features, leaving the rest of the channels unaltered, as depicted in Figure 3c. During the process of accessing sequential memory, the initial or terminal contiguous channels within the input feature graph are chosen for computation, ensuring that the input and output feature maps maintain an identical number of channels. The computational complexity of PConv is quantified by FLOPs (floating point operations per second), denoted as $h \times w \times k^2 \times c_p^2$, where h denotes the height, w signifies the width, k represents the number of kernels, and c indicates the number of channels of the feature map, respectively. If the value of the partial ratio is set to $r = c_p/c = 1/4$, the required FLOPs amount to merely 1/16th of those needed for a full-channel convolution, thereby significantly enhancing computational efficiency. From Formula (6), it is evident that the module memory access is reduced to one-quarter of that required by traditional convolution.

$$h \times w \times 2c_p + k^2 + c_p^2 \approx h \times w \times 2c_p \quad (6)$$

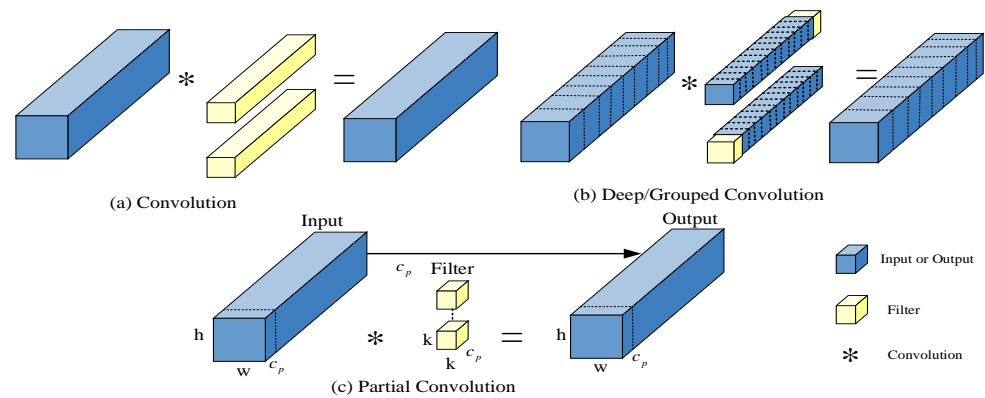


Figure 3. Schematic Diagram of the PConv Operation.

PWConv is capable of integrating channel information in a more comprehensive and efficient manner, and it synergizes with the depth dimensions of the feature maps produced by PConv. The fusion of PWConv with PConv yields a T-shaped convolutional effect on the input feature maps, as depicted in Figure 4.

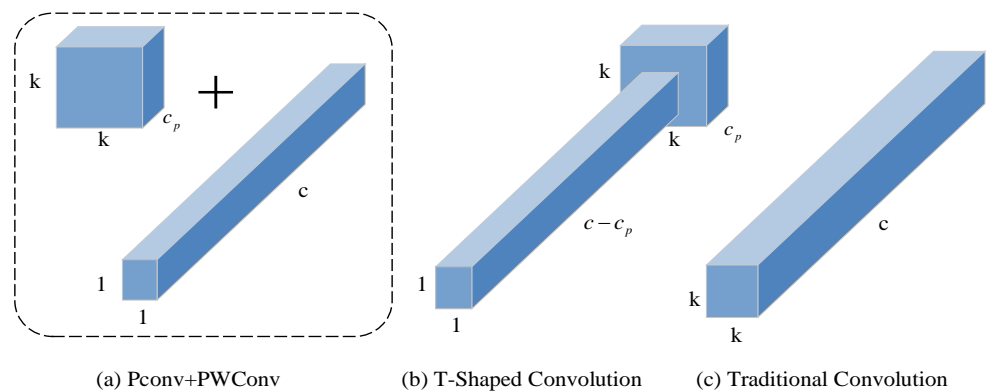


Figure 4. Convolutional structure comparison diagram.

To ascertain the efficacy of feature extraction and computational efficiency in traditional convolution, T convolution, and combined convolution, we quantified the importance of each position for all the filters in the pre-trained ResNet50 [8] model. The findings indicate that the significance of the central position's weight surpasses that of the surrounding

positions, aligning with the T-shaped structure's emphasis on the central area's information processing capabilities. Nevertheless, if the process can be broken down into two stages, the potential computational redundancy among filters can be substantially mitigated, thereby significantly decreasing the number of FLOPs needed for model computation without compromising the model's performance. Given the same input $I \in R^{c \times h \times w}$ and output $O \in R^{c \times h \times w}$, the FLOPs of the T-shaped convolution are computed to be $h \times w \times (k^2 \times c_p \times c + c \times (c - c_p))$, surpassing the FLOPs value $h \times w \times (k^2 \times c_p^2 + c^2)$ of the combined convolution. There exists a correlation between the number of parameters in both convolutions, denoted as $(k^2 - 1)c > k^2 c_p$. The conventional convolutional approach uniformly processes the entire feature region, failing to accentuate positions with more pronounced features. Consequently, the weights assigned are not meaningful, and the computational efficiency is inferior to that of T-shaped convolution and combined convolution.

3.3. Explicit Visual Center (EVC)

Typically, when the Feature Pyramid Network (FPN) tackles tasks involving dense prediction, it tends to concentrate on the interaction of features across different layers. However, this emphasis may result in a deficiency of detailed features within individual layers, potentially causing the network to overlook fine-grained feature information [3]. Given that deep features are abstractions of shallow specific features, we adopt a top-down strategy to globally and centrally modulate the extracted feature pyramid. During this process, the deep features derived from the explicit visual center information serve to constrain all the shallow features.

The EVC module primarily comprises two parallel-connected components, as illustrated in Figure 5. These components include a lightweight multilayer perceptron (MLP) and a learnable visual center mechanism. The lightweight MLP is designed to capture global long-distance dependencies, while the learnable visual center mechanism retains local critical regional information from the input image. This dual-component structure enables the EVC module to effectively acquire both global and local feature information, thereby enriching the visual center information for subsequent feature interactions. By optimizing the feature extraction pyramid, the EVC module enhances the efficiency of cross-layer feature interactions. Given that the EVC module can capture both global and local feature information, it significantly improves the model's ability to identify useful feature information in the presence of noisy data, thereby enhancing the model's robustness. During the feature extraction process, noisy data often introduce irrelevant or redundant information. By refining the feature extraction method, the EVC module minimizes the impact of such irrelevant or redundant information on model performance, thus improving the model's accuracy. Through training on noisy datasets and leveraging the optimization capabilities of the EVC module, the model can better learn the essential features of the data, thereby enhancing its generalization ability. Consequently, the model demonstrates improved predictive and classification performance when confronted with new, unseen data.

