electronics

MDPI

Article

Few-Shot Object Detection with Memory Contrastive Proposal Based on Semantic Priors

Linlin Xiao *, Huahu Xu, Junsheng Xiao and Yuzhe Huang

School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
* Correspondence: xiaolinlin@shu.edu.cn

Abstract: Few-shot object detection (FSOD) aims to detect objects belonging to novel classes with few training samples. With the small number of novel class samples, the visual information extracted is insufficient to accurately represent the object itself, presenting significant intra-class variance and confusion between classes of similar samples, resulting in large errors in the detection results of the novel class samples. We propose a few-shot object detection framework to achieve effective classification and detection by embedding semantic information and contrastive learning. Firstly, we introduced a semantic fusion (SF) module, which projects semantic spatial information into visual space for interaction, to compensate for the lack of visual information and further enhance the representation of feature information. To further improve the classification performance, we embed the memory contrastive proposal (MCP) module to adjust the distribution of the feature space by calculating the contrastive loss between the class-centered features of previous samples and the current input features to obtain a more discriminative embedding space for better intra-class aggregation and inter-class separation for subsequent classification and detection. Extensive experiments on the PASCAL VOC and MS-COCO datasets show that the performance of our proposed method is effectively improved. Our proposed method improves nAP50 over the baseline model by 4.5% and 3.5%.

Keywords: object detection; few-shot learning; semantic fusion; contrastive learning; memory contrastive proposal

check for updates

Citation: Xiao, L.; Xu, H.; Xiao, J.; Huang, Y. Few-Shot Object Detection with Memory Contrastive Proposal Based on Semantic Priors. *Electronics* **2023**, *12*, 3835. https://doi.org/10.3390/electronics12183835

Academic Editors: Yiqi Wu, Dejun Zhang and Yilin Chen

Received: 10 August 2023
Revised: 7 September 2023
Accepted: 8 September 2023
Published: 11 September 2023

1. Introduction

Deep learning-based object detection algorithms rely on large amounts of training data to achieve excellent performance [1–4]. In real-world scenarios, such as in the medical and aviation fields, there is a problem of scarce sample data. With a restricted sample size, existing deep learning algorithms are not very effective when it comes to object detection. In contrast, humans can learn new knowledge quickly with very few sample examples. FSOD [5] draws on human learning capabilities to rapidly learn novel category knowledge based on prior knowledge for more accurate classification and detection in a few novel class sample instances.

The FSOD is currently studied mainly based on meta-learning [5–8] and fine-tuning [9–11] methods. Meta-learning utilizes a learn-to-learn methodology and allows for rapid adaptation without relying on additional training, which exhibits lower performance. The fine-tuning-based method has achieved better performance by directly fine-tuning box models on a few sample datasets. For example, TFA [9] is a two-stage fine-tuning training model based on the Faster R-CNN [12] framework. In the first stage, a large amount of base-class data is involved in pre-training. In the second stage, a balanced dataset is involved in fine-tuning the box model to further improve the model's performance. The model also demonstrates the effectiveness of the fine-tuning-based method.

When visual information about a novel class is limited, we consider introducing rich semantic information to improve the diversion of the detector to the novel class.

Indeed, both semantic and visual information describes the detection object in different ways [13]. The visual information extracted by the feature extractor is limited, especially in the presence of noise and intricate background samples, which are more likely to be identified incorrectly. In contrast to the visual features, the semantic information of the class label of an image is invariant. The motivation for introducing semantic information is displayed in Figure 1. We use the word-embedding model glove [14] to embed the semantic representation into the visual space, complementing the visual features via the interaction between them to achieve a deeper focus on the significant features for more accurate recognition.

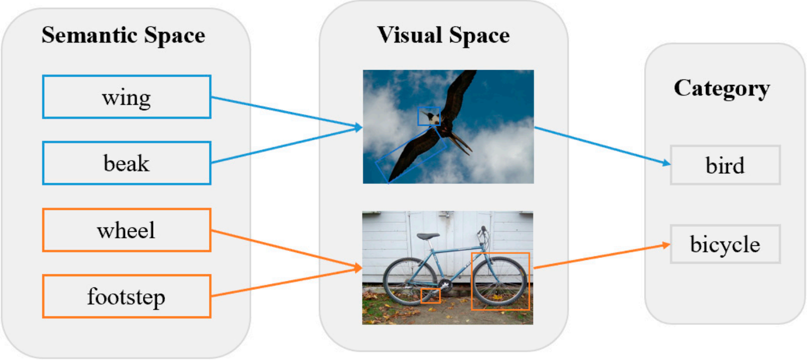


Figure 1. The illustration diagram shows the motivation for introducing semantic information. In the few-shot case, the visual features of the novel object are inadequate. With the embedding of semantic information, we can supplement the feature information of novel objects and enhance the expression of feature information to achieve accurate discrimination.

Our aim was to further improve the detection of FSOD to solve the misclassification problem in the classification stage, where the discrimination of novel and base categories is prone to confusion. We needed a restriction that enables features from the same category to be aggregated and those from different categories to be separated. Therefore, we introduced the memory contrastive proposal (MCP) module to adapt the feature distribution. During the training phase, we constantly update the class-centered features using a dynamic update with the help of a memory bank. We calculate the contrastive loss between the current input features and the class-centered features in the previous memory bank to adapt the feature space distribution, thus enabling the classifier to learn a more discriminative embedding space and achieve more accurate classification.

The main contributions of this paper are as follows:

- We introduce semantic information and design the semantic fusion (SF) module, in which we achieve an interaction between semantic and visual information to enhance the representation of feature information and achieve a deeper focus on significant features.
- We designed the memory contrastive proposal (MCP) module to learn increasingly accurate feature representations by continuously updating the class-centered features. Thus, it enhances the intra-class embedding capability, allowing for a more discriminative embedding space to improve the performance of the detector.
- Experiments based on the PASCAL VOC and MS-COCO datasets are conducted to validate the effectiveness of our proposed method. Moreover, we use the visualization methods Grad-CAM and t-SNE for a more intuitive presentation of the experimental results.

