electronics

MDPI

Article

Few-Shot Object Detection with Local Feature Enhancement and Feature Interrelation

Hefeng Lai and Peng Zhang *

School of Electronics and Communication Engineering, Shenzhen Campus of Sun Yat-Sen University, Shenzhen 518107, China; laihf3@mail2.sysu.edu.cn

* Correspondence: zhangpeng5@mail.sysu.edu.cn

Abstract: Few-shot object detection (FSOD) aims at designing models that can accurately detect targets of novel classes in a scarce data regime. Existing research has improved detection performance with meta-learning-based models. However, existing methods continue to exhibit certain imperfections: (1) Only the interacting global features of query and support images lead to ignoring local critical features in the imprecise localization of objects from new categories. (2) Convolutional neural networks (CNNs) encounter difficulty in learning diverse pose features from exceedingly limited labeled samples of unseen classes. (3) Local context information is not fully utilized in a global attention mechanism, which means the attention modules need to be improved. As a result, the detection performance of novel-class objects is compromised. To overcome these challenges, a few-shot object detection network is proposed with a local feature enhancement module and an intrinsic feature transformation module. In this paper, a local feature enhancement module (LFEM) is designed to raise the importance of intrinsic features of the novel-class samples. In addition, an Intrinsic Feature Transform Module (IFTM) is explored to enhance the feature representation of novel-class samples, which enriches the feature space of novel classes. Finally, a more effective cross-attention module, called Global Cross-Attention Network (GCAN), which fully aggregates local and global context information between query and support images, is proposed in this paper. The crucial features of novel-class objects are extracted effectively by our model before the feature fusion between query images and support images. Our proposed method increases, on average, the detection performance by 0.93 (nAP) in comparison with previous models on the PASCAL VOC FSOD benchmark dataset. Extensive experiments demonstrate the effectiveness of our modules under various experimental settings.

Keywords: few-shot object detection; object detection; local feature enhancement; cross-attention mechanism

Citation: Lai, H.; Zhang, P. Few-Shot Object Detection with Local Feature Enhancement and Feature Interrelation. *Electronics* **2023**, *12*, 4036. https://doi.org/10.3390/electronics12194036

Academic Editor: Silvia Liberata Ullo

Received: 19 August 2023
Revised: 19 September 2023
Accepted: 20 September 2023
Published: 25 September 2023

1. Introduction

Object detection (OD) via deep learning approaches [1–5] in computer vision has experienced tremendous progress in recent years. However, existing object detection models rely heavily on a substantial amount of annotated data and require long training times to achieve exceptional performance. Demonstrating good performance with a limited number of annotated samples is challenging. Object detection via few-shot learning methods, called few-shot object detection (FSOD), is a promising research branch to overcome data scarcity.

Few-shot learning [6–10] aims at designing models that can successfully operate in a limited data regime. By leveraging few-shot learning, significant achievements have been made in few-shot classification (FSC) tasks [11–14]. FSOD is a more challenging research area that requires the simultaneous accomplishment of both novel object classification and localization. Recently, most FSOD research has focused on meta-learning approaches [15–17], which leverage support images to guide the detector in the classification and localization of novel-class objects. One of the crucial research branches in

meta-learning-based FSOD is how to finish the aggregation of features from support images and query images effectively.

Existing meta-learning-based FSOD methods [18–23] aggregate query features and support features, which are generated by a feature extractor called the backbone. However, it is worth noting that the features extracted by the backbone network are coarse and may not highlight the key feature information of the samples. In other words, these methods do not effectively utilize precise features to accomplish feature aggregation, which leads to unsatisfactory detection performance. Training a model with a limited number of annotated samples means that the model must quickly focus on the recognizable feature representations of novel-class objects. How to highlight the key feature representations of objects (particularly the objects from different classes that have a high degree of similarity) becomes one of the challenges in FSOD. In fact, several feature processing methods proposed in object detection or image segmentation address the problem of feature enhancement from various perspectives. SENet [24] leverages the learning of channel weights to determine the importance of each channel and performs a weighted average of multiple feature maps during the fusion process, taking into account the channel weights. GCNet [25] utilizes a self-attention mechanism to integrate and capture global contextual information and local features during the fusion of multiple feature maps. DANet [26], based on both channel and spatial attention mechanisms, leverages the channel weights and spatial weights to determine the importance of each channel and each spatial position. Inspired by the above methods, a local feature enhancement module (LFEM) is designed in this work, which utilizes different-sized convolution kernels to extract local feature information and calculate the spatial weights to highlight the crucial regions.

A limited number of annotated samples in a novel-class dataset implies that the dataset contains only a finite amount of feature information of objects from novel classes. Traditional object detection methods are prone to overfitting and poor generalization in data scarcity situations. The limited availability of annotated samples hampers the model's ability to learn robust representations and generalize well to unseen samples of novel classes. Several commonly used geometric transformation methods for data augmentation include flipping, rotation, cropping, scaling, and translation at the picture level, which are usually used in data pre-processing. Performing data augmentation during the data pre-processing stage can increase the diversity of inputs effectively to a certain extent. Some data-augmentation-based few-shot learning methods train a hallucinator [6,27] to generate proposals or images containing novel-class objects by transferring the knowledge from base classes to novel classes. Although still finishing data augmentation at the picture level, the knowledge learned from base classes is fresh and novel for unseen classes. This greatly enriches the training data of novel classes by transferring knowledge through generation. A spatial transformer network (STN) [28] allows neural networks to actively manipulate and reason about the spatial transformations within input data. This enables a neural network to learn spatial invariance and perform geometric transformations on its input data, such as translation, rotation, scaling, and cropping. This method has inspired us to propose a novel approach for performing data augmentation within neural networks. Indeed, unlike picture-level data augmentation, we aim to perform data augmentation to enrich novel-class data at the feature level. Therefore, an intrinsic feature transform module (IFTM) is proposed to transform the crucial local features extracted by LFEM and increase the diversity of features from novel classes, thereby enriching the learnable feature information of novel classes.

Several existing methods [18,29,30] mainly leverage the global attention modules to aggregate features from the query and support images before the procedure of classification and bounding box regression. Channel-wise attention, used widely in global attention mechanisms, encodes the support features into a vector that contains the global contextual information of the support image. Although interacting the channel information with query feature maps helps the model extract the correlation between different channels, the interacting features, only in a global context, ignore the importance of some local pixels

in query and support images. The interaction between query and support features in [20] indicates that the crucial local context improves the performance of FSDO. As a result, a challenge in feature aggregation is finding a balance between local and global context. Depth-wise convolution is used to enhance the query feature encoder and the support feature encoder, which helps to extract local features efficiently. In the meanwhile, we retain the aggregation on global context using a global attention branch. The global and local attention are combined into the glocal cross-attention network (GCAN), which is used to generate the aggregation features for the detector.

In this paper, we combine the three aforementioned modules into the Faster-RCNN detector and employ contrastive learning on the second training stage.

On the PASCAL VOC dataset, we significantly improved the performance in the majority of shot conditions. To demonstrate the value of our suggested modules and the positive impacts of meta contrastive learning in FSOD, we conducted a number of extensive experiments.

The main contributions of this paper can be summarized as follows:

- The local feature enhancement module (LFEM), which can deeply extract the local feature representations on query and support images, is proposed.
- The intrinsic feature transform module (IFTM) is proposed, which transforms the feature extracted by LFEM and enriches the features information of novel classes.
- The Global Cross-Attention Network (GCAN) is proposed by integrating the global and spatial attention mechanisms. We build a balance between the interaction of global and local features and provide high-quality aggregation features for the detector.
- The aforementioned modules are put together in the Faster RCNN to create an exceptional FSOD network, which achieves excellent performance on the PASCAL VOC dataset.

2. Related Works

2.1. Object Detection

Localizing and identifying objects of interest in an image is the problem of object detection. Each object's bounding box and the correct object category must be predicted by the object detector. With the advent of deep learning, CNN-based methods, which have been divided into two groups, two-stage and single-stage detectors, have emerged as the dominant paradigm in the field of object identification. Two-stage detectors demonstrate that the models generate region proposals, including RCNN [31] and its variants [1,32–35], using a separate module called the region proposal module. The first technique that made use of CNN to boost detection performance was RCNN [31]. The region proposal module generates proposals, some of which are picked out by selective search, considering they are very likely to contain objects. Features of proposals will be extracted into vectors by CNN and finally classified by the SVM classifier. SPP Net [32] puts the convolution layers before the region proposal module, reducing the operation of uniforming the input size of images, which avoids the object deformation input warping. Both RCNN and SPP Net work slowly due to the separated training process of multiple components. Fast RCNN [33] was proposed to solve this issue by designing an end-to-end trainable network. Fast RCNN replaces the pyramidal pooling layers with an RoI pooling layer, which associates the feature maps with proposals. Faster RCNN [1] introduces anchor boxes by proposing a fully CNN-based network called the region proposal network (RPN), thereby making the detector run faster. R-FCN [34] is proposed to solve the issue of translation invariance in CNN and share most of the computations within the model in the meantime. Mask RCNN [35] replaces the RoI pooling layer with RoI Align on Faster RCNN and adds a mask head parallel to the classifier and boxes regressor head for classifying each pixel in proposals. Single-stage detectors such as YOLO [36], its variants [37,38] and SSD [39] finish the classification and boxes regression in the meantime. YOLO [36] regards the detection task as a regression problem, thereby using a fully connected layer to classify and locate the objects. SSD [39] concatenates multiple hierarchical feature maps after feature extraction

and performs regression for the object's position coordinates and classification. While single-stage detectors perform more quickly than two-stage detectors, two-stage detectors offer advantages in accuracy. In this paper, we adopt Faster RCNN as the base detector.