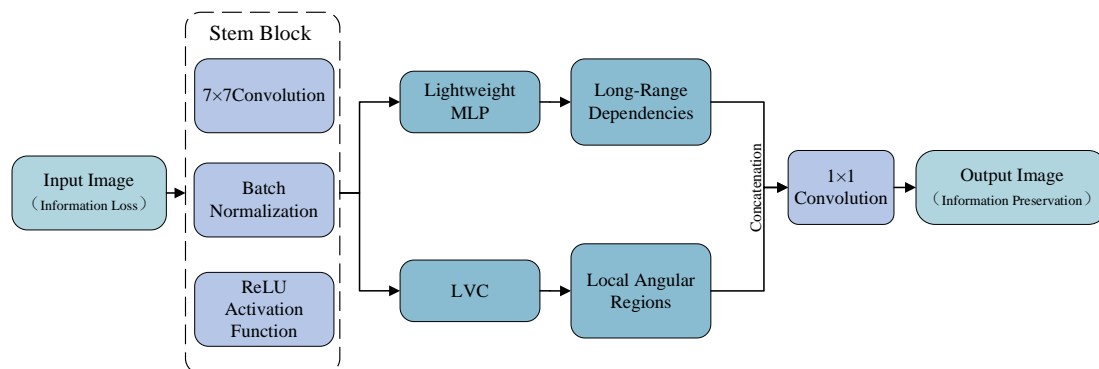


Figure 5. The EVC module.

The lightweight multi-layer perceptron (MLP) is responsible for capturing the global long-range dependencies within the deep feature X_4 , effectively extracting the global contextual information. To preserve the local angular region information, a Learnable Visual Center mechanism (LVC) is incorporated to aggregate the feature information from local regions at the same level of X_4 . The features are smoothed by a stem block architecture that encompasses a 7×7 convolutional layer with 256 output channels, a batch normalization layer, and an activation function layer. The output X_{in} of the stem block is determined using Formula (7), wherein $Conv_{7 \times 7}$ signifies the 7×7 convolution operation with a stride of 1, and the channel count is established at 256 [9] within the experimental setup. BN refers to the batch normalization layer, while σ denotes the ReLU (Rectified Linear Unit) activation function.

We concatenated the feature maps produced by the lightweight MLP and LVC modules along the channel dimension, employing them as the output for the EVC module's subsequent identification task. The splicing procedure is depicted in Equation (8). X denotes the output of the EVC module, while cat refers to the concatenation of feature graphs along the channel dimension. Furthermore, $MLP(X_{in})$ and $LVC(X_{in})$ symbolize the feature outputs subsequent to processing by the respective lightweight MLP and LVC modules.

$$X_{in} = \sigma(BN(Conv_{7 \times 7}(X_4))) \quad (7)$$

$$X = cat(MLP(X_{in}); LVC(X_{in})) \quad (8)$$

The Lightweight MLP (depicted in Figure 6), a pivotal element of the EVC module, comprises two residual modules: one based on deep convolution, namely the residual module [32], and the other, a channel MLP-based residual module. The output features from the former serve as the input [33] for the latter. These two residual modules improve the generalization ability of features by channel scaling [34] and regularized DropPath [35].

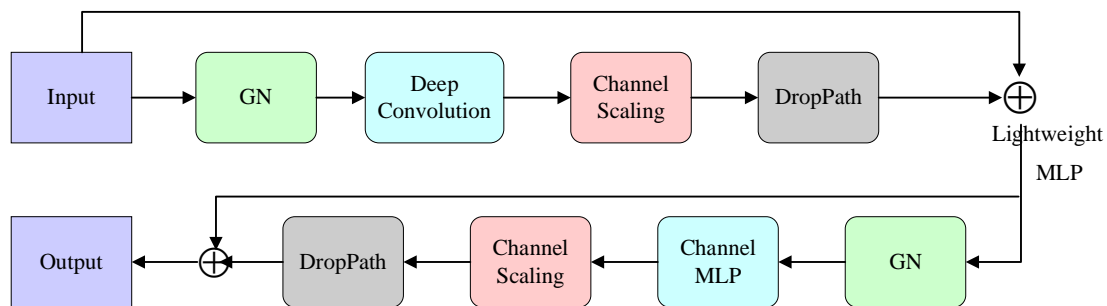


Figure 6. Lightweight MLP module.

Within the residual module that employs deep convolution, the feature X_{in} , produced by the stem block, is fed into the deep convolutional layer. Following group normalization, the feature map is segmented into several subgroups along the channel dimension for normalization computations. This process enhances the representational power of the features and significantly alleviates the computational load. The incorporation of a channel scaling module and a DropPath module can more effectively leverage computational resources, enhance model performance, mitigate the overfitting phenomenon, and bolster the model's generalization capabilities and robustness. Finally, the aforementioned operation is repeated, and the residual connection pertaining to the characteristic X_{in} is executed.

The output feature X_{in}^D , which is based on the deep convolution module, is depicted in Equation (9). Here, GN represents the group normalization operation, and $DConv$ denotes the deep convolution layer with a kernel size of 1×1 .

$$X_{in}^D = DConv(GN(X_{in})) + X_{in} \quad (9)$$

The output feature X_{in}^D from the deep convolution-based residual module is normalized and serves as the input for the channel MLP-based residual module. This module then performs the channel MLP operation on the normalized features, followed by channel scaling and regularization using DropPath. The processed features are subsequently combined with the original input feature X_{in}^D via a residual connection. The outcome is depicted in Equation (10), where CMLP represents the channel MLP operation. In comparison to the spatial MLP, which concentrates solely on the spatial dimension of images, the channel MLP not only significantly reduces computational complexity, but also satisfies the demands of a variety of intricate visual tasks.

$$MLP(X_{in}) = CMLP(GN(X_{in}^D)) + X_{in}^D \quad (10)$$

The LVC encoder module, as depicted in Figure 7, encompasses an internal dictionary and two principal components. These components are as follows: (1) The native code book $B = \{b_1, b_2, \dots, b_k\}$, which embodies a collection of predefined features, with the aggregate spatial dimensions of these features being $N = H \times W$, where H and W denote the height and width of the feature map, respectively. (2) The scale factor set $S = \{s_1, s_2, \dots, s_k\}$. The feature X_{in} output by the stem block is convolved with kernels of size 1×1 , 3×3 , and 1×1 , and the encoded result then undergoes deep feature extraction via a CBR (Convolutional-Batch Normalization-ReLU) block, which comprises a convolutional layer with 3×3 filters, a batch normalization layer, and a ReLU activation function. The extracted feature X_{in} is input into the intrinsic codebook for deep information interaction. By applying a series of predefined scale factors s , the components of feature X_i^D and feature b_k are mapped to their respective location data. The information encapsulated within the k -th visual code word of an image can be derived using Equation (11). Here, X_i^D represents the i -th pixel of the image, b_k denotes the k -th learnable visual code word, s_k signifies the k -th scaling factor, $X_i^D - b_k$ embodies the pertinent information between the pixel and its corresponding code word, and K corresponds to the aggregate count of visual centers.