2. Related Works

2.1. General Object Detection

The task of object detection is to identify all interesting parts within an image and determine their class and location. Deep learning-based object detection algorithms fall into two categories, which are the one-stage object detection algorithms and the two-stage object detection algorithms. The one-stage framework, typified by the YOLO [15–17] series, directly predicts the location and class of objects based on features without region proposals. The major problems are anchors deviating from the object area and feature information from inaccurate locations with lower detection accuracies [18,19]. The two-stage framework, typified by the R-CNN [12,20,21] series, primarily performs classification and localization tasks on these region proposals after first generating several potential region proposals. In contrast, the two-stage algorithms are inferior to one-stage in real-time but usually have superior detection accuracy. Both types of detection algorithms rely on a large amount of annotated data.

2.2. Few-Shot Object Detection

Few-shot learning is a challenging machine learning task that is extremely difficult to study for the bias in the distribution of the visible and invisible classes. Nowadays, most of the research is carried out based on classification [22,23]. Few-shot object detection (FSOD) is a core problem in few-shot learning, which involves classification and location tasks that are more difficult to perform than classification [24,25]. FSOD research is mainly based on meta-learning and fine-tuning methods. Among them, FSRW [5] fully learns the basic feature information based on meta-learning methods to rapidly learn novel objects by predicting feature weights. Meta R-CNN [7] proposes a meta-learning paradigm that incorporates feature information from a support set dataset into a network model in a supervised manner with a loss function. TFA [9] used a two-stage training strategy to fine-tune the detector, which achieves better performance than the meta-learning-based method. FSCE [10] proposed a training strategy for contrastive proposal encoding to improve detection performance by contrastive learning to obtain more significant visual representations. MPSR [11] adds FPNs to the backbone network to address the problem of sparse scale distribution in FSOD. Considering both simplicity and effectiveness, we adopt the two-stage fine-tuning method. We introduce a semantic representation into the FSOD task and propose a new method of contrastive learning.

2.3. Semantic in Visual Tasks

Semantic information has been widely applied in the zero-shot learning [26–28] task to learn the interactive projection of visual space and semantic space. Zero-shot learning has made meaningful progress by aiding with semantic information. In contrast, the study of semantic information is relatively less in FSOD. KGTN networks [29] represent the semantic associations among base classes and novel classes in the form of the knowledge graph, where the graph nodes represent classification weights and the edges represent semantic relationships, learning deeply feature information. SRR-FSOD [30] combines semantic relations and visual information to perform relational reasoning in the semantic space for learning novel objects rapidly. In contrast to the above methods, we project semantic information into the visual space for fusion to further enhance the representation of visual features.

2.4. Contrastive Learning

Recent studies apply contrastive learning to the self-supervised learning domain, where the core idea is to learn a better feature representation space utilizing data augmentation that effectively distinguishes between positive and negative samples [31,32]. In contrast to self-supervised learning, the idea of contrastive learning in supervised learning is to maximize the consistency between images of the same category in the embedding space and minimize the similarity between images of different categories. Supervised contrastive learning [33] introduces label information to extend contrastive learning by allowing multiple positive pairs for one instance, and the proposed supervised contrastive loss function effectively improves classification accuracy. While contrastive learning is increasingly used in image classification tasks [23], using contrastive learning ideas in

FSOD work is rare. FSCE [10] first applied contrastive learning in FSOD, which achieved performance improvement by facilitating the formation of a more discriminative feature space via contrastive proposal encoding. CAReD [34] proposed a contrastive learning method incorporating attention mechanisms that significantly improved detection performance. In contrast to the above methods, we propose the concept of class-centered features to adjust the feature distribution in the embedding space by calculating the contrastive loss between the current input sample features and the previous class-centered features to achieve better detection.

3. The Proposed Method

In this section, we first define the basic problem setting of FSOD. Then, we introduced the semantic fusion (SF) module and memory contrastive proposal (MCP) module. Finally, we introduced the two-stage training strategy. We propose the method shown in Figure 2.

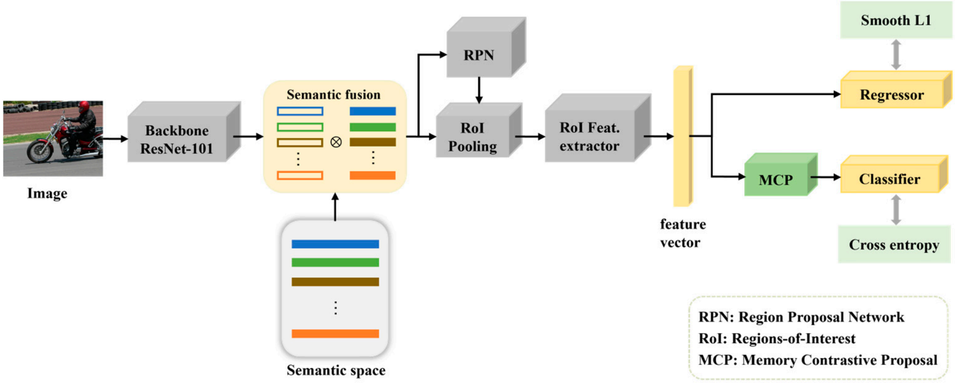


Figure 2. The framework we propose for FSOD. We add semantic fusion (SF) and memory contrastive proposal (MCP) modules. The SF module embeds semantic information in the semantic space for interaction in the visual space. The MCP module aims to learn a more discriminative embedding space between different categories. The combination of the two modules is designed to achieve more accurate detection under limited sample conditions.

3.1. Problem Definition

We reference recent studies on FSOD [9,10] for problem definition. The dataset involved in training contains two parts: a base set $D_{base} = \{(x_i, y_i), i \in \{1, 2, \ldots, N_b\}\}$ containing a large number of labeled samples and a novel set $D_{novel} = \{(x_i, y_i), i \in \{1, 2, \ldots, N_n\}\}$ containing a small number of novel class samples, where x_i is the image and y_i is the corresponding class label. The number of samples per class in D_{novel} sets k. Notably, the class set of base class data is indicated by C_{base}, while the class set of novel class data is denoted by C_{novel}, and $C_{base} \cup C_{novel} = \varnothing$. FSOD aims to learn extensively from D_{base} and generalize rapidly to D_{novel}. We follow a classical two-stage training strategy. In the first base training stage, training is performed similarly to a traditional object detector. In the second novel class fine-tuning stage, we sample from $D_{base} \cup D_{novel}$ to form a balanced dataset to participate in training and further fine-tune the parameters of the base detector. With the two-stage training strategy, the detectors learn more class parameters and aim to detect all objects belonging to $C_{base} \cup C_{novel}$ in the test set.