2.2. Few-Shot Learning

Recent deep learning methods require a lot of computation and resources since they train models with a great deal of data. Few-shot learning (FSL) refers to machine learning methods that can learn new knowledge or concepts with only a few training data examples, which is wisely used in few-shot classification (FSC). Transferring knowledge from the domain of base-class domain to that of novel-class domain is the core goal of FSL. Meta-learning [40] is employed in most few-shot learning methods and is considered as the basic technique for FSL. Metric-based methods [8–10,41] leverage the learning of distance function to measure the distance between two samples. Siamese neural net (SNN) [41] uses a couple of weighted-shared CNNs and takes a pair of samples as input for image recognition. The network is trained to determine whether two samples belong to the same category. The match network [9] computes the cosine similarity between the embeddings of support and query images unlike the L1 distance used in SNN. The prototypical network [10] encodes the query and support images into embedding vectors, and the prototype of each class is defined by the average of embedding vectors from support images in this class. The network makes predictions by calculating the squared Euclidean distance between the query's and each class's embedding vector, which represents the similarity between the query image and each class. The relation network [8] utilizes a CNN to measure the similarity score instead of calculating the similarity metric by the distance function. Optimization-based methods [42–44] aim to achieve good performance by optimizing the model on limited training data. LSTM Meta-Learner [42] is modified from long short-term memory (LSTM) and first generates the parameters on the training dataset and optimizes them on the test dataset. Model-Agnostic Meta-Learning (MAML) [43] is proposed to find good initialization parameters that make the model adapt new tasks with a few shot samples quickly. Meta-Transfer Learning (MTL) [44] employs a pretrained deep neural network (DNN) to extract features and completes the meta-training procedure on the last layer of the classifier. Some model-based methods [45,46] design the model framework according to the task in particular. Some fine-tuning-based methods [47,48] transfer the knowledge from a related task that has been trained on the model, leveraging transfer learning [49].

2.3. Few-Shot Object Detection

Leveraging a vast amount of annotated images, general object detection networks perform excellently. The difficult task of learning to detect novel classes of things using just one or a few examples per class is known as few-shot object detection (FSOD). Since localization is an additional assignment, FSOD is more complicated than FSC. Existing FSOD methods can be divided into two categories: fine-tuning-based methods and meta-learning methods. Fine-tuning-based methods [50–53], also called transfer learning-based methods, aim to improve the detection performance of novel classes by transferring the knowledge learned from base classes to novel classes. TFA [50] employs a two-stage framework, Faster RCNN, and considers that the features extracted by the backbone and RPN are class-agnostic. Since the weights of the feature extractor are fixed in the second step, only the parameters of the box classifier and box regressor need to be fine-tuned after the entire framework has been trained on base-class data in the first stage. MPSR [51] is proposed to use an independent branch to process the object and resize its feature maps to various scales. The model finally refines the predictions with multi-scale positive samples. FSCE [52] introduces a contrastive head parallel to the box classifier and box regressor to measure the similarity scores between proposal embeddings. Leveraging the contrastive head and contrastive proposal encoding loss, FSCE enlarges distances between different clusters and increases the generalizability of the model. Meta-learning

methods [18–21,54] use a siamese network with a query and a support branch to improve the generalizability. FSRW [19] aims to perform the learning of reweighting coefficients with a few samples by measuring the intrinsic importance of novel-class features on a end-to-end YOLOv2 framework. MetaDet [54] finetunes a weight prediction meta-model to predict the parameters of class-specific components from a few examples of novel classes. Meta RCNN [18] applies meta-learning over RoI features and introduces a predictor-head remodeling network (PRN) containing a shared backbone with Faster RCNN. The PRN employs channel-wise soft-attention to generate the attentive vectors of each class that are used to remodel RoI features. DCNet [20] and DAnA [21] improves the detection performance by proposing attention-based aggregation modules. DAnA highlights the relevant semantic features of support images by the dual-awareness attention and incorporates the spatial correlations between query and support features, while DCNet utilizes a similar co-attention module.

3. Method

In this section, the preliminaries of the FSOD task will be first described. Then, the following shows the specifics of the modules in our proposed model for FSOD.

3.1. Preliminaries

Our meta-learning-based method leverages a two-stage training paradigm, which learns generalizable feature knowledge from base set D_{base} with an extensive corpus of annotated samples and learns to detect novel-class objects from novel set D_{novel} that only includes a few annotated examples about novel classes. Note that the base classes C_{base} and the novel classes C_{novel} are disjointed, namely, $C_{base} \cap C_{novel} = \emptyset$.

The training strategy of meta-learning is employed in this paper. Each batch of training data includes query data $Q(x, y)$, where x is an image with novel objects, y is the corresponding label consisting of class annotations and bounding boxes, and a support set S contains a series of support images cropped from the box annotations for each class of D_{base}. In both training stages, the model trains the detector based on lots of query–support data pairs. In the first stage, the base detector learns to detect the objects in the query images x_q that belong to classes C_{base}. In the second stage, the model will finetune the detector with the query–support batch from $D_{base} \cup D_{novel}$ to adapt to detect all novel-class objects in the image pairs from $C_{base} \cup C_{novel}$. Note that the procedures of the two training stages are separate and identical, apart from the difference in iteration numbers. The overall process follows the standard meta-learning of Equation (1).

$$M_{inital} \xrightarrow{D_{base}} M_{base} \xrightarrow{D_{novel}} M_{novel} \tag{1}$$

3.2. Network Overview

Our proposed few-shot object detection method is shown in Figure 1. The two-stage detector Faster-RCNN [1] for FSOD can be divided into four main parts: the backbone, the feature aggregation part, the region proposal network (RPN) [1], and the Region of Interest (RoI) [1,33] head. The input images of D_{base} and D_{novel} are fed to the backbone to generate feature maps and sent to the feature aggregation network designed to refine local features and aggregate global features between query image and support images. Then, RPN generates proposals based on the similarity between the query and support features, and the RCNN head commences the process of classification and refining bounding box regression.

Query–support images pairs constitute the input of our method, similarly to the meta-learning based methods mentioned above [18–21]. A support set $S = \{S_1, S_2, \ldots, S_N\}$ consisting of a group of support images is organized before the model training procedure, where support image S_i indicates an object of a specific class, and N is the number of classes included in the support set S. We input a query–support piar (Q, S) and use the ResNet [3] as the backbone to generate the query features and support features, which are denoted as

$F_Q \in \mathbb{R}^{C\times H\times W}$ and $F_S \in \mathbb{R}^{N\times C\times H'\times W'}$, where C denotes the number of channels, H and W denote the height and width of F_Q, and H' and W' denote the height and width of F_S.

Next, we present the detailed feature aggregation network.

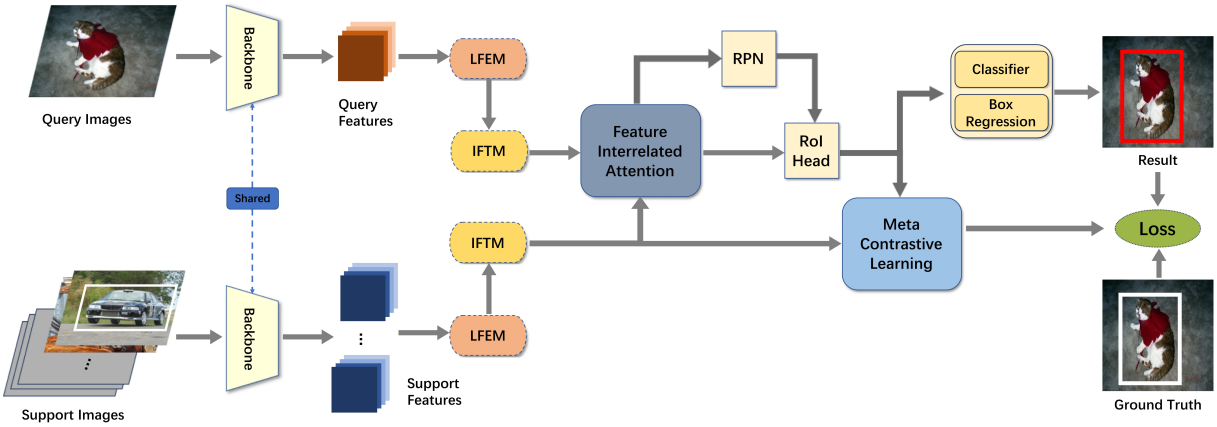


Figure 1. The framework of the proposed architecture. The LFE module, the IFT module and the FIA network are plugged into the standard into Faster-R-CNN. The shared backbone extracts the coarse features of the query images and support images.