$$e_k = \sum_{i=1}^N \frac{e^{-s_k \|x_i^D - b_k\|^2}}{\sum_{j=1}^K e^{-s_j \|x_i^D - b_j\|^2}} (x_i^D - b_k) \quad (11)$$

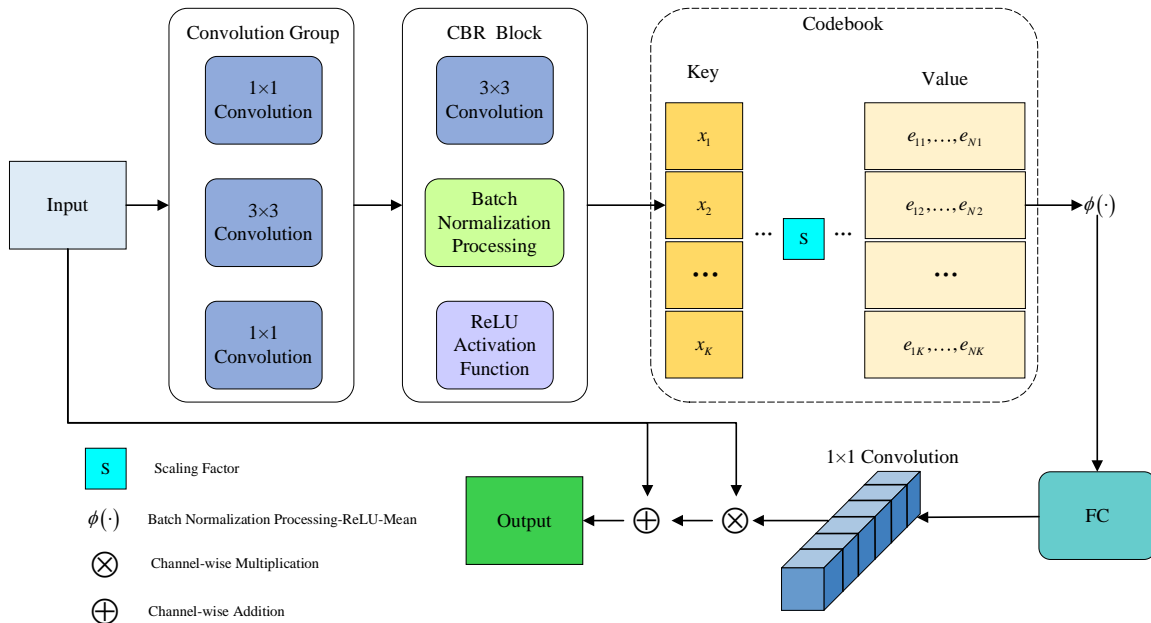


Figure 7. The LVC module.

The combination function Φ serves to integrate all the visual codeword information e_k , thereby generating comprehensive information e , as illustrated in Equation (12).

To further pinpoint and refine the key features of the category, the data are transmitted to a fully connected and 1×1 layer convolutional network. The scaling factor coefficient δ is applied to each channel of the feature X_{in} to obtain the local angular region feature Z , as illustrated in Equation (13). $Conv_{1 \times 1}$ denotes the 1×1 layer convolutional component, and δ signifies the Sigmoid activation function, while \otimes corresponds to the channel-wise multiplication process. By summing Z and X_{in} , we derive the output characteristics of the LVC module, as illustrated in Equation (14). Furthermore, \oplus signifies the channel-wise addition operation.

$$e = \sum_{k=1}^K \phi(e_k) \quad (12)$$

$$Z = X_{in} \otimes (\delta(Conv_{1 \times 1}(e))) \quad (13)$$

$$LVC(X_{in}) = X_{in} \oplus Z \quad (14)$$

3.4. Enhanced Spatial Grouping

To alleviate the impact of background noise on target detection efficacy, we performed lightweight SGE processing in the RPN. The SGE module finely tunes the significance of each subfeature by generating tailored attention factors for every spatial location within each semantic group. This empowers each group to independently bolster its acquired representations while mitigating potential noise. In doing so, SGE leverages the similarity between global and local feature descriptors to steer the attention factors, thereby optimizing the distribution of feature weights. During implementation, the SGE module encounters hurdles, including ensuring the precision and efficiency of attention factor generation, as well as attaining substantial performance gains without escalating the number of parameters or computational load. Despite these challenges, the SGE module has demonstrated remarkable improvements in both image classification and object detection tasks.

To adaptively modify the feature importance across various spatial locations within each semantic grouping, unique attention weights were assigned to each location across all the groups. This process enhances the model's ability to learn expressions by reinforcing critical subfeatures, while simultaneously suppressing potential noise that could impact performance [36], as depicted in Figure 8. This process necessitates minimal additional parameters and does not involve complex computational procedures. Instead, it operates by processing subfeatures in parallel, leveraging the similarity between global statistical features and local positional features to guide attention. Consequently, this approach enhances the features and yields a semantic feature representation that is evenly distributed across the spatial domain.

The input set of convolutional feature maps, characterized by a channel size of C and a dimension size of $H \times W$, is segmented into G groups along the channel dimension. The spatial position of each cluster corresponds to a vector, $X = \{x_1, x_2, \dots, x_m\}$, $x_i \in RC/G$, $m = H \times W$. Given that the convolutional layer is prone to noise interference during the processing of input images, the similarity among image features results in an uneven feature response. To enhance the model's capacity for learning and recognizing semantic features within critical areas, this paper employs the global contextual information of feature grouping to construct an approximate vector value that mirrors the semantic representation of the feature group. This approach aims to extract semantic features that are more balanced and discriminative. The spatial average aggregation function, denoted as F_{gp} , which represents the global statistical feature, is illustrated in Equation (15).

$$g = F_{gp}(X) = \frac{1}{m} \sum_{i=1}^m x_i \quad (15)$$

Utilizing the aforementioned global characteristics, significance weights for each local feature can be derived through the dot product, as illustrated in Equation (16).