3.2. Semantic Fusion

In the few-shot setting, the feature extractor extracts restricted features, and we propose to adopt semantic knowledge to complement visual features in a top–down manner with more attention weights distributed for critical features. We extract visual features f_q by a backbone feature extractor and encode the features as a feature vector $v(v \in R^{d_v})$ using a fully connected (FC) layer, where d_v denotes the dimensionality of the feature vector. We choose the glove [14] word-embedding model to process the class label information to obtain the word vector $W_e = \left\{ W_e^c \in R^{d_s} \right\}_{c=1}^{N}$, which represents the semantic information. Here, W_e^c denotes the word vector corresponding to category c, d_s is the dimensionality of the word embedding, and N is the number of object classes. To interact with the semantic and visual information, we design a semantic fusion (SF) module, which transforms via a multilayer perceptron (MLP) to project the word vector information from the semantic space into the visual space $h : R^{N \times ds} \rightarrow R^{N \times dv}$. We obtain the visual attention of different spatial regions by nonlinear changes. This process could be understood as paying more intense attention to critical feature information by fusing it with semantic information, achieving an enhancement of feature information. As shown in Equation (1), the SF module can be indicated as

$$SF(v) = \frac{1}{N} \sum_{i=1}^{N} softMax\left(h\left(W_e^i\right)\right) \otimes v_i \tag{1}$$

The SF module, performing mapping through the softMax function, uses the form of $\otimes$ a dot product to achieve the interaction from semantic information to visual information, which guides the model to pay attention to more critical feature information via semantic information, thus obtaining enhanced visual features.

3.3. Memory Contrastive Proposal

After obtaining the visual features fused with semantic information through the SF module, the features are sent to the region proposal network (RPN) [12] to generate a series of region proposals $p = RPN(\hat{v})$, which is followed by extracting the region proposal feature information $\hat{f}_q$ through the region of interest (RoI) and encoding the features into a feature vector $v_{RoI} \in R^{P \times D_1}$ using the fully connected (FC) layer. Among them, P denotes the number of RoI features, and D_1 denotes the dimensionality of the RoI feature vector, which is often set to $D_1 = 1024$. Previous FSOD work [9,10] generally performs classification and regression directly after extracting the feature information, which is usually not highly accurate. It is often due to confusion between the novel class and the base class, which results in misclassification. We introduced the memory contrastive proposal (MCP) module to solve this problem, as shown in Figure 3. We first project the input proposal features into the embedding space, observe the feature distribution, and find the corresponding class-centered features by distance calculation to store in the memory bank. The new input features are then learned by comparing them with previous features in the memory bank, distancing the distance in the feature representations of cases from different categories and closing the distance in the feature representations of samples from the same category, thus forming tighter clusters between similar classes and keeping longer distances around various classes, further forming a more discriminative embedding space.

Our proposed MCP module relies on an external memory bank that stores the feature centers of previous sample features with different classes corresponding to distinct centers, which we call class-centered features. We perform contrastive learning between the feature representation of the current sample and the class-centered features of the potential spatial features from the repository, calculating the memory contrastive proposal loss to adjust the feature distribution to learn a well-embedding space.

Specifically, we first initialize the all-zero vector as the class-centered features of each category into the memory bank M (in the case of the balanced dataset constructed by PASCAL VOC, which contains 20 categories, it corresponds to 20 all-zero vectors in our M), $M_0 = \left[m_i^0\right]$, where $i \in C_{base} \cup C_{novel}$. Then, we obtain the novel class-centered features by the weighted average of the feature representation of the current input MCP module and the class-centered features in the memory bank and update the memory bank in parallel. We denote the feature v_{ROI}^i of class I extracted by RoI as θ_{v_i} and the corresponding class-centered feature m_i in the memory bank M as θ_{m_i}, we and update θ_m by the following equation:

$$\theta_{m_i}^{k+1} \leftarrow \mu\theta_{v_i}^{k} + (1-\mu)\theta_{m_i}^{k} \tag{2}$$

where $\mu \in [0,1]$ is a momentum factor and $\theta_{m_i}^{k+1}$ represents the class-centered features of the feature representation of the i category in the kth iteration. Only the parameter $\theta_{v_i}^{k}$ is involved in the back-propagation update. The above momentum Equation (2) makes the variation $\theta_{m_i}^{k+1}$ smoother. In this paper, we use the form of the contrastive loss function called InfoNCE to represent the contrastive loss between the previous class-centered feature and the current input feature representation. In Equation (3), θ_m^+ represents the class-centered feature of v_{RoI}^{i} in M corresponding to the input category and τ represents the temperature hyperparameter.

$$L_{mcp} = -\log \frac{\exp\left(\theta_{v_i} \cdot \frac{\theta_m^+}{\tau}\right)}{\sum_{j=1}^{C_b} \exp\left(\theta_{v_i} \cdot \frac{\theta_m^j}{\tau}\right)} \tag{3}$$