3.3. Local Feature Enhancement Module

Gaining query features F_Q and support features F_S from the backbone, we hope the features contain more crucial local information before aggregation. Some classes in novel set D_{novel} have a high degree of similarity with the classes in base set D_{base}; as a result, the features extracted by the backbone also have a high degree of similarity. So, how to help the FSOD model extract the discriminative features of these classes is a key point of our work. Naturally, we design a module named the local feature enhancement module, as shown in Figure 2.

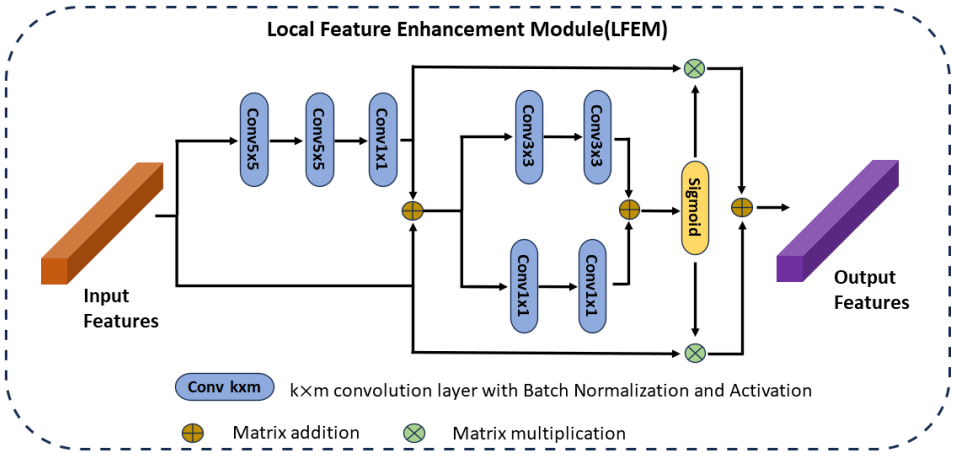


Figure 2. The detailed framework of the local feature enhancement module (LFEM).

We put the features $F \in \{F_Q, F_S\}$ into the LFEM,

$$F' = f_{conv5\times5}(f_{conv5\times5}(f_{conv1\times1}(F))), \tag{2}$$

$$F_{mix} = F + F', \tag{3}$$

$$F'_{mix} = f_{conv3\times3}(f_{conv3\times3}(F_{mix})) + f_{conv1\times1}(f_{conv1\times1}(F_{mix})), \tag{4}$$

$$w_F = f_{sigmoid}(F'_{mix}), w_{F'} = 1 - w_F, \tag{5}$$

$$F_{ouput} = w_F \times F + w_{F'} \times F', \tag{6}$$

where $f_{convk\times m}$ refers to a complete function with a convolution layer, a Batch Normalization (BN) layer and a ReLU layer. Its mathematical expression is shown as follows:

$$f_{convk\times m} = f_{ReLU} \cdot f_{BN} \cdot Conv_{k\times m}, \tag{7}$$

where k, m is the size of the convolution kernel. In this module, we use two channels to operate the input features; the first channel is used to extract the fine features. Motivated by ResNet, we add fine features to original features and then generate local attention weight w_F, which makes the local features more attractive in the output features F_{output}.

3.4. Intrinsic Feature Transform Module

After features enhancement, we propose a method to enrich the representation of features. The depiction of characteristics is monotonous considering there are just a few annotated samples of fresh classes. We propose the intrinsic feature transform module to alleviate this problem. Motivated by the spatial transform network (STN) [28], which shows that different scales and angles of feature representation are different in CNN, IFTM helps to transform features to increase the diversity of the training data. A detailed structure of the IFTM is shown in Figure 3.

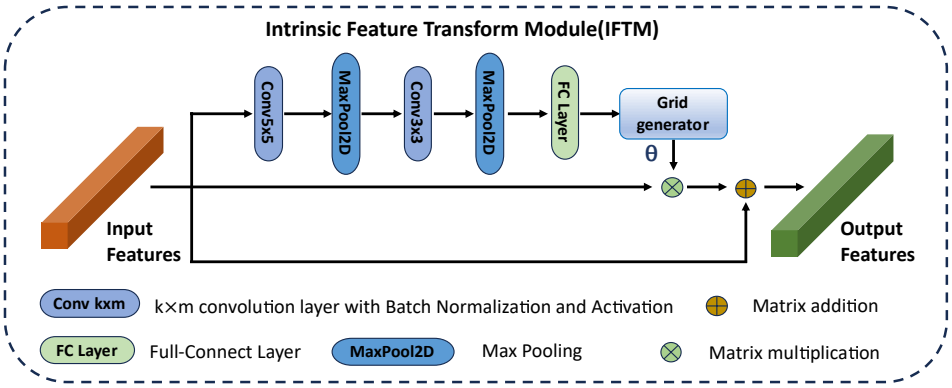


Figure 3. Detailed structure of intrinsic feature transform module (IFTM).

$$F' = f_{FC}(f_{maxpool}(f_{conv3\times3}(f_{maxpool}(f_{conv5\times5}(F))))), \tag{8}$$

$$\theta = f_{lt}(F'), \tag{9}$$

$$F_{output} = F \times \mathcal{T}_G(\theta) + F, \tag{10}$$

where the definition of $f_{convk\times m}$ is the same as that shown in Equation (7), $f_{maxpool}$ refers to the function of maxpooling, f_{FC} refers to the function of the Full-connect Layer, θ refers to the angle to transform features, f_{lt} refers to the function Local Transformation that generates the parameter θ, and $\mathcal{T}_G$ refers to the transformation matrix with the input θ. For clear exposition, we assume that $\mathcal{T}_G(\theta)$ is a 2D affine transformation matrix A_θ; the pointwise transformation of the feature map is

$$\begin{pmatrix} x_i^{out} \\ y_i^{out} \end{pmatrix} = A_\theta \begin{pmatrix} x_i^{in} \\ y_i^{in} \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^{in} \\ y_i^{in} \\ 1 \end{pmatrix}, \tag{11}$$

where (x_i^{in}, y_i^{in}) are the source coordinates of the regular grid in the input feature map, and (x_i^{out}, y_i^{out}) are the target coordinates of the regular grid in the output feature map. In this module, the transformation not only operates the feature map with rotation but also allows scale. Then, we obtain the output features with different feature representation information of novel-class objects.

3.5. Global Cross-Attention Network

By utilizing LFEM and IFTM, the model extracts further local information from the original feature maps of the support images and query images. The next critical issue is how to fully utilize these traits for the purpose of detecting novel-class objects. A series of images including each class to be recognized makes up the support–image input branch in meta-learning-based FSOD models. In other words, all of the objects in the query image belong to the support classes. The FSOD models focus on the key regions of the query image through the interaction of query features F_Q and support features F_S. The local pixel similarity between different features is captured by the spatial–feature interaction attention. However, spatial–feature interaction attention disregards the significance of the global context. The channel-wise attention mechanism captures the global context of support features to gain the channel information correlations. However, it comes at the cost of relinquishing spatial information. Consequently, we combine the global and spatial attention mechanism and design a Global Cross-Attention Network containing a spatial feature attention mechanism, as illustrated in Figure 4.

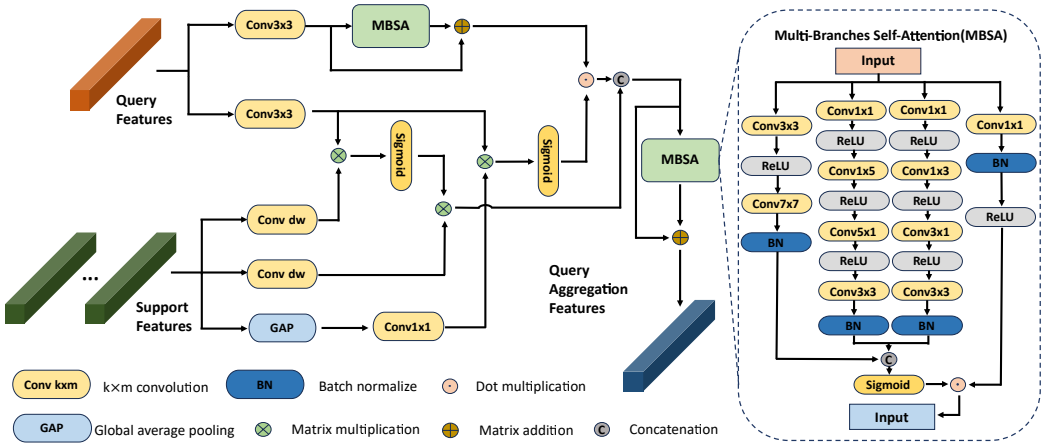


Figure 4. The detailed framework of the Global Cross-Attention Network (GCAN).