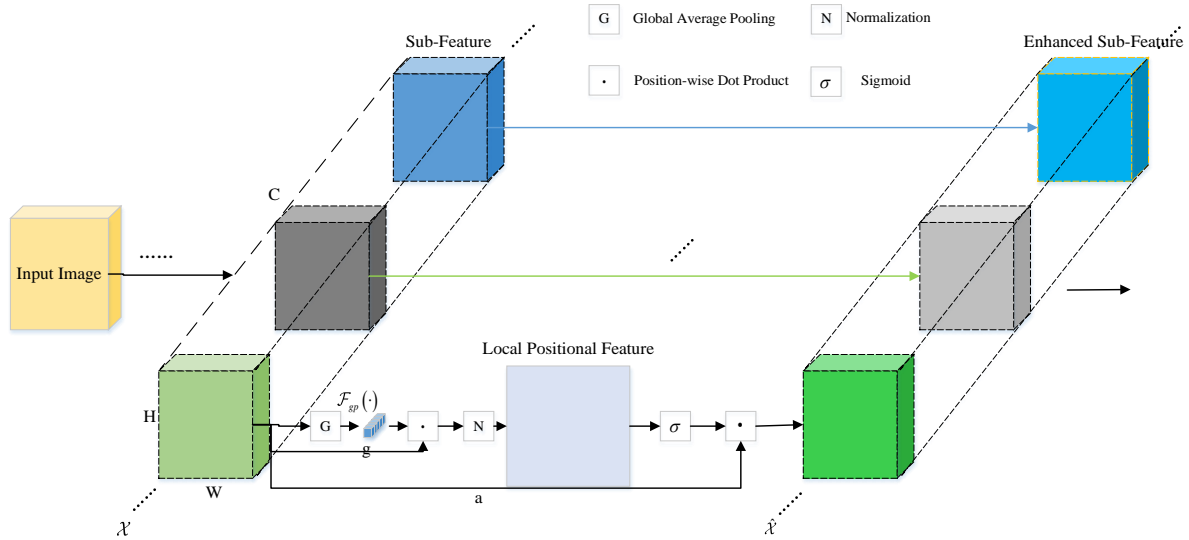


Figure 8. Lightweight SGE module diagram.

The dot product serves as a measure of similarity between the global semantic feature g and the local features x_i . Additionally, the concept of c_i can be extended to $\|g\| \|x_i\| \cos(\theta_i)$, and θ_i represents the angle between features g and x_i . When the local eigenvector is substantial and its orientation aligns with g , the initial weight value is likely to increase. Concurrently, the reinforcement mechanism for key regional features should also be substantiated.

$$c_i = g \cdot x_i = \|g\| \|x_i\| \cos(\theta_i) \quad (16)$$

To eliminate amplitude discrepancies among coefficients from various samples and ensure balanced comparisons across spatial dimensions, c is normalized using Equation (17), where ε represents an additional, small positive value introduced to ensure numerical stability. When integrating the network architecture, to achieve an approximate identity transformation of the normalized layer, the learnable zoom parameter γ and offset parameter β are introduced for each weight coefficient c'_i , and the importance coefficient α_i is then assigned to it, as illustrated in Equation (18). Within the entire SGE module, if there are only two additional parameters, their number can align with the number of groups. Subsequently, the sigmoid function σ , typically assuming small integer values like 32 or 64, is applied to the coefficient α_i . This process scales the original feature vector x_i in the spatial dimension, resulting in the enhanced subfeature vector x'_i and the enhancer feature group X' , as illustrated in Equations (19) and (20).

$$c'_i = \frac{c_i - \mu_c}{\sigma_c + \varepsilon}, \mu_c = \sum_j^m c_j, \sigma_c^2 = \frac{1}{m} \sum_j^m (c_j - \mu_c)^2 \quad (17)$$

$$\alpha_i = \gamma c'_i + \beta \quad (18)$$

$$x'_i = x_i \cdot \sigma(\alpha_i) \quad (19)$$

$$X' \equiv \{x'_{1...m}\}, x'_i \in R^{\frac{C}{c}}, m = H \times W \quad (20)$$

4. Experimental Analysis

4.1. Experimental Dataset

We selected the most commonly utilized dataset in the Few-Shot Object Detection (FSOD) research field as the training set and validation dataset of this method, to ensure the diversity and integrity of data parameters and to verify the advantages of this method. The experimental datasets presented in this paper comprise MS COCO as the source domain, with ArTaxOr, UODD, and DIOR serving as the target domains. We selected three major categories and 30 subcategories, with a total of 35,648 images for the training set and 6889 images for the test set. The datasets for these three target domains are detailed in Table 1.

Table 1. Target domain datasets.

Dataset	Data Field	Category Number	Training Images	Test Pattern
ArTaxOr [10]	Biology	7	13,991	1383
UODD [11]	Underwater	3	3194	506
DIOR [12]	Aviation	20	18,463	5000

The MS COCO dataset [9] features instances of everyday objects situated within their natural surroundings, comprising 80 distinct categories. Specifically, the dataset is partitioned into a training set with 118,000 images, a validation set with 5000 images, and a test set with 20,000 images. Among these, 20 categories that coincided with those in the PASCAL VOC dataset were designated as new classes, while the remaining 60 categories served as the foundational classes.

The ArTaxOr dataset [10] encompasses 1.3 million distinct arthropod classes, including centipedes, millipedes, and isopods, each featuring a variety of species that differ in size, shape, and coloration. Each species is represented by a minimum of 2000 individual specimens, and each image within the dataset depicts between 1 and 50 specimens.

The UODD dataset [11] comprises samples of sea urchins, sea cucumbers, and scallops captured in various underwater scenarios with differing levels of light scattering. The collection process took into account factors such as shooting angle, scene complexity, scene contrast, and target size. The dataset includes approximately 19,000 objects, represented by 3194 images. The distribution of these images into training, validation, and test sets was as follows: 2688 for training, 128 for validation, and 506 for testing.

The DIOR dataset [12] encompasses 20 common categories of infrastructure. It comprises a collection of 23,463 remote sensing images, captured across various weather conditions, seasons, and imaging scenarios, spanning over 80 countries. These images account for a total of 192,472 annotated object instances. Within each category, objects exhibit variations in color, size, and scale, while there is a significant semantic overlap among fine-grained objects across different classes.

4.2. Setting of the Experimental Parameters

The experimental environment was the NVIDIA GeForce RTX 3090 GPU and the model framework was the standard Detectron2 [24].

In the context of Cross-Domain Few-shot Object Detection, the configuration of hyperparameters and training specifics is pivotal for optimizing model performance. An initial learning rate of 0.001 was established, with an exponential decay strategy employed to facilitate both the convergence and enhancement of the model's performance. The Adam optimizer, widely recognized for its efficacy in FSOD, was selected to dynamically adjust the learning rate throughout the training process, thereby expediting the model's convergence. Considering the scale of the dataset, a batch size of 32 was determined to minimize computational expenses. For the creation of both the student and teacher

models, we employed a Faster R-CNN, Feature Pyramid Network (FPN), and ResNet50. The confidence threshold was established at $\delta = 0.7$, and the EMA ratio α was set to 0.999. Table 2 presents the maximum number of training iterations for the three datasets within the target domains, with sample sizes of $K = 1$, $K = 5$, and $K = 10$. The evaluation metric employed is the mean Average Precision (mAP).

Table 2. Maximum iteration number setting.