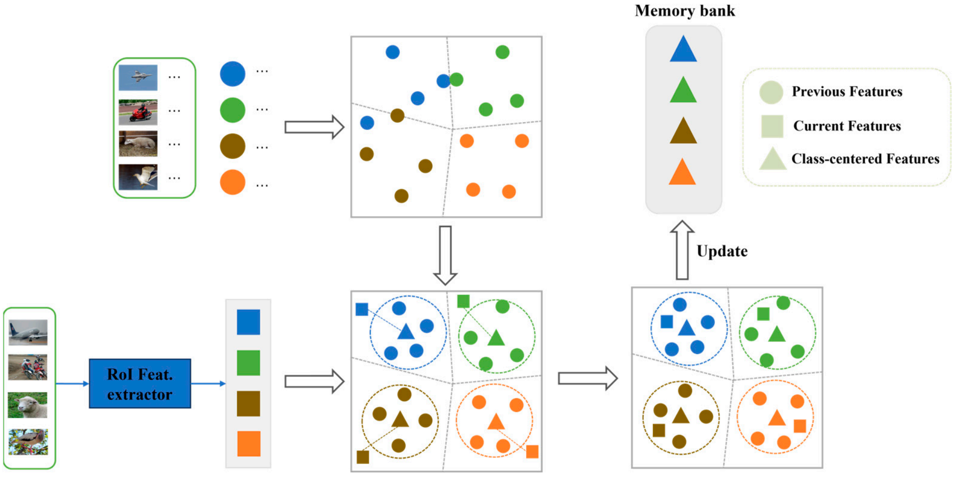


Figure 3. Detailed architecture of the memory contrastive proposal module. Different colors represent different categories. By constantly updating the class-centered features in the memory bank, obtain a more discriminative feature distribution.

In addition, Algorithm 1 provides a detailed training procedure for the MCP module. Through iterative computation, the class-centered features in the memory bank are increasingly accurate class representations. It also helps to achieve a good feature distribution for the input features and to prepare them for subsequent classification.

Algorithm 1 Memory contrastive proposal

```
Input: Base set $D_{base}$, novel set $D_{novel}$, image $x_i$, feature vector $v_{RoI}$ is represented as $\theta_{v_i}$, momentum factor $\mu$, temperature parameters $\tau$. Initialize memory bank $M = [m_i]$, where $m_i = [0]^{1\times d}$ is represented as $\theta_{m_i}$, $i \in C_{base} \cup C_{novel}$ and $d$ represents the dimension of the feature vector.
Output: Update the memory bank M and contrastive loss $L_{mcp}$.
for $x_i \in D_{base} \cup D_{novel}$ do
    RoI extract feature information for RPN: $\theta_{v_i}$
    Update the memory bank M by the feature representation $\theta_{v_i}$ with a momentum
    factor $\mu$: $\theta_{m_i} \leftarrow \mu\theta_{v_i} + (1-\mu)\theta_{m_i}$
    Calculate memory contrastive proposal loss: $L_{mcp} = -\log \frac{\exp(\theta_{v_i} \cdot \theta_m^+ / \tau)}{\sum_{j=1}^{C_b} \exp(\theta_{v_i} \cdot \theta_m^j / \tau)}$
    where $\theta_{\mathrm{m}}^+$ represents the class-centered features corresponding to $\theta_{\mathrm{v_i}}$
end for
```

3.4. Training Strategy

Flowing the two-stage fine-tuning training strategy, the training process is the same as the general object detection method in the first base training stage. In addition, the semantic fusion (SF) module is added after our feature extractor to enrich the input feature information. The joint training objective of our method is defined in Equation (4).

$$L = L_{rpn} + L_{cls} + L_{reg} \tag{4}$$

where L_{rpn} is the binary cross-entropy loss for training the base detector to generate the proposed RPN headers from the bounding box anchors, L_{cls} is the cross-entropy loss for the bounding box classifier, and L_{reg} is the smoothed L1 loss for the bounding box regression headers.

In the fine-tuning stage, we use L_{mcp} to learn a more discriminative embedding space. We perform memory contrastive proposal (MCP) learning after extracting RoI features. The joint training objective of our method is defined in Equation (5).

$$L = L_{rpn} + L_{cls} + L_{reg} + \lambda L_{mcp} \tag{5}$$

where λ is set to 0.5 to control the balance between L_{mcp} and other losses.

4. Experiments

Our experiments build on two datasets, PASCAL VOC [35] and MS-COCO [36]. First, we present the basic setup of the experiments and then conduct comparative experiments. We perform an ablation study to validate the effect of the proposed modules and hyperparameters in our method on the experimental results. We conclude with visualization to further validate the validity of the method.

4.1. Implementation Details

We use PyTorch to implement our proposed network model, based on the Faster-RCNN [12] framework, with ResNet-101 as the backbone network. All training utilized a stochastic gradient descent (SGD) optimizer with momentum set to 0.9, weight decay set to 0.0001, learning rate set to 0.001, and input batch size set to 16. For word embedding, we used the 300-dimensional glove [14] vector from the language model trained on Wikipedia. All experiments were conducted on a computer with a GTX3090Ti GPU.

Implementation on PASCAL VOC. We follow previous work on FSOD [9] with the same setup and class division rules. We consider three random splits, where each split randomly selects five categories as novel classes and the remaining 15 classes as base classes: that is, Split 1 (bird, bus, cow, motorbike, sofa/other), Split 2 (plane, bottle, cow, horse, sofa/other), and Split 3 (boat, cat, motorbike, sheep, sofa/other). Based on the two-stage training strategy, the first stage pre-trained the base class data (15 classes) and the second stage used a sampling rule of K-shot (K = 1, 2, 3, 5, 10) random sampling to sample from a balanced dataset consisting of novel classes and base classes to participate in the training. We report AP50 ($AP_{50} = \int_0^1 p(r)dr$) for novel classes (nAP50) and base classes (bAP50) on the PASCAL VOC 2007 test set, i.e., the average precision of PR curves computed by integration with the IOU threshold set to 0.5.

Implementation on MS-COCO. We followed the previous setup [9] and selected 60 categories in the MS-COCO dataset that did not intersect with PASCAL VOC as base classes, while the other 20 categories were used as novel classes. Compared to the PASCAL VOC dataset, the MS-COCO dataset contains image samples with more complex scenes that are more challenging to identify. We trained the images using K-shot (K = 10, 30). We used AP, AP50, and AP75 to detect the performance of the novel classes based on the MS COCO 2014 minival set.

4.2. Results

To validate the advancement and effectiveness of the proposed method, we selected the existing mainstream FSOD baseline methods for contrastive. These include Meta-Det [37], Meta R-CNN [6], TFA [9], MPSR [11], FSCE [10], FSRW [5], FSOD-UP [38], CAReD [34], QSAM [39] and KFSOD [40].