Query feature F_Q is encoded into query value feature map V_q and query key feature map K_q by two parallel 3×3 convolutional layers. In the meanwhile, support features F_S are decoded into a pair of value and key feature maps V_s, K_s by two parallel depth-wise convolutional layers. Here, we use depth-wise convolution to be the encoder instead of standard convolution. Depth-wise convolution seperates the sigle filter into two parts called depth-wise convolution and pointwise convolution. This approach extracts powerful local feature representations with a local receptive field and combines them with a pointwise convolution. In addition, we need a global average pooling branch to gain the global context of F_S, which is donated as F_S^G. Equations (12)–(14) show the encoding calculation.

$$V_q, K_q = Conv_{3\times3}(F_Q), \tag{12}$$

$$V_s, K_s = Conv_{Dw}(F_S), \tag{13}$$

$$F_S^G = Conv_{1\times1}(f_{GAP}(F_S)), \tag{14}$$

where $Conv_{3\times3}$ refers to standard convolution, $Conv_{Dw}$ refers to depth-wise convolution and f_{GAP} refers to the operation of global average pooling. In order to focus on crucial feature representations, we use the multi-branches self-attention module (MBSA) to process query value feature map V_q.

$$V_{q1} = f_{BN} \cdot Conv_{7\times7} \cdot f_{ReLU} \cdot Conv_{3\times3}(V_q), \tag{15}$$

$$V_{q2} = f_{BN} \cdot Conv_{3\times3} \cdot f_{ReLU} \cdot Conv_{5\times1} \cdot f_{ReLU} \cdot Conv_{1\times5} \cdot f_{ReLU} \cdot Conv_{1\times1}(V_q), \tag{16}$$

$$V_{q3} = f_{BN} \cdot Conv_{3\times3} \cdot f_{ReLU} \cdot Conv_{3\times1} \cdot f_{ReLU} \cdot Conv_{1\times3} \cdot f_{ReLU} \cdot Conv_{1\times1}(V_q), \tag{17}$$

$$V_{q4} = f_{ReLU} \cdot f_{BN} \cdot Conv_{1\times1}(V_q), \tag{18}$$

$$V'_q = f_{sigmoid}(Cat(V_{q1}, V_{q2}, V_{q3})) \odot V_{q4}, \tag{19}$$

where $Cat(\cdot)$ refers to concatenation operation and $\odot$ refers to dot multiplication. Note that the operation of reducing the channels of feature maps is used in the calculation of the query feature correlation map $V_{qi} \in \{V_{q1}, V_{q2}, V_{q3}\}$. In these three process streams, the first convolution layer is used to reduce the channels of feature maps and preserve feature information as much as possible. In detail, channels of the first branch $C_{V_{q1}} = \frac{N}{2}$, channels of the first branch $C_{V_{q2}} = \frac{N}{4}$ and channels of the second branch $C_{V_{q3}} = \frac{N}{4}$. The purpose of the first process branch is extracting the global features. In the meanwhile, the second and third branches are utilized to measure the local feature correlation. In order to balance the local and global feature information during the concatenation process, the channels is divided into two parts on average. Different sizes of convolution layers represent different receptive fields. $Conv_{5\times1}$ and $Conv_{3\times1}$ are employed to measure the horizontal feature correlation; naturally $Conv_{1\times5}$ and $Conv_{1\times3}$ are employed to assess the vertical feature correlation. The concatenation output contains the local and global feature correlation information, which is employed to process the query self-attention feature map V'_q.

Then, we first compute the key attention matrix M_k with K_q and K_s:

$$M_k = f_{sigmoid}(Re^{\top}(K_q) \otimes Re^{\top}(K_s)), \tag{20}$$

where Re refers to the feature map reshape operation, $\top$ is the matrix transpose operation and $\otimes$ is the multiplication of the matrix. We sequentially obtain a query–feature–interaction support feature map K'_s:

$$K'_s = Re^{\top}(M_k \otimes Re^{\top}(V_s)), \tag{21}$$

which K'_s raises the query local feature aware at the pixel level. Meanwhile, we utilize the support–image global context matrix F_S^G and query key feature map K_q to generate the global attention matrix M_q:

$$M_q = f_{sigmoid}(Re^{\top}(K_q) \otimes Re^{\top}(F_S^G)). \tag{22}$$

Global attention matrix M_q indicates the similarity between the query feature and the support features of N classes. Moving forward, we can enhance the crucial regions of the query value feature map, which have high correlation to support features. The cross-image enhanced feature maps V''_q are as follows:

$$V''_q = \sum_{i=1}^{N} Re^{\top}({M_q}^i \odot Re^{\top}(V'_q)). \tag{23}$$

We integrate the global category information into V'_q; then, we concatenate V'_q and K'_s together and input this global attention feature map into MBSA:

$$F'_Q = Cat(V'_q, K'_s) + f_{MBSA}(Cat(V'_q, K'_s)), \tag{24}$$

where f_{MBSA} represents the function of the MBSA module, and F'_Q represents the query aggregation features. Finally, we combine the original query features F_Q and query aggregation features F'_Q :

$$\hat{F}_Q = Conv_{1\times1}(Cat(F_Q, F'_Q)). \tag{25}$$

Throughout the process, we integrate global and local attention mechanisms while importing the cross-attention mechanism to aggregate query features and support features.

3.6. Meta-Contrastive Learning

Output query features $\hat{F}_Q$ are fed into the RPN head and RoI head. The RPN head can generate a series of region proposals. After that, the RoI head crops the regions in query features into fixed size feature maps by leveraging the RoI module [33], and then, it encodes them as vector embeddings named RoI features F_{RoI}. The detector has never seen novel classes in a query image ever before. Thus, previous methods [55–58] directly conduct the task of classification and regression on RoI features F_{RoI}, producing lots of mistakes in classification results when accurately locating the objects of novel classes. Obviously, some similar classes have close decision boundaries, which when not well separated lead to a misclassification problem. Contrastive learning methods [59,60] are usually used in self-supervised learning, especially in scenarios with limited data availability. Meta-contrastive learning methods make a suitable output for constructing the similarity between query and support images. We divide the contrastive learning function into two parts: one is an intra-class supervised contrastive part aiming to increase the correlation features of the same class; the other one is an inter-class contrastive part aiming to separate similar classes at a suitable distance by decreasing the similarity of different classes.

With the information regarding the class labels, we measure the similarity between F_{RoI} with a loss function based on supervised methods. Therefore, we build a supervised contrastive loss function to make F_{RoI} belong to the same class into a close cluster. The intra-class supervised contrastive loss L_{icsc} is as follows:

$$L_{icsc} = \frac{1}{N_p}\sum_{i=1}^{N_p} L_{icsc}(z_i), \tag{26}$$

where N_p is the number of positive proposals, $i \in I \equiv \{1...N_p\}$,

$$L_{icsc}(z_i) = -\sum_{a\in A(i)} \log \frac{\exp(z_i \cdot z_a/\tau)}{\sum_{b\in B(i)} \exp(z_i \cdot z_b/\tau)}, \tag{27}$$

where z_i refers to the i-th embedding vector of the i-th positive proposal, τ is a scalar temperature parameter, $A(i) \equiv I\backslash\{i\}$ and $B(i) \equiv I\backslash\{i\}$. In addition, $z_i = \frac{p_i}{\|p_i\|}$, where p_i represents the i-th positive proposal of RoI features and $\|\cdot\|$ represents the L2 norms. Additionally, the scalar temperature parameter τ is a hyperparameter, which is set to 0.20 experimentally.

It is not enough to make classification decision boundaries of different classes with a high level of comparability at suitable distance. Support features from different categories provide assistance in reducing the similarity between similar categories. By contrasting the RoI features with the support features which include the common class representations, we can construct an inter-class contrastive loss function. The query–support contrastive loss L_{qsc} is as follows:

$$L_{qsc} = \frac{1}{N_p} \sum_{i=1}^{N_p} L_{qsc}(z_i), \tag{28}$$

where

$$L_{qsc}(z_i) = -\sum_{k=1}^{N} \log \frac{\exp(z_i \cdot R_k/\tau)}{\sum_{j=1}^{N} \exp(z_i \cdot R_j/\tau)}, \tag{29}$$

where $R_k = \frac{r_k}{\|r_k\|}$ refers to the category representation of the i-th class and N is the number of classes in the support dataset. The inter-class contrastive loss focuses on the correlation between the query image and support image and enlarges the inter-class discrepancy to isolate each class cluster.

Finally, the two contrastive learning branches are combined into the meta-contrastive learning method in this work. The meta-contrastive learning function is as follows:

$$L_{Meta} = L_{icsc} + L_{qsc}. \tag{30}$$

where L_{Meta} is the meta-contrastive loss. L_{Meta} enhances the model's ability to distinguish new categories, which effectively alleviates the problem of misclassification.