Dataset/Sample Number	1	5	10
ArTaxOr [10]	5000	10,000	12,500
UODD [11]	10,000	15,000	15,000
DIOR [12]	8000	12,500	15,000

4.3. Experiment Comparison

To ascertain the efficacy of this approach, we selected ten models for few-sample target detection, including CD-FSOD [8], for comparison purposes, including: A-RPN [37], H-GCN [22], Meta R-CNN [24], TFA [38], FSCE [39], DeFRCN [40], FRCN-ft, Detic-FT [41], and GOAT [42]. In particular, FRCN-ft undergoes pre-training using the CD-FSOD foundational dataset, followed by fine-tuning on K target instance samples. For the comparative experiments, the sample sizes across various orders of magnitude [8] are illustrated, represented as 1, 5, and 10. The outcomes are presented in Table 3, featuring bold and underlined optimal and suboptimal values. The experimental outcomes demonstrate that the method exhibited substantial benefits subsequent to the integration of the combined convolutional module, the explicit visual center module, and the spatial grouping enhancement module. It outperformed alternative methods across the three experimental datasets within the target domain.

Table 3. Contrasting experimental results on the three target domains.

Method/ Sample Number	ArTaxOr			UODD			DIOR		
	1	5	10	1	5	10	1	5	10
A-RPN [37]	2.5	8.1	13.9	3.3	8.4	10.8	7.5	17.1	20.3
Meta R-CNN [24]	2.8	8.5	10	3.6	8.8	11.2	7.8	17.7	20.6
H-GCN [22]	2.6	8.2	12	3.8	7.7	11.0	7.9	18.0	20.9
TFA w/cos [38]	3.1	8.8	18	4	8.7	11.8	8.0	18.1	20.5
FSCE [39]	3.7	10.2	15.9	3.9	9.6	12.0	8.6	18.7	21.9
DeFRCN [40]	3.6	9.9	15.5	5	9.9	12.1	9.3	18.9	22.9
FRCN-ft [2]	3.4	9.3	15.2	1	9.2	12.3	8.4	18.3	21.2
CD-FSOD [8]	5.1	12.5	18.1	5.9	12.2	15	10.5	19.1	26.5
Detic-FT [41]	3.2	8.7	12.0	2	10.4	14	1	12.1	15.4
GOAT [42]	5.7	11.1	21.2	3.5	9.5	16.6	11.3	19.8	25.4
Ours	7.2	13.3	19.2	6.2	12.7	15.3	12.1	20.0	27.7

The experimental results show that the detection accuracy increases by 1.67% and 1.87% compared with the best GOAT [42] method for sample sizes 1 and 5. Since the method mainly monitors small samples, the detection effect does not improve much when the sample size is 10. This also illustrates the advantages of this method in small-sample monitoring.

The prediction outcomes of the current approach, when compared with CD-FSOD, GOAT, and TFA, are depicted in Figure 9. It is evident from the figure that the current approach surpasses the other contrastive models in terms of both classification accuracy and localization accuracy. When the image encompasses multiple objects (as seen in the first line of Figure 9) or multiple objects of the same type (as depicted in the third line of Figure 9), this method can accurately detect all the objects within the image. The experimental results

prove that the method has satisfactory results for the detection of target images. In contrast, TFA fails to detect any objects, while CD-FSOD and GOAT can only identify some of the objects, resulting in a certain degree of omission. For the objects that are detected, the boundary box positioning accuracy and recognition confidence of these two methods are inferior to those of the present method. When the background noise within the image data is excessively high (as seen in the second row of Figure 9), the positioning accuracy of this method is markedly superior to that of the other methods. The converse is true when the significance of features within the image's object apart from the color are considered (refer to the fourth row of Figure 9).

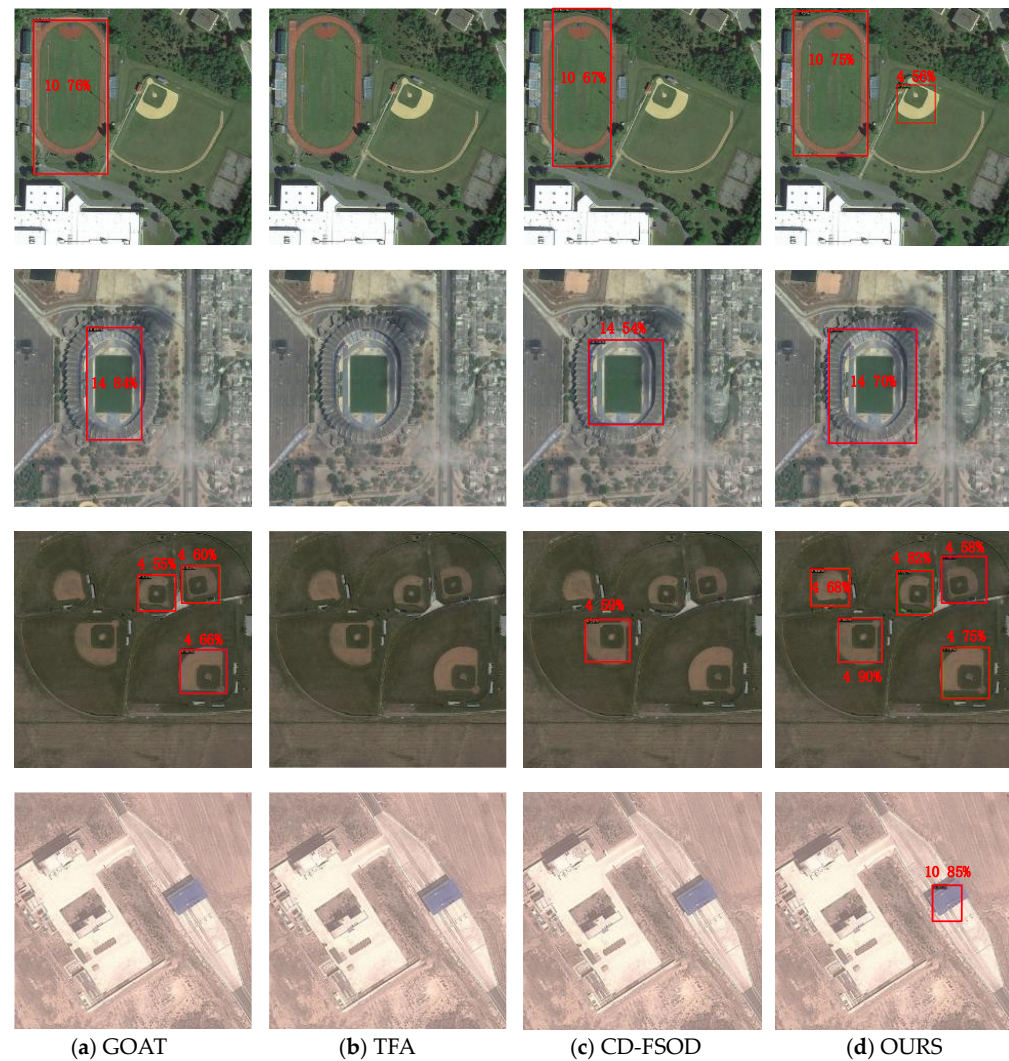


Figure 9. Prediction results.