Results on PASCAL VOC. As shown in Table 1, the performance of our proposed method was evaluated against recent research work for novel classes (nAP50) on the three random splitting settings of PASCAL VOC for different K-Shot (K = 1, 2, 3, 5, 10) settings. The experimental results are all averaged over multiple random novel class sampling sets, which indicates that our method outperforms the results of other models on the PASCAL VOC dataset in most cases. Based on the experimental results, we also found that the performance of our proposed method is more significant in low-shot settings. For example, at the one-shot setting of split 1, our method achieves a result 13.7% higher than the same MPSR based on two-stage fine-tuning, while it is only 6.6% higher than the 5-shot. It also shows that our proposed method can still achieve effective identification even when there is little available data and identification is relatively difficult.

Table 1. Performance in the PASCAL VOC dataset based on three random splits for different k-shot (1,2,3,5,10) settings. Bold indicates the optimal result of the test.

Method/Shots	Novel Split 1					Novel Split 2					Novel Split 3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
Meta-Det	17.1	19.1	28.9	35.0	48.8	18.2	20.6	25.9	30.6	41.5	20.1	22.3	27.9	41.9	42.9
Meta R-CNN	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
TFA	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
MPSR	31.5	40.6	48.3	55.2	60.8	24.4	-	39.2	39.9	47.8	35.6	-	42.3	48.0	49.7
FSCE	32.9	44.0	46.8	52.9	59.7	27.3	30.6	38.4	43.0	48.5	22.6	33.4	39.5	47.3	54.0
FSRW	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
FSOD-up	34.6	43.1	49.3	53.4	59.7	29.5	29.9	37.2	39.4	46.3	32.5	38.7	41.9	45.6	51.5
CAReD	36.5	45.2	47.1	50.8	58.8	26.4	31.0	37.9	43.5	**51.1**	20.2	33.8	41.6	48.3	55.3
QSAM	31.1	36.1	39.2	50.7	59.4	22.9	29.4	32.1	35.4	42.7	24.3	28.6	35.0	50.0	53.6
KFSOD	44.6	-	54.4	60.9	65.8	**37.8**	-	43.1	48.1	50.4	34.8	-	44.1	52.7	53.9
Ours	**45.2**	**49.7**	**55.3**	**61.8**	**66.4**	32.4	**34.1**	**43.6**	**48.7**	50.1	**34.9**	**41.9**	**46.3**	**53.8**	**56.2**

Results on MS-COCO. The detection performed on the MS-COCO 2014 minival set shows the experimental results in Table 2. Based on different K-shot (K = 10, 30) settings, we evaluated our method in contrast to recent research work. The AP, AP50, and AP75 in this table illustrate the average accuracy of novel class instances at various thresholds. It is demonstrated experimentally that our results outperform most previous research work. At the 10/30-shot setting, our method respectively achieved a 4.5% and 4.0% increase in AP values compared to FSCE. All experimental results of our method are optimal compared to other baselines with significant performance improvements achieved.

Table 2. Performance evaluation of AP, AP50 and AP75 for MS-COCO dataset with different k-shot (k = 10,30) settings. Bold indicates the optimal result of the test.

Method	10-Shot			30-Shot		
	AP	AP50	AP75	AP	AP50	AP75
Meta-Det	7.1	14.6	6.1	11.3	21.7	8.1
Meta R-CNN	8.7	19.1	6.6	12.4	25.3	10.8
TFA	10.0	-	9.3	13.7	-	13.4
MPSR	9.8	17.9	9.7	14.1	25.4	14.2
FSCE	11.1	-	9.8	15.3	-	14.2
FSRW	5.6	12.3	4.6	9.1	19.0	7.6
FSOD-up	11.0	-	10.7	15.6	-	15.7
CAReD	15.5	25.1	14.9	18.4	30.1	17.7
QSAM	13.0	24.7	12.1	15.3	29.3	14.5
Ours	**15.6**	**26.4**	**15.7**	**19.3**	**33.6**	**18.9**

4.3. Ablation Experiment

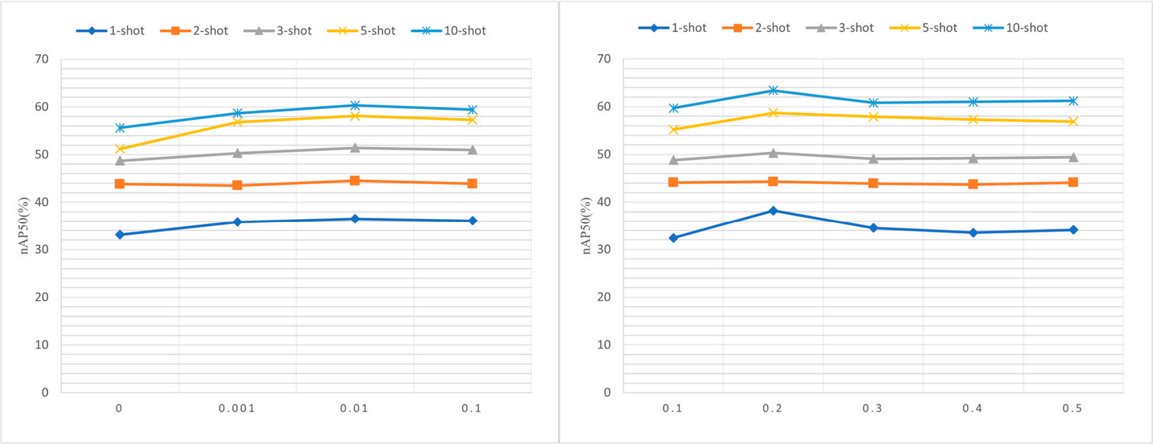
In this part, we conduct ablation research on various hyperparameters of our proposed method, and the experiments are based on PASCAL VOC split1.

Analysis of SF and MCP modules. We conducted an ablation study of the SF module and the MCP module to demonstrate the impact of the corresponding modules on the overall performance. As shown in Table 3, the SF module significantly improved the performance of nAP50 at the 1/2/3-shot settings and remained essentially flat at the 5/10-shot settings compared to the baseline, quantitatively demonstrating the effectiveness of the interaction between semantic and visual information. In particular, at lower shots, where the models do not have much visual information to rely on, the embedding of semantic information effectively compensates for the lack of visual information to guide the models to make correct discriminations. The introduction of the MCP module improves the model performance significantly at all settings compared to the baseline. It suggests that learning a more discriminative embedding space for classification by the classifier through the MCP module is effective. We also see that the combination of the SF module and the MCP module can further enhance the model's overall performance.