3.7. Total Loss Function

The total loss function of our method consists of the binary cross-entropy loss L_{rpn} for the RPN, the cross-entropy loss for the L_{cls} classification, the smoothed-L1 loss L_{reg} for the bound box regression head and the meta-contrastive loss L_{Meta}:

$$L_{total} = L_{rpn} + L_{cls} + L_{reg} + \lambda \times L_{Meta}, \tag{31}$$

where λ is a scaling factor and is experimentally set to 0.15.

4. Experiments

4.1. Dataset

We follow the settings in generic FSOD methods to make a fair comparison. We evaluate our model on PASCAL VOC [61] and compare it with plenty of baselines using the same data splits.

PASCAL VOC: The dataset consists of 20 classes, and we evaluate the model on three different class splits. Therefore, we randomly divide the 20 classes into 15 base classes C_{base} and 5 novel classes C_{novel} in each split. Each novel-class split includes $k \in \{1, 2, 3, 5, 10\}$ examples batched from the PASCAL VOC2007 trainval sets and the PASCAL VOC2012 trainval sets. In other words, we use the PASCAL VOC2007 and PASCAL VOC2012 trainval sets for training, and we use the PASCAL VOC2007 test set to evaluate the performance of the models.

In the experiments, we utilize novel average precision (nAP) as the key metric to evaluate the performance. In addition, we use nAP_{50} as the metric, which indicates the mean average precision of novel classes with 0.5 of the intersection over union (IoU) threshold. The following is the calculation formulas of nAP:

$$nAP = \frac{\sum_{i=1}^{N_{novel}} AP(c^i_{novel})}{N_{novel}}, \tag{32}$$

where N_{novel} is the number of classes in D_{novel}; here, we fix it to 5 in the experiments on VOC dataset.

4.2. Implementation Details

We construct the proposed model with the Pytorch library in Python on an RTX 3090 GPU. We choose the Faster R-CNN framework with the ResNet-101 backbone, RPN and RoI head. We do not fix the size of input images; instead, we set the minimum side length

to 400 and the maximum side length to 666. Before the experiments, we crop objects from base classes by utilizing the bounding box annotations of the images to form the support set $D_{support}$. The number of objects of D_{base} in each class is fixed to 200. Additionally, we transform the input images of the support set into a uniform size 192 $\times$ 192. We select the stochastic gradient descent (SGD) to optimize the model using a momentum of 0.9 and weight decay of 0.0001. In experiments, we train the model with a 0.002 learning rate in both the base and fine-tune training stage.

4.3. Comparison with Baselines

We compare our method with the recent approaches to evaluate its performance and effectiveness. We select some famous fine-tuning based methods, e.g., TFA [50], MPSR [51] and FSCE [52]. Our approach is based on meta-learning; therefore, we need to show that the proposed method outperforms current meta-learning methods, e.g., FSRW [19], MetaDet [54], Meta R-CNN [18], FSDetView [62], TIF [63], DCNet [20], DAnA [21] and DGDI [23]. For fair comparison, most baselines use the ResNet-101 as the backbone and Faster R-CNN as the framework detector. Only FSRW (meta-learning based) uses the DarkNet-19 as the backbone and chooses a one-stage detector, YOLO-v2.

To ensure the accuracy and reliability of our results, we run the experiments three times and take the average value as the final results. The experiment results are shown in Table 1. Obviously, the evaluation results indicate that our proposed method outperforms most recent state-of-the-art FSOD baselines. In detail, our method is better than all three famous finetuning-based methods on 1/2/3/5/10 shots of all three splits of novel classes. Especially, our method is 10.4% higher than FSCE for the 2-shot of novel-class split 2, which means the feature information from support images helps detect the objects of novel classes. In other words, meta-learning-based methods are more suitable than finetuning-based methods in the FSOD task. Then, we conduct a detailed comparison with three meta-learning methods (DCNet, DAnA, and DGFI). Our method is much better than DCNet and DAnA in low-shot different class splits. For example, the nAP of ours in the 2-shot of split 2 is 12.4% higher than DCNet and 12.9% higher than DAnA. In novel-class split 3, our method achieves better performance than DGFI. Despite having four metrics that are not as good as those of DGFI, our approach outperforms it overall in terms of performance. Based on the experimental results presented in the table, we can reasonably infer that our approach is capable of more effectively enhancing novel-class feature representations.

4.4. Qualitative Results and Analysis

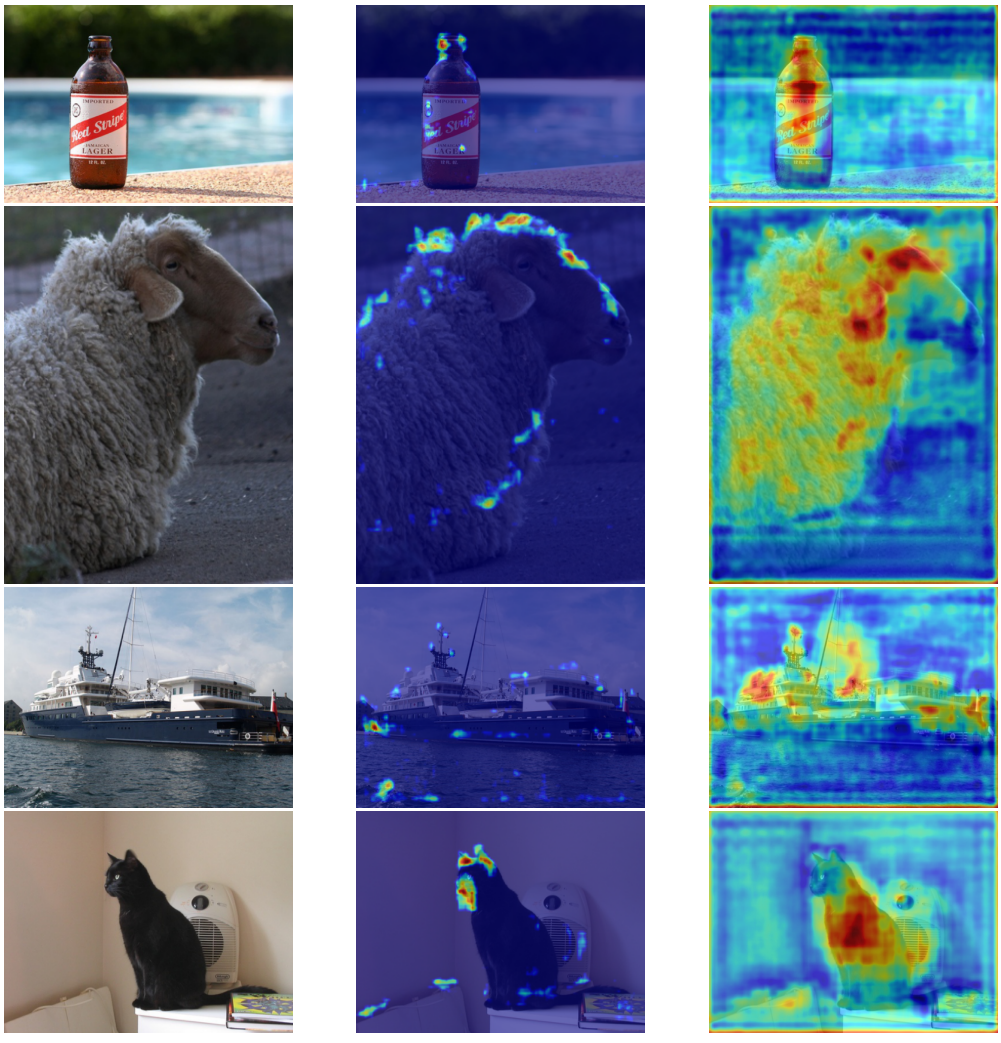
In this subsection, we will conduct analysis of the visual experimental results to further evaluate the performance and effectiveness of our proposed model. Through a comprehensive examination of the visualizations, we can gain a deeper understanding of the model's feature extraction capabilities and its ability to accurately and efficiently detect novel-class targets.

Table 1. Experiments comparison with FSOD baselines in the nAP(%) metric of novel classes on the PASCAL VOC benchmark. We divide the methods into finetuning-based and meta-learning based two parts. Note that all these methods use ResNet-101 as the backbone and follow the two-stage detection framework of Faster R-CNN except for FSCE, which utilizes DarkNet-19 as the backbone and follows the one-stage detector YOLO-v2.