The outcomes of the current methodology when applied to the datasets of the target domains, namely ArTaxOr, UODD, and DIOR, are depicted in Figure 10, Figure 11 and Figure 12, respectively.

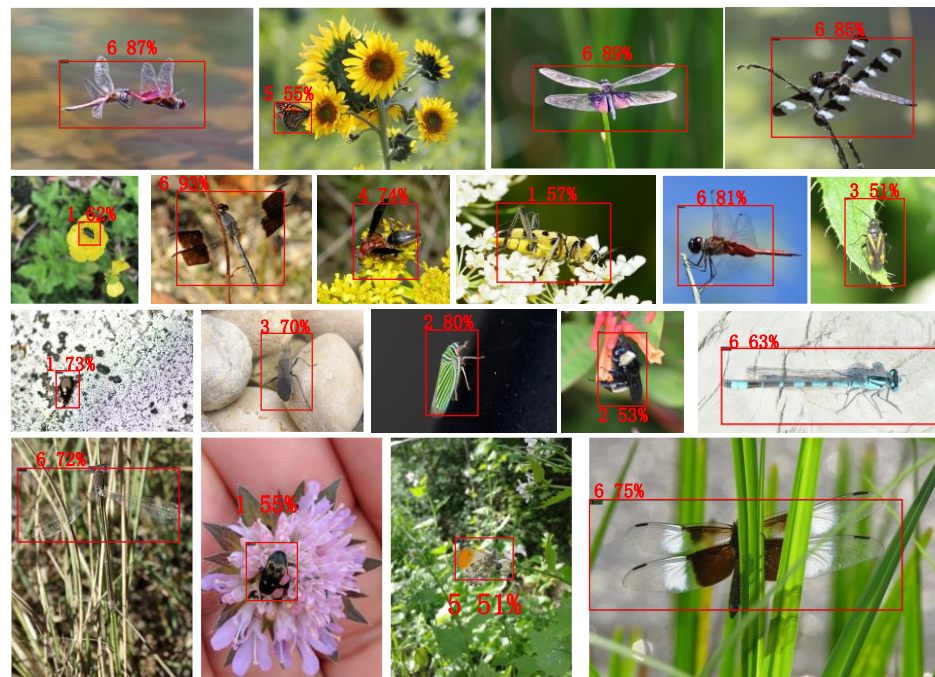


Figure 10. ArTaxOr dataset prediction results.

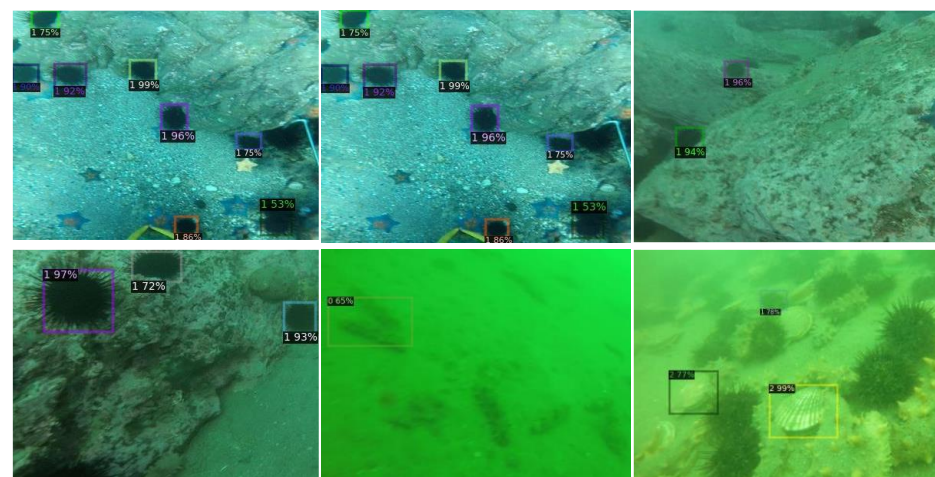


Figure 11. Results for the UODD dataset.

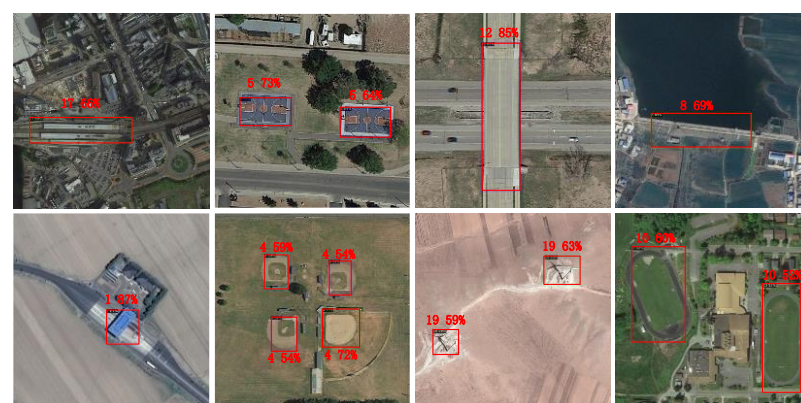


Figure 12. Cont.



Figure 12. DIOR dataset prediction results.

4.4. Ablation Experiments

The EVC (Enhanced Visual Context) module has emerged as the most influential component in ablation studies due to its unique advantages. It captures global long-range dependencies through an explicit visual center mechanism and integrates local corner region information, achieving effective fusion of global and local information. Simultaneously, it utilizes a global centralized regulation method to adjust shallow features, significantly enhancing the representation capability of the feature pyramid. These characteristics enable the EVC module to excel in complex scenarios and under varying lighting conditions, particularly in the detection of small objects. Therefore, compared to methods such as CConv and SGE, the EVC module demonstrates significant advantages in improving detection accuracy, providing strong support for high-precision object detection applications.

EMA (Exponential Moving Average) technology is widely employed in optimizing the parameters of the teacher model within the teacher–student model framework, owing to its effectiveness in smoothing the parameter update process, minimizing performance fluctuations, and swiftly adapting to data variations. By bolstering the stability of the teacher model, EMA technology ensures more dependable output predictions, which, in turn, optimizes the knowledge distillation process. This gradual evolution of the teacher model fosters improved learning and mimicry by the student model, mitigates the risk of overfitting, and ultimately results in substantial performance enhancements.