Table 3. Ablation experiment based on SF module and MCP module for different k-shot (k = 1, 2, 3, 5, 10) settings.

SF	MCP	1-Shot	2-Shot	3-Shot	5-Shot	10-Shot
-	-	31.6	41.2	48.3	50.4	55.7
√	-	34.3	44.4	50.2	50.9	56.2
-	√	36.1	43.8	51.7	56.3	60.2
√	√	39.2	45.1	52.4	59.7	61.3

Analysis of hyperparameters μ and τ. The memory contrastive proposal loss in the MCP module includes two hyperparameters, namely μ and τ. The value of μ affects the update rate of the memory bank. From Equation (2), the larger the value of μ, the more the memory bank prefers to learn the feature representation of the current sample, and the smaller the value of μ, the more the memory bank prefers to preserve the feature representation of the previous sample. As shown in Figure 4a, our method achieves the best performance at all shot settings when μ = 0.01. The value of temperature parameter τ directly affects the degree of similarity between the input features and the previous class-centered features. We can see that a change in the temperature parameter τ does not significantly affect the value of nAP50 at the 2/3-shot setting. Therefore, we determined the optimal τ based primarily on the 1/5/10-shot. We can observe that the metric nAP50 slowly increases and then gradually decreases. As shown in Figure 4b, the maximum value can be reached when τ = 0.2. Finally, for the sake of integration, we determined τ = 0.2.



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

Figure 4. Hyperparameter ablation results of μ and τ. Column (**a**) represents the experimental results of hyperparameter μ; Column (**b**) represents the experimental results of hyperparameter τ.

Analysis of scale factor λ. The scaling factor λ controls the balance between memory contrastive proposal loss and other losses. As shown in Figure 5, the scale factor λ slowly increases and then gradually decreases in the overall trend of the metric nAP50 in the 5/10-shot setting, with relatively smooth changes, especially taking relatively high values around λ = 0.5/0.6. The maximum value is reached at λ = 0.5 on the 1/2/3-shot setting. The above observation demonstrates that giving a certain weight to the memory contrastive proposal loss may lead to learning a better embedding space. However, too much weight will also damage the nAP50. Comprehensively, we determine λ = 0.5.

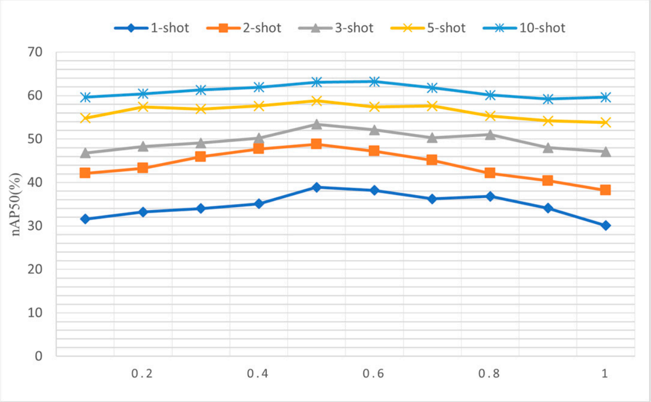


Figure 5. Hyperparameter ablation result of λ. Perform experiments at different k-shot (k = 1, 2, 3, 5, 10) settings to find the optimal λ.

4.4. Visualization Results

To understand our proposed method more intuitively, we show it through visualization. Analogous to the attention mechanism approach, we explain our semantic fusion (SF) module by the Grad-CAM [41] visualization results as an example. As shown in Figure 6, feature learning is inadequate and incomplete without semantic information embedding. With the addition of the SF module, the model was guided to pay more attention to the critical feature information of the novel class samples and learn a more comprehensive visual feature.

To validate the impact of our proposed memory contrastive proposal (MCP) loss, we visualized the embedding space by t-SNE [42]. All test images are from the PASCAL VOC 2007 test set with different colors representing different categories. As shown in Figure 7, we can see the significant changes in the feature distribution after MCP learning. Before the MCP calculation, many similar classes feature distributions would overlap each other. After MCP computation, the feature distributions between similar categories are distanced from each other, resulting in learning a more discriminative embedding space.

In addition, we also provide the object detection results of the novel class cases. As shown in Figure 8, it mainly consists of three typical detection error phenomena: missing detection, low detection confidence score, and classification error. From the detection results, the model is effective by combining the improvements of SF and MCP modules. In the case of challenging detection objects, the proposed method achieves stable and effective detection, which effectively improves the detection performance over the baseline model.