Model	Type	Class Split 1					Class Split 2					Class Split 3				
		1 Shot	2 Shot	3 Shot	5 Shot	10 Shot	1 Shot	2 Shot	3 Shot	5 Shot	10 Shot	1 Shot	2 Shot	3 Shot	5 Shot	10 Shot
TFA	Finetune	25.3	36.4	42.1	47.9	52.8	18.3	27.5	30.9	34.1	39.5	17.9	27.2	34.3	40.8	45.6
MPSR	Finetune	30.0	39.1	46.8	55.2	60.3	18.7	29.1	29.5	38.2	44.6	18.9	32.8	39.3	43.9	52.6
FSCE	Finetune	33.1	40.3	46.9	51.6	59.7	24.2	26.8	37.2	41.7	48.5	22.6	33.4	39.5	47.3	54.1
FSRW	Meta	14.8	15.5	26.7	33.9	47.2	15.7	15.2	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
MetaDet	Meta	18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
MetaRCNN	Meta	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
FsDetView	Meta	24.2	35.3	42.2	49.1	57.4	21.6	24.6	31.9	37.0	45.7	21.2	30.0	37.2	43.8	49.6
TIP	Meta	27.2	36.5	43.3	50.2	59.6	22.7	30.1	33.8	40.9	46.9	21.7	30.6	38.1	44.5	50.9
DCNet	Meta	33.9	37.4	43.7	51.1	59.6	23.2	24.8	30.6	36.7	46.6	32.3	34.9	39.7	42.6	50.7
DAnA	Meta	31.0	41.7	47.8	51.2	54.8	23.3	24.3	35.8	37.5	44.0	32.1	38.5	43.2	50.1	52.0
DGFI	Meta	35.9	43.7	**50.7**	56.2	**61.3**	**26.9**	27.8	39.0	43.2	**51.1**	34.8	41.2	44.2	51.4	56.8
Ours	Meta	**36.1**	**44.2**	50.0	**56.3**	59.9	25.3	**37.2**	**43.9**	**44.0**	48.7	**35.1**	**41.8**	**44.8**	**52.9**	**56.9**

4.4.1. Visualization on Attention Map

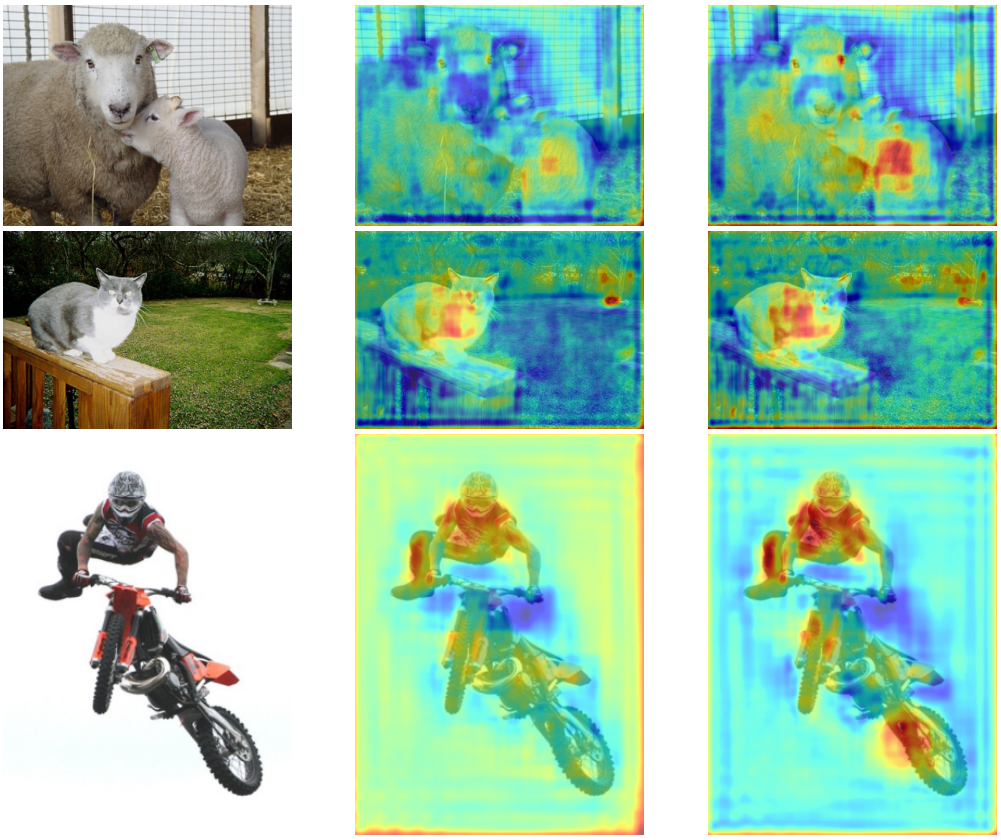
Before being sent into the RPN and RoI head, we can regard the feature maps as attention maps. We generalize the visual attention map by employing Score-based Class Activation Mapping (ScoreCAM) [64], which is a class activation mapping method that highlights the important regions of an image for a particular class by assigning a weight to each pixel based on its contribution to the class score. Figure 5 shows the visual attention maps on some novel-class images. According to these experiment results, the attention maps only have high scores in a few pixels, which belong to objects. In other words, the features before being sent into our proposed network (containing LFEM, IFTM and GCAN) contain little information of novel-class objects. The last column in Figure 5 highlights the important regions of an image, which make a great contribution to locating and classifying the novel-class objects. It illustrates that our proposed network performs well when it comes to enhancing feature representations and increasing its attention on important areas of novel-class query images.



(**a**) (**b**) (**c**)

Figure 5. The ScoreCAM visualization of novel-class query images. The first column presents an original novel-class image. The second column shows the attention map on the features before being sent into our proposed network. The last column depicts the visual attention map after the network. The red color represents high scores of the regions, and the blue color represents low scores of the regions. (**a**) image. (**b**) ScoreCAM before our network. (**c**) ScoreCAM after our network.

To further demonstrate our proposed model's ability, some experiment results on ScoreCAM are compared with the baseline DGDI shown in Figure 6. Figure 6a contains the original query images. Both the first and second image contain the novel classes "sheep" and "cat". The third image contains the novel class "motorbike" and the base class "people". Figure 6b shows the attention maps on ScoreCAM by baseline DGFI and Figure 6c shows attention maps by our method. In comparison, our attention maps obtain more high-scoring areas over the objects to be detected. In detail, DGFI does not even focus on the two "sheep" in the first image. In the third image, both DGFI and our method show the high-scoring attention areas on the base class "people". However, our model performs better when focusing on the novel class "motorbike" object. Through these qualitative comparison experiments, we demonstrate that our model outperforms the baseline DGFI especially in detecting the novel-class objects.

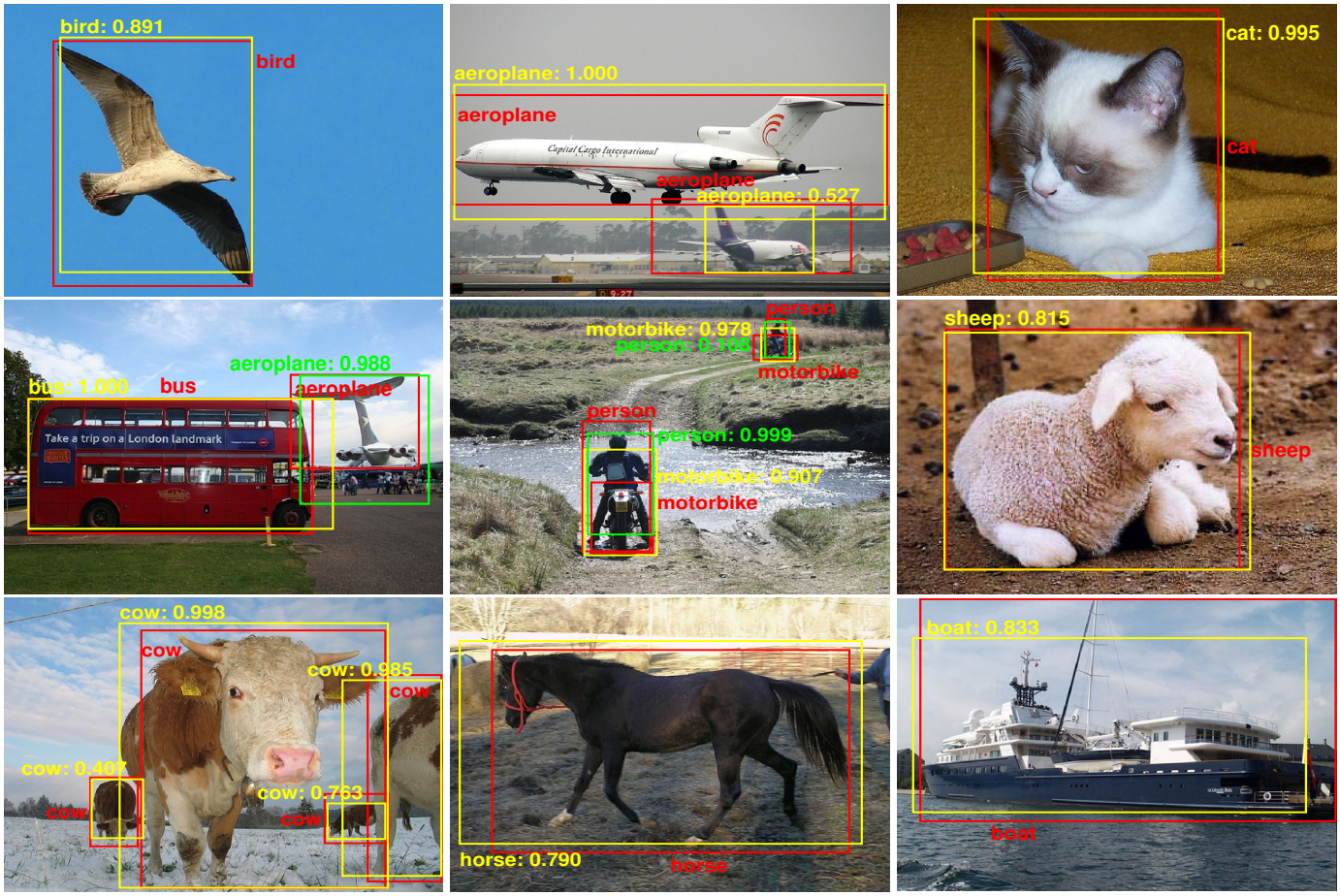


(**a**) (**b**) (**c**)

Figure 6. Visual ScoreCAM comparison on baseline DGFI and ours. Column (**a**) represents the query images. Column (**b**) contains the ScoreCAM results of DGFI, and column (**c**) shows the results from ours.