We performed ablation experiments to ascertain the necessity of the EMA strategy, including the distillation stage, CConv, EVC, and SGE core component modules within this approach. The outcomes of the ablation studies for the EMA strategy and distillation, with sample sizes of 1, 5, and 10, are presented in Table 4. Here, ‘S’ denotes the student model, while ‘T’ represents the teacher model. As illustrated in the table, both the Exponential Moving Average (EMA) strategy and distillation enhance model performance. Notably, the EMA strategy demonstrates superior efficacy in performance enhancement compared to the distillation method, and it can additionally refine the model’s capabilities. As the teacher model’s performance is enhanced, the student model’s performance correspondingly improves, thoroughly demonstrating that a mutually beneficial learning relationship can be established between the student and teacher models. Moreover, the implementation of EMA and knowledge distillation techniques can effectively mitigate the issue of overfitting in FSOD. Upon employing EMA, the teacher model emerges as a refined amalgamation of the weights from various training phases of the student model, thereby enhancing the overall stability of the model. The distillation loss can be viewed as a form of regularization, which fortifies the student model’s learning process and enhances its capacity to generalize to previously unseen samples.

Table 4. mAP effects of EMA and distillation.

EMA	Distillation	1		5		10	
		S	T	S	T	S	T
Yes	No	8.9	9.7	18.7	19.2	21.6	25.4
No	Yes	9.0	9.6	18.9	19.1	22.9	23.7
Yes	Yes	10.5	12.1	19.1	20.0	25	27.7

We also conducted experiments using various EMA ratios α , with specific values of 0.5, 0.7, 0.9, 0.999, and 0.9999. The corresponding mAP results are presented in Table 5. When a smaller ratio is employed, such as $\alpha = 0.5$, the student model exerts a significant influence on the teacher model during each iteration, resulting in a low and unstable mAP for the teacher model. The instability of the ratio is mitigated and enhanced as the EMA ratio α increases. This occurs because at higher EMA ratios, the teacher model's weights tend to favor smoother historical averages, thereby enhancing the model's stability. The model achieved its peak mAP when the EMA ratio α was set to 0.999. Nevertheless, as the EMA ratio α ascends further, the efficacy of the teacher model diminishes. This phenomenon arises because the teacher model's weight predominantly depends on the weights from a remote past time window, rather than being influenced by the current learning outcomes of the student model.

Table 5. Changes in EMA ratios and mAP under different samples.

Sample Number/ α	0.5	0.7	0.9	0.999	0.9999
1	6.7	8.2	9.9	12.1	8.6
5	15.5	17.0	18.9	20.0	19.4
10	22.3	27	25.4	27.7	26.8

The pseudo-label threshold serves to filter out prediction bounding boxes with low confidence levels, playing a pivotal role in optimizing distillation loss. The selection of an appropriate pseudo-label threshold is essential to ensure that the model strikes a balance between effective prediction and sufficient training. The mAP values at the threshold levels of 0.6, 0.7, 0.8, and 0.9 are presented in Table 6. It can be observed that with a lower threshold, the model tends to learn from less reliable bounding boxes, which consequently leads to a diminished mAP for the model. On the contrary, if the threshold is set excessively high, the enhancement of model performance is constrained, because it becomes impossible to use a sufficient quantity of bounding boxes for training purposes. In this paper, the highest mAP threshold was established at 0.7, a value that was consistently applied across all the experiments.

Table 6. Pseudo-tag thresholds and mAP for different samples.

Sample Number/Threshold	0.6	0.7	0.8	0.9
1	9.8	12.1	10.0	10.3
5	17.9	20.0	18.5	18.2
10	25.7	27.7	22	21

Table 7 showcases the efficacy of each component of the method proposed in this paper. It is evident that the ECA module plays a more significant role in enhancing model performance compared to both the CConv and SGE modules. The EVC module is capable of capturing the long-range dependencies of features, aggregating the local key regions of the input image, and extracting more comprehensive features. This significantly enhances the model's detection accuracy.

Table 7. Validation of the different components.

CConv	EVC	SGE	mAP		
			1	5	10
No	No	No	10.5	19.1	26.5
Yes	No	No	10.7	19.4	26.8
No	Yes	No	11.5	19.7	27.3
Yes	Yes	No	11.7	19.9	27.5
No	No	Yes	11.3	19.8	27.0
Yes	Yes	Yes	12.1	20.0	27.7

4.5. Discussion

Upon comparing the experimental results, it is apparent that our method demonstrates superior performance relative to other methods across all three experimental datasets. It is worth noting that, specifically when the number of samples K is 1 and 5, our method achieves improvements in detection accuracy of 1.67% and 1.87%, respectively, compared to the state-of-the-art method. Ablation experiments demonstrated that employing CConv to diminish model redundancy and memory access, while simultaneously lowering the high feature similarity in the channel dimension, enhances the model's detection performance. The SGE module is capable of significantly mitigating the impact of noise on images, thereby further enhancing the model's detection performance. When the background information within an image closely resembles the target information, pixel-level EVC (Edge-Enhanced Visual Computing) can significantly enhance the feature representation of local critical regions.

5. Conclusions

In this work, we introduce a supervised and distillation-driven approach for detecting targets with limited samples, addressing the challenges of weak correlation and substantial data discrepancies in cross-domain scenarios. Initially, during the feature extraction phase, the utilization of combined convolution serves to diminish the computational redundancy within the feature channel dimension, thereby effectively curbing the model's memory access requirements. In the construction of the feature pyramid, an explicit visual center is employed to capture and integrate long-range dependencies among features of varying scales. This process enhances the feature representation of local angular regions, allowing the model to achieve a comprehensive and discriminative feature representation. In the end, the generation of potential areas through the use of combined features allows for a significant reduction in the influence of data noise on detection outcomes, thanks to the enhancement provided by spatial grouping, which leads to more accurate localization of targets. Through comprehensive ablation experiments and qualitative analysis, we proved that the proposed method effectively improves the detection accuracy of the model with data from different fields. Certainly, the network model presented in this paper requires enhancement in numerous areas. The subsequent research step will concentrate on how to further streamline the model and diminish its complexity. Furthermore, our approach will be expanded to encompass large-scale outdoor environmental monitoring. Given that real-world scenes often contain objects with complex background environments, and these objects can be highly confounded with the background information, a key focus will be on how to mitigate the impact of background noise and enhance the accuracy of cross-domain target detection.

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

Funding: This research was funded by the National Natural Science Foundation of China, grant number 62272426; Shanxi Province's Major Science and Technology Special Program 'Unveiling the List and Leading the Way' Project, grant number 202201150401021; the National Natural Science Foundation of Shanxi, under grant number 202303021211153 and 202303021212372; and the Foundation of Shanxi Key Laboratory of Machine Vision and Virtual Reality, grant number 447-110103.