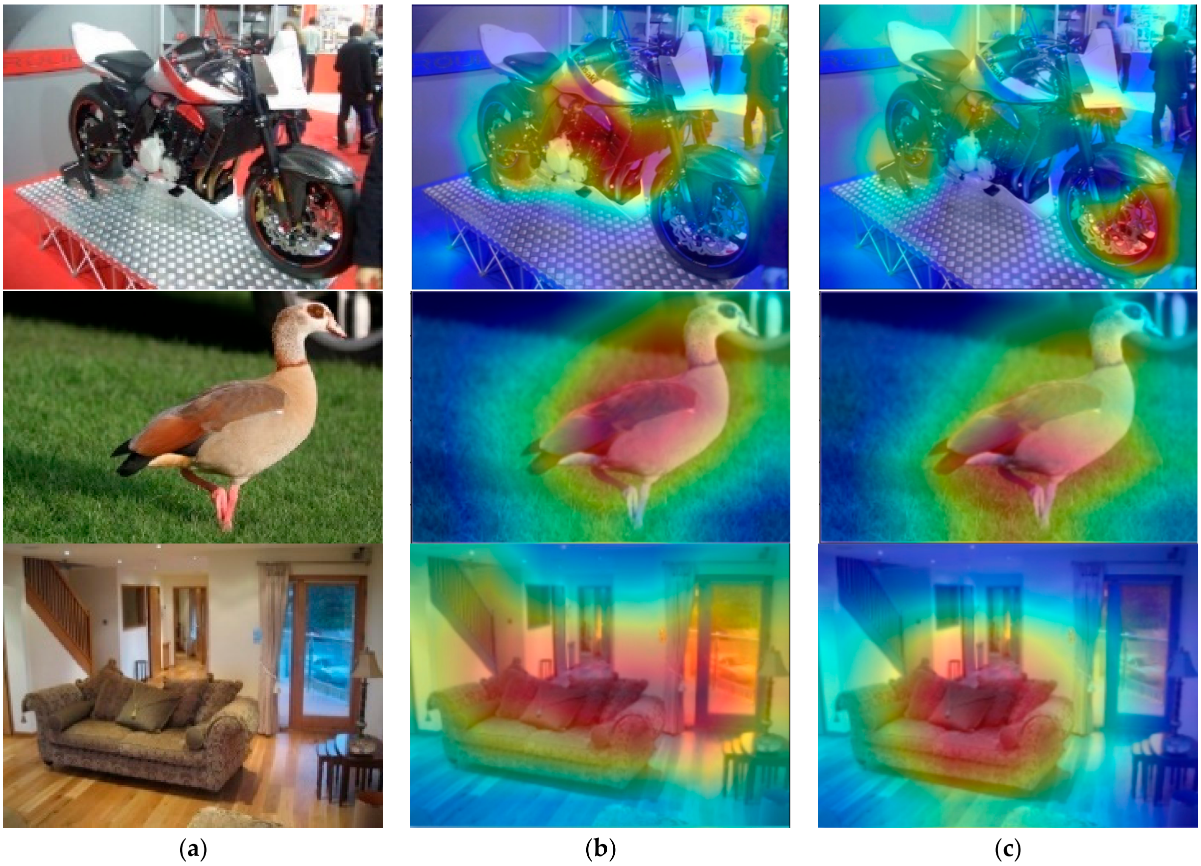


Figure 6. Grad-CAM visualization results. Column (**a**) represents the detection image; Column (**b**) describes the visualization results without the SF module; Column (**c**) represents the visualization results with the SF module. The visualization results in columns (**b**,**c**) demonstrate the effectiveness of the SF module.

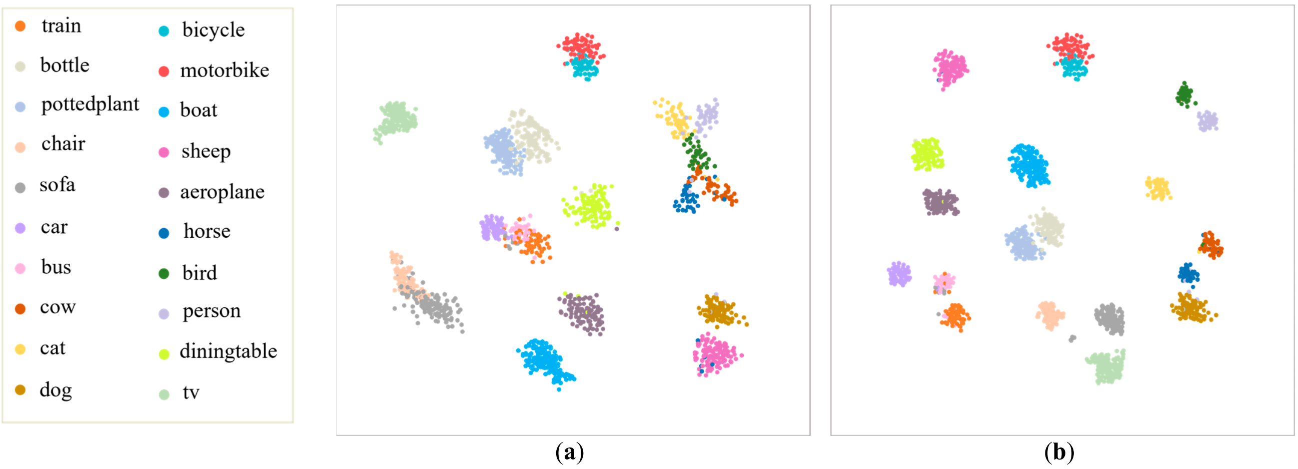


Figure 7. T-SNE visualization results. Column (**a**) represents the feature distribution before the embedding of the MCP module, column (**b**) represents the feature distribution after the embedding of the MCP module. The visualization results in columns (**a**,**b**) show that the distribution of the feature space changes before and after the learning of the MCP module, and the introduction of the MCP module learns a more discriminative embedding space.

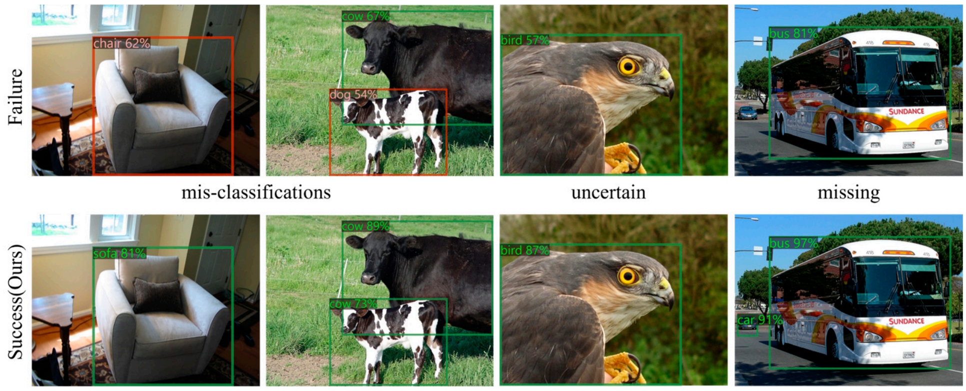


Figure 8. Visualization of detection results for novel class cases on PASCAL VOC split1. The first row is from the baseline model detecting failure cases, and the second row is the successful detection results of our model for the above failure cases. The visualization of the detection results shows that our proposed method achieves correct detection for the failed detection cases of the baseline.