4.4.2. Visualization on Preditions

In Figure 7, we present qualitative visual experiments of 10-shot detection results on several images from VOC data setups. For easy explanation, we use yellow and green boxes to distinguish novel classes and base classes objects, while red boxes represent the ground truth of all class objects. Most prediction results are closely aligned with the ground truth, exhibiting a high degree of accuracy. Additionally, the confidence scores associated with these predictions are also quite high, further emphasizing the reliable classification of our model. The two prediction results both contain novel classes and base classes objects, which illustrates that our model retains the ability to detect base classes objects while adapting to detect novel-class objects with very constrained annotated samples.



(**a**) (**b**) (**c**)

Figure 7. Visualization of our model's 10-shot object detection results on PASCAL VOC setups. For simplicity, the detections of objects illustrated are mostly belonging to novel classes. Note that the red boxes refer to the ground truth of the objects, the yellow boxes refer to the predictions on novel-class samples and the green boxes refer to the predictions on base-class samples. The scores represent the confidence of the boxes. The qualitative experimental results demonstrate that our proposed model can effectively detect novel objects with constrained annotated samples. (**a**) VOC split 1 (10 shot). (**b**) VOC split 2 (10 shot). (**c**) VOC split 3 (10 shot).

5. Ablation Studies

We conduct some comprehensive ablation studies to verify the effectiveness of our model. The ablation studies are divided into two parts: (i) ablation study on different modules, (ii) ablation study on meta-learning. The experiments of ablation studies are run on the 10-shot class split 3 of PASCAL VOC.

5.1. Ablation Study on Modules

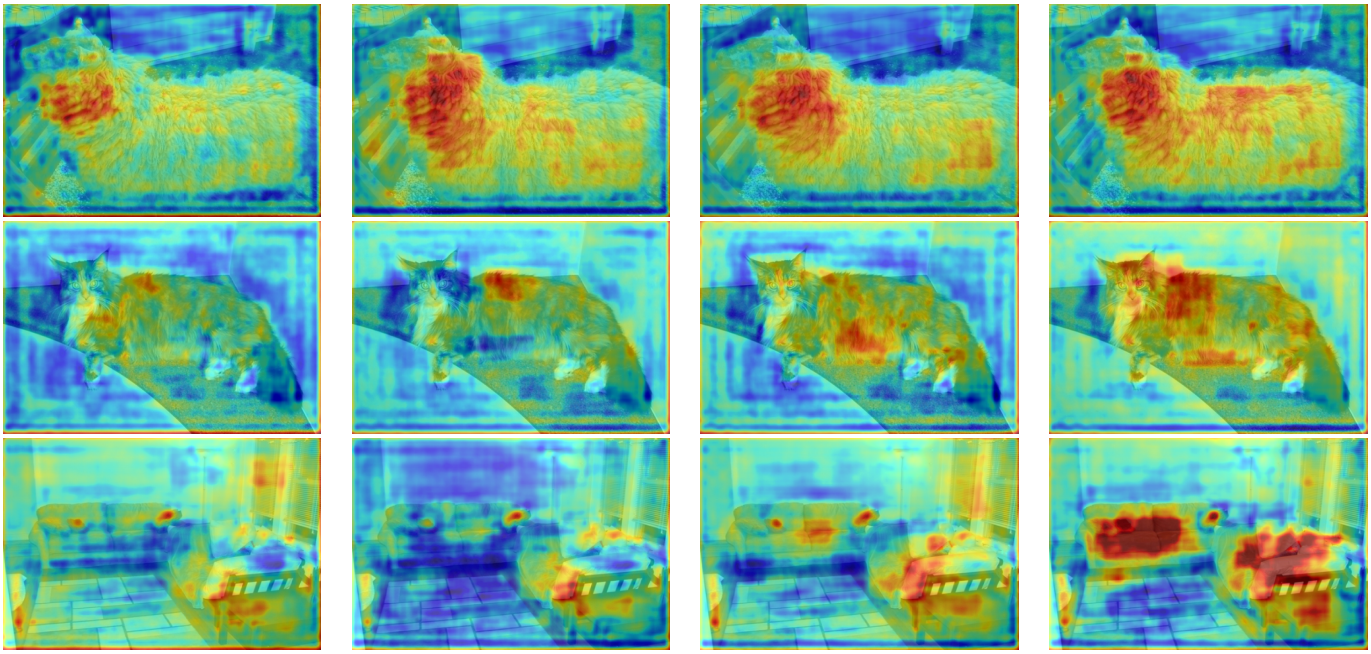
The module for the aggregation of query features and support features is necessary in meta-based FSOD; thus, we keep the GCAN module fixed in the model during the ablation study stage. Here, we conduct the ablation studies on the effectiveness of the LFEM and the IFTM modules, as shown in Table 2. We find that only utilizing LFEM or IFTM will reduce the detection performance of novel classes. However, we obtain great performance improvements when the model both contains LFEM and IFTM. Especially, we obtain 4.8% improvement in the nAP metric on the 2-shot class split 3. These ablation study results demonstrate that LFEM and IFTM cannot independently improve the model's performance. It is evident that there exists a strong correlation between these two mod-

ules, and their combined usage is essential for achieving significant improvements in our proposed method.

Table 2. This is the ablation study result of different modules in the nAP(%) metric of novel classes. The experiments run on PASCAL VOC class split 3.

LFEM	IFTM	GCAN	1 Shot	2 Shot	3 Shot	5 Shot	10 Shot
		✔	32.1	37.0	43.1	49.0	54.6
	✔	✔	30.0	33.7	34.4	44.1	45.9
✔		✔	32.7	32.3	40.5	42.8	49.4
✔	✔	✔	**35.1**	**41.8**	**44.8**	**52.9**	**56.9**

Meanwhile, we conduct some visualization experiments to showcase the detection performance of the model when different modules are employed. The comparison experiments are shown in Figure 8. In addition, the experiments correspond to Table 2 and we choose 10-shot setups for good presentation. From the comparison, the model that only contains GCAN obtains less high-scoring areas on the objects to be detected. When adding the second module (LEFM or IFTM), the model can focus on the right areas that should be detected and classified. It demonstrates that the model with LFEM and IFTM is effective for improving the performance of detecting novel-class objects. The last column presents that the complete model outperforms than other models without one or two crucial components. In a word, the results demonstrate the effectiveness of each module, and the model achieves the best performance when the proposed modules are employed simultaneously.



(**a**) (**b**) (**c**) (**d**)

Figure 8. ScoreCAM visualization ablation studies of different modules. These experiments correspond to Table 2. Column (**a**) shows the ScoreCAM of the model without LFEM and IFTM. Column (**b**) shows the ScoreCAM of the model without IFTM. Column (**c**) shows the ScoreCAM of the model without LFEM. Column (**d**) shows the ScoreCAM of the complete model in this paper.

5.2. Ablation Study on Meta-Contrastive Learning

We conduct an ablation study to analyze the contribution of each loss function in meta-contrastive learning to the performance of our method, as shown in Table 3. By systematically removing individual loss functions, we are able to assess their impact on

the overall performance of our approach. Note that we use all three modules and L_{base} in experiments, where $L_{base} = L_{rpn} + L_{cls} + L_{reg}$. These results indicate that L_{icsc} or L_{qsc} can obtain performance improvements independently. L_{qsc} is more effective than L_{icsc} in improving performance on average. We conjecture that it is more important to separate the decision boundaries of different classes in learning to detect novel-class objects. In the meanwhile, their combined usage leads to more improvements. We are surprised to find that we obtain a 10% raise on 2-shot class split 3 and a 12.7% raise on 10-shot class split 3 when using both L_{icsc} and L_{qsc}. It is evident that these two loss functions possess independence and can be transferred to other models.

Table 3. This is the ablation study result of each loss function of meta-contrastive learning in the nAP(%) metric of novel classes. The experiments run on PASCAL VOC class split 3.