Data Availability Statement: The MS COCO dataset presented in this study is openly available on the website. Available online: <http://cocodataset.org/#home>, accessed on 19 July 2023. The ArTaxOr dataset presented in this study is openly available on the website. Available online: <https://www.kaggle.com/datasets/mistag/arthropod-taxonomy-orders-object-detection-dataset>, accessed on 19 July 2023. The UODD dataset presented in this study is openly available on the website. Available online: <https://github.com/LehiChiang/Underwater-object-detectiondataset>, accessed on 19 July 2023. The DIOR dataset presented in this study is openly available on the website. Available online: <http://www.escience.cn/people/gongcheng/DIOR.html>, accessed on 19 July 2023.

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

References

1. Zhang, B.; Chen, T.; Wang, B.; Li, R. Joint distribution alignment via adversarial learning for domain adaptive object detection. *IEEE Trans. Multimed.* **2021**, *24*, 4102–4112. [CrossRef]
2. Everingham, M.R.; Eslami, S.M.A.; Gool, L.J.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes Challenge. *Int. J. Comput. Vis.* **2015**, *11*, 98–136. [CrossRef]
3. Inoue, N.; Furuta, R.; Yamasaki, T.; Aizawa, K. Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018. [CrossRef]
4. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
5. Sakaridis, C.; Dai, D.; Van Gool, L. Semantic Foggy Scene Understanding with Synthetic Data. *Int. J. Comput. Vis.* **2018**, *126*, 973–992. [CrossRef]
6. Gao, Y.; Yang, L.; Huang, Y.; Xie, S.; Li, S.; Zheng, W. croFOD: An Adaptive Method for Cross-Domain Few-Shot Object Detection. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022. [CrossRef]
7. Gao, Y.; Lin, K.-Y.; Yan, J.; Wang, Y.; Zheng, W.-S. AsyFOD: An Asymmetric Adaptation Paradigm for Few-Shot Domain Adaptive Object Detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
8. Xiong, W. CD-FSOD: A Benchmark for Cross-domain Few-shot Object Detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023. [CrossRef]
9. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Proceedings, Part V 13; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
10. Drange, G. Arthropod Taxonomy Orders Object Detection Dataset. 2019. Available online: <https://www.kaggle.com/datasets/mistag/arthropod-taxonomy-orders-object-detection-dataset> (accessed on 19 July 2023).
11. Jiang, L.; Wang, Y.; Jia, Q.; Xu, S.; Liu, Y.; Fan, X.; Li, H.; Liu, R.; Xue, X.; Wang, X. Underwater Species Detection using Channel Sharpening Attention. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021.
12. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
13. 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, Glasgow, UK, 23–28 August 2020; pp. 456–472.
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
15. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
16. Hua, J.; Liu, X.; Zhao, Y. Target detection of target detection based on feature fusion. *Comput. Sci.* **2023**, *50*, 209–213.
17. Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; Darrell, T. Few-Shot Object Detection via Feature Reweighting. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV'19), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8419–8428.
18. Wang, Y.X.; Ramanan, D.; Hebert, M. Meta-Learning to Detect Rare Objects. In Proceedings of the International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019. [CrossRef]

19. Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; Lin, L. Meta R-CNN: Towards General Solver for Instancelevel Few-shot Learning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019. [\[CrossRef\]](#)
20. Xiao, Y.; Lepetit, V.; Marlet, R. Few-shot Object Detection and Viewpoint Estimation for Objects in the Wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 3090–3106. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Karlinsky, L.; Shtok, J.; Harary, S.; Schwartz, E.; Aides, A.; Feris, R.; Giryes, R.; Bronstein, A.M. RepMet: Representative-based metric learning for classification and one-shot object detection. *arXiv* **2018**. [\[CrossRef\]](#)
22. Fan, Q.; Zhuo, W.; Tang, C.K.; Tai, Y.-W. Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector. *arXiv* **2019**. [\[CrossRef\]](#)
23. Li, B.; Yang, B.; Liu, C.; Liu, F.; Ji, R.; Ye, Q. Beyond Max-Margin: Class Margin Equilibrium for Few-shot Object Detection. *arXiv* **2021**. [\[CrossRef\]](#)
24. Han, G.; He, Y.; Huang, S.; Ma, J.; Chang, S.-F. Query Adaptive Few-Shot Object Detection with Heterogeneous Graph Convolutional Networks. *arXiv* **2021**. [\[CrossRef\]](#)
25. Chen, H.; Wang, Y.; Wang, G.; Qiao, Y. LSTD: A Low-Shot Transfer Detector for Object Detection. *arXiv* **2018**. [\[CrossRef\]](#)
26. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot multibox Detector. In *Computer Vision—ECCV 2016: Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part I 14; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
27. Wu, A.; Han, Y.; Zhu, L.; Yang, Y. Universal-Prototype Augmentation for Few-Shot Object Detection. *arXiv* **2021**. [\[CrossRef\]](#)
28. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
29. Chen, J.; Kao, S.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; Chan, S.-H.G. Run, Don't walk: Chasing higher FLOPS for faster neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031.
30. Quan, Y.; Zhang, D.; Zhang, L.; Tang, J. Centralized feature pyramid for object detection. *IEEE Trans. Image Process.* **2023**, *32*, 4341–4354. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Wang, J.; Ju, N.; Tie, Y.; Bai, Y.; Ge, H. A framework for identifying the onset of landslide acceleration based on the exponential moving average (EMA). *J. Mt. Sci.* **2023**, *20*, 1639–1649. [\[CrossRef\]](#)
32. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9729–9738.
33. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:17004861.
34. Tolstikhin, I.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. Mlp-mixer: An all-mlp architecture for vision. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24261–24272.
35. Yu, W.; Luo, M.; Zhou, P.; Si, C.; Zhou, Y.; Wang, X.; Feng, J.; Yan, S. Metaformer is actually what you need for vision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10819–10829.
36. Li, X.; Hu, X.; Yang, J. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. *arXiv* **2019**, arXiv:1905.09646.
37. Abhishek AV, S.; Kotni, S. Detectron2 object detection & manipulating images using cartoonization. *Int. J. Eng. Res. Technol. (IJERT)* **2021**, *10*, 1–5.
38. Han, G.; Huang, S.; Ma, J.; He, Y.; Chang, S.-F. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 780–789. [\[CrossRef\]](#)
39. Wang, X.; Huang, T.E.; Darrell, T.; Gonzalez, J.E.; Yu, F. Frustratingly Simple Few-Shot Object Detection. *arXiv* **2020**. [\[CrossRef\]](#)
40. Sun, B.; Li, B.; Cai, S.; Yuan, Y.; Zhang, C. FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding. *arXiv* **2021**. [\[CrossRef\]](#)
41. Qiao, L.; Zhao, Y.; Li, Z.; Qiu, X.; Wu, J.; Zhang, C. DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection. *arXiv* **2021**. [\[CrossRef\]](#)
42. Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; Misra, I. Detecting twenty-thousand classes using image-level supervision. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 350–368.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.