5. Limitations

Our proposed method combines SF and MCP modules to learn enhanced feature representations and achieve effective improvement over existing baseline detection performance. Meanwhile, we also identified some limitations in our research. FSOD models generally lack attention to temporal complexity. Reducing the temporal complexity of the model while improving its detection performance is a critical issue. Our method is more suitable for non-real-time detection scenarios to ensure accuracy, and we leave the real-time detection scenarios for future work to explore.

6. Conclusions

In this paper, we introduce semantic information into the FSOD framework to complement the visual information of novel objects, which helps to learn novel categories better. We propose an FSOD method with semantic fusion (SF) and memory contrastive proposal (MCP) modules. Semantic information and visual features are fused to guide the model to focus on critical features of novel class samples. We adjusted the proposed visual feature distribution through the MCP module to make the feature distribution of the same category more compact and the feature distribution of different classes more distant. Then, we drove the classifier to learn more discriminative embedding spaces. We have demonstrated through extensive experiments that our method achieves state-of-the-art results compared to previous detection methods.

In future work, we will perform more experimental research and design more lightweight frameworks to optimize our approach, improve the performance, and reduce the time complexity of the model. Specifically, we will explore how to achieve deeper interactions between semantic and visual information and design more robust and efficient class-centered feature update mechanisms for more efficient feature extraction and spatial optimization to improve detection performance and efficiency.

Author Contributions: Conceptualization, H.X.; methodology, L.X. and J.X.; software, L.X. and Y.H.; validation, L.X. and J.X.; writing—original draft preparation, L.X.; writing—review and editing, L.X. and J.X.; visualization, L.X.; supervision, H.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Kong, T.; Yao, A.; Chen, Y.; Sun, F. Hypernet: Towards accurate region proposal generation and joint object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 845–853.
2. Yu, G.; Fan, H.; Zhou, H.; Wu, T.; Zhu, H. Vehicle target detection method based on improved SSD model. *J. Artif. Intell.* **2020**, *2*, 125. [CrossRef]
3. Micheal, A.A.; Vani, K.; Sanjeevi, S.; Lin, C.-H. Object detection and tracking with UAV data using deep learning. *J. Indian Soc. Remote Sens.* **2021**, *49*, 463–469. [CrossRef]
4. Zhang, X.; Wan, F.; Liu, C.; Ji, X.; Ye, Q. Learning to match anchors for visual object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3096–3109. [CrossRef] [PubMed]
5. 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.
6. 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.
7. 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, Nashville, TN, USA, 20–25 June 2021; pp. 10185–10194.
8. 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]
9. Wang, X.; Huang, T.E.; Darrell, T.; Gonzalez, J.E.; Yu, F. Frustratingly simple few-shot object detection. *arXiv* **2020**, arXiv:2003.06957.
10. 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, Nashville, TN, USA, 20–25 June 2021; pp. 7352–7362.
11. 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; Proceedings, Part XVI 16, 2020; pp. 456–472.
12. 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**, *39*, 91–99. [CrossRef] [PubMed]
13. Schwartz, E.; Karlinsky, L.; Feris, R.; Giryes, R.; Bronstein, A. Baby steps towards few-shot learning with multiple semantics. *Pattern Recognit. Lett.* **2022**, *160*, 142–147. [CrossRef]
14. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
15. 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.
16. 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.
17. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
18. 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; Proceedings, Part I 14, 2016; pp. 21–37.
19. Chen, X.; Yu, J.; Kong, S.; Wu, Z.; Wen, L. Joint anchor-feature refinement for real-time accurate object detection in images and videos. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 594–607. [CrossRef]
20. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
21. 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.
22. Li, G.; Zheng, C.; Su, B. Transductive distribution calibration for few-shot learning. *Neurocomputing* **2022**, *500*, 604–615. [CrossRef]
23. Cui, Z.; Wang, Q.; Guo, J.; Lu, N. Few-shot classification of façade defects based on extensible classifier and contrastive learning. *Autom. Constr.* **2022**, *141*, 104381. [CrossRef]
24. Li, B.; Wang, C.; Reddy, P.; Kim, S.; Scherer, S. Airdet: Few-shot detection without fine-tuning for autonomous exploration. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 427–444.
25. Zhou, Z.; Li, S.; Guo, W.; Gu, Y. Few-Shot Aircraft Detection in Satellite Videos Based on Feature Scale Selection Pyramid and Proposal Contrastive Learning. *Remote Sens.* **2022**, *14*, 4581. [CrossRef]
26. Chen, L.; Zhang, H.; Xiao, J.; Liu, W.; Chang, S.-F. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1043–1052.

27. Chen, S.; Xie, G.; Liu, Y.; Peng, Q.; Sun, B.; Li, H.; You, X.; Shao, L. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 16622–16634.
28. Li, Y.; Wang, D.; Hu, H.; Lin, Y.; Zhuang, Y. Zero-shot recognition using dual visual-semantic mapping paths. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3279–3287.
29. Chen, R.; Chen, T.; Hui, X.; Wu, H.; Li, G.; Lin, L. Knowledge graph transfer network for few-shot recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 10575–10582.
30. Zhu, C.; Chen, F.; Ahmed, U.; Shen, Z.; Savvides, M. Semantic relation reasoning for shot-stable few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8782–8791.
31. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 1597–1607.
32. Zeng, H.; Cui, X. Simclrt: A simple framework for contrastive learning of rumor tracking. *Eng. Appl. Artif. Intell.* **2022**, *110*, 104757. [CrossRef]
33. 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.
34. Quan, J.; Ge, B.; Chen, L. Cross attention redistribution with contrastive learning for few shot object detection. *Displays* **2022**, *72*, 102162. [CrossRef]
35. 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]
36. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13, 2014; pp. 740–755.
37. 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, Korea, 27 October–2 November 2019; pp. 9925–9934.
38. Wu, A.; Han, Y.; Zhu, L.; Yang, Y. Universal-prototype enhancing for few-shot object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9567–9576.
39. Lee, H.; Lee, M.; Kwak, N. Few-shot object detection by attending to per-sample-prototype. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 2445–2454.
40. Zhang, S.; Wang, L.; Murray, N.; Koniusz, P. Kernelized few-shot object detection with efficient integral aggregation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19207–19216.
41. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
42. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

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.