Full Modules	L_{icsc} [1]	L_{qsc} [2]	1 Shot	2 Shot	3 Shot	5 Shot	10 Shot
✔			29.8	30.8	39.2	43.0	44.2
✔	✔		29.0	33.0	40.5	44.7	45.6
✔		✔	33.4	36.4	40.9	44.3	48.0
✔	✔	✔	**35.1**	**41.8**	**44.8**	**52.9**	**56.9**

[1] L_{icsc} refers to the intra-class supervised contrastive loss of meta-contrastive learning. [2] L_{qsc} refers to the query–support contrastive loss of meta-contrastive learning.

6. Conclusions

To increase local feature attention and capture transformational features, the local feature enhancement module (LFEM) and intrinsic feature transform module (IFTM) are described in detail. These modules can assist in putting more emphasis on local features. Then, we improve the existing support–query feature fusion network by designing a Global Cross-Attention Network (GCAN), which facilitates the aggregation of both global and local contextual information of query and support features. These modules are then concatenated into the two-stage detector Faster-RCNN. In order to make the clusters of each class tighter, a meta-contrastive loss function is designed for few-shot object detection inspired by contrastive learning. Extensive experiments on the PASCAL VOC dataset indicate that our proposed method effectively improves the performance on detecting the novel classes compared with some well-known baselines. In other words, the crucial local features extracted by our modules and the interactive contextual information between query features and support features play an important role in learning to detect novel-class objects. Ablation studies on modules demonstrate the substantial link between LFEM and IFTM and the effectiveness of their combined usage in improving performance. In the meanwhile, ablation studies on meta-contrastive learning indicate that each branch of meta-contrastive loss functions can produce satisfactory improvements. The results also show that using both intra-class supervised contrastive loss and query–support contrastive loss results in significant nAP (%) gains. In the future, we will be committed to exploring a lightweight few-shot object detection model, which can balance processing speed and accuracy on novel classes.

Author Contributions: Conceptualization, P.Z. and H.L.; methodology, H.L.; software, H.L.; validation, P.Z. and H.L.; formal analysis, P.Z. and H.L.; writing—original draft preparation, H.L.; writing—review and editing, P.Z. and H.L.; funding acquisition, P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Shenzhen Science and Technology Program (Grant No. KQTD20190929172704911) and the Science and Technology Planning Project of Guangdong Science and Technology Department under Grant Guangdong Key Laboratory of Advanced IntelliSense Technology (2019B121203006).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the authors upon reasonable request.

Acknowledgments: The authors sincerely appreciate the helpful comments and constructive suggestions given by the academic editors and reviewers.

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

References

1. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
2. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), Computational and Biological Learning Society, San Diego, CA, USA, 7–9 May 2015.
3. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
4. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
5. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
6. Wang, Y.X.; Girshick, R.; Hebert, M.; Hariharan, B. Low-shot learning from imaginary data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7278–7286.
7. Wu, J.; Dong, N.; Liu, F.; Yang, S.; Hu, J. Feature hallucination via maximum a posteriori for few-shot learning. *Knowl.-Based Syst.* **2021**, *225*, 107129. [CrossRef]
8. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T.M. Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1199–1208.
9. Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D. Matching networks for one shot learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3630–3638.
10. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4077–4087.
11. Xie, J.; Long, F.; Lv, J.; Wang, Q.; Li, P. Joint distribution matters: Deep brownian distance covariance for few-shot classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7972–7981.
12. Yang, Z.; Wang, J.; Zhu, Y. Few-shot classification with contrastive learning. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 293–309.
13. Guo, Y.; Du, R.; Li, X.; Xie, J.; Ma, Z.; Dong, Y. Learning calibrated class centers for few-shot classification by pair-wise similarity. *IEEE Trans. Image Process.* **2022**, *31*, 4543–4555. [CrossRef]
14. Bendou, Y.; Hu, Y.; Lafargue, R.; Lioi, G.; Pasdeloup, B.; Pateux, S.; Gripon, V. Easy—Ensemble augmented-shot-y-shaped learning: State-of-the-art few-shot classification with simple components. *J. Imaging* **2022**, *8*, 179. [CrossRef]
15. Chi, Z.; Gu, L.; Liu, H.; Wang, Y.; Yu, Y.; Tang, J. Metafscil: A meta-learning approach for few-shot class incremental learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14166–14175.
16. Feng, Y.; Chen, J.; Xie, J.; Zhang, T.; Lv, H.; Pan, T. Meta-learning as a promising approach for few-shot cross-domain fault diagnosis: Algorithms, applications, and prospects. *Knowl.-Based Syst.* **2022**, *235*, 107646. [CrossRef]
17. Lee, K.; Maji, S.; Ravichandran, A.; Soatto, S. Meta-learning with differentiable convex optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10657–10665.
18. Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; Lin, L. Meta r-cnn: Towards general solver for instance-level low-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9577–9586.
19. Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; Darrell, T. Few-shot object detection via feature reweighting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8420–8429.
20. Hu, H.; Bai, S.; Li, A.; Cui, J.; Wang, L. Dense relation distillation with context-aware aggregation for few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 10185–10194.
21. Chen, T.I.; Liu, Y.C.; Su, H.T.; Chang, Y.C.; Lin, Y.H.; Yeh, J.F.; Chen, W.C.; Hsu, W. Dual-awareness attention for few-shot object detection. *IEEE Trans. Multimed.* **2021**, *25*, 291–301. [CrossRef]
22. Zhang, G.; Luo, Z.; Cui, K.; Lu, S.; Xing, E.P. Meta-DETR: Image-level few-shot detection with inter-class correlation exploitation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–12.

23. Huang, L.; Dai, S.; He, Z. Few-shot object detection with dense-global feature interaction and dual-contrastive learning. *Appl. Intell.* **2023**, *53*, 14547–14564. [CrossRef]
24. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
25. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
26. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
27. Zhang, W.; Wang, Y.X. Hallucination improves few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 13008–13017.
28. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 2017–2025.
29. Hsieh, T.I.; Lo, Y.C.; Chen, H.T.; Liu, T.L. One-shot object detection with co-attention and co-excitation. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 2721–2730.
30. Fan, Q.; Zhuo, W.; Tang, C.K.; Tai, Y.W. Few-shot object detection with attention-RPN and multi-relation detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4013–4022.
31. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
33. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
34. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 379–387.
35. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
36. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
37. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
38. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
39. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
40. Schaul, T.; Schmidhuber, J. Metalearning. *Scholarpedia* **2010**, *5*, 4650. [CrossRef]
41. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese Neural Networks for One-Shot Image Recognition. Master's Thesis, University of Toronto, Toronto, ON, Canada, 2015.
42. Ravi, S.; Larochelle, H. Optimization as a model for few-shot learning. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
43. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.
44. Sun, Q.; Liu, Y.; Chua, T.S.; Schiele, B. Meta-transfer learning for few-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 403–412.
45. Munkhdalai, T.; Yu, H. Meta networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 2554–2563.
46. Cai, Q.; Pan, Y.; Yao, T.; Yan, C.; Mei, T. Memory matching networks for one-shot image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4080–4088.
47. Wang, Y.; Chao, W.L.; Weinberger, K.Q.; Van Der Maaten, L. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv* **2019**, arXiv:1911.04623.
48. Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J.B.; Isola, P. Rethinking few-shot image classification: A good embedding is all you need? In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 266–282.
49. Torrey, L.; Shavlik, J. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*; IGI Global: Hershey, PA, USA, 2010; pp. 242–264.
50. Wang, X.; Huang, T.; Gonzalez, J.; Darrell, T.; Yu, F. Frustratingly Simple Few-Shot Object Detection. In Proceedings of the International Conference on Machine Learning, Virtual, 12–18 July 2020; pp. 9919–9928.
51. Wu, J.; Liu, S.; Huang, D.; Wang, Y. Multi-scale positive sample refinement for few-shot object detection. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 456–472.

52. Sun, B.; Li, B.; Cai, S.; Yuan, Y.; Zhang, C. Fsce: Few-shot object detection via contrastive proposal encoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 7352–7362.
53. Qiao, L.; Zhao, Y.; Li, Z.; Qiu, X.; Wu, J.; Zhang, C. Defrcn: Decoupled faster r-cnn for few-shot object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 8681–8690.
54. Wang, Y.X.; Ramanan, D.; Hebert, M. Meta-learning to detect rare objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9925–9934.
55. Liu, W.; Wen, Y.; Yu, Z.; Yang, M. Large-margin softmax loss for convolutional neural networks. *arXiv* **2016**, arXiv:1612.02295.
56. Wang, F.; Cheng, J.; Liu, W.; Liu, H. Additive margin softmax for face verification. *IEEE Signal Process. Lett.* **2018**, *25*, 926–930. [CrossRef]
57. Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W. Cosface: Large margin cosine loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5265–5274.
58. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 4690–4699.
59. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
60. Wang, X.; Qi, G.J. Contrastive learning with stronger augmentations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5549–5560. [CrossRef]
61. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
62. 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]
63. Li, A.; Li, Z. Transformation Invariant Few-Shot Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 3094–3102.
64. Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; Hu, X. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 24–25.