remote sensing

MDPI

Article

Few-Shot Aircraft Detection in Satellite Videos Based on Feature Scale Selection Pyramid and Proposal Contrastive Learning

Zhuang Zhou [1,2,3], Shengyang Li [1,2,3,*], Weilong Guo [1,2,3] and Yanfeng Gu [4,5]

[1] Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China
[2] Key Laboratory of Space Utilization, Chinese Academy of Sciences, Beijing 100094, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
[4] School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China
[5] Heilongjiang Province Key Laboratory of Space-Air-Ground Integrated Intelligent Remote Sensing, Harbin 150001, China
* Correspondence: shyli@csu.ac.cn

Abstract: To date, few-shot object detection methods have received extensive attention in the field of remote sensing, and no relevant research has been conducted using satellite videos. It is difficult to identify foreground objects in satellite videos duo to their small size and low contrast and the domain differences between base and novel classes under few-shot conditions. In this paper, we propose a few-shot aircraft detection method with a feature scale selection pyramid and proposal contrastive learning for satellite videos. Specifically, a feature scale selection pyramid network (FSSPN) is constructed to replace the traditional feature pyramid network (FPN), which alleviates the limitation of the inconsistencies in gradient computation between different layers for small-scale objects. In addition, we add proposal contrastive learning items to the loss function to achieve more robust representations of objects. Moreover, we expand the freezing parameters of the network in the fine-tuning stage to reduce the interference of visual differences between the base and novel classes. An evaluation of large-scale experimental data showed that the proposed method makes full use of the advantages of the two-stage fine-tuning strategy and the characteristics of satellite video to enhance the few-shot detection performance.

Keywords: satellite videos; aircraft detection; few-shot learning; feature scale selection pyramid; proposal contrastive learning

Citation: 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. https://doi.org/10.3390/rs14184581

Academic Editors: Pedram Ghamisi, Danfeng Hong, Xin Wu and Sicong Liu

Received: 15 August 2022
Accepted: 7 September 2022
Published: 14 September 2022

1. Introduction

The emergence of satellite video provides new opportunities for real-time, continuous Earth observations; such video streams are becoming a novel data source for remote sensing dynamic monitoring. Compared with traditional still images, object detection based on staring data obtained from satellite videos expands a series of innovative applications in industry, agriculture, traffic and other fields, and has become one of the most popular research topics in the field of remote sensing [1–3].

In recent years, the performance of object detection methods based on deep learning, i.e., Faster R-CNN [4], RetinaNet [5], YOLO [6], and others have greatly improved in terms of accuracy and efficiency compared with traditional, hand-crafted, feature based methods. Impressive progress has also been made in the application of object detection in remote sensing images [7–15]. Notably, the DOTA [16,17], DIOR [18], NWPU VHR-10 [19], and other large-scale labeled datasets meet the network training needs of available models. For satellite video, although a few datasets such as SatSOT [20] and VISO [21] have been published, due to the huge cost of sequence frame labeling, their richness and scale are far less than those of still image datasets. In addition, compared with high-resolution remote sensing images, satellite video has lower spatial resolution and contrast, and phenomena such as motion blur and distortion reduce the quality of sequence frames [20,22].

The need to use only a small number of annotations of unseen objects to achieve a suitable model is reflected in satellite videos. Based on the employed detection dataset, novel class objects can be detected with only a few corresponding annotations. By designing training strategies, network structures and loss functions, a detection model with generalization performance can be obtained, which greatly reduces the labeling cost of object detection [23,24]. Compared with few-shot image classification, few-shot object detection is more challenging. The former tries to recognize novel image categories with a few shot samples, and related methods focus on how to obtain features that can better represent the image and how to quickly learn a generalized model for new, unseen tasks without numerous labeled samples [25–27]. The latter task is more complex: the network must distinguish between the object and background, as well as between different objects in the same image. It is necessary not only to extract the high-level semantic features of the image in order to classify different objects, but also to obtain low-level information about the image to accurately locate the object [28,29].

The public remote sensing image datasets used in object detection have the characteristics of high spatial resolution, a large proportion of objects, and rich and salient appearance features. As satellite video has low spatial resolution and small-sized objects and lacks visual features and texture information, objects are difficult to identify, especially in few-shot conditions. Owing to the small amount of satellite video data and sparse objects, it is necessary to rely on external data sources as the base class. Visual changes between the base and novel classes due to domain differences also make detection more challenging.

In this paper, we propose few-shot aircraft detection method using satellite videos based on two-stage fine-tuning via a feature scale selection pyramid and proposal contrastive learning to solve the problems of few annotations, small object size, weak object distinguishability, and domain differences between base and novel classes. The contributions of this paper are as follows.

(1) To the best of our knowledge, the method proposed herein is the first few-shot object detection method for satellite videos. As such, it may serve as a significant reference for related research in the field.
(2) FSSPN and proposal contrastive learning loss are constructed to improve the two-stage fine-tuning method based on Faster R-CNN, making it more suitable for the small size objects and weak object distinguishability of aircraft in satellite videos. In addition, in the fine-tuning stage, we expand the network freezing parameters to reduce the interference of visual differences due to domain shift.
(3) We evaluated the proposed method with large-scale satellite video data. The experimental results showed that our technology makes full use of the advantages of a two-stage fine-tuning strategy and the characteristics of satellite video to enhance the few-shot detection performance.

The rest of this paper is organized as follows. The related work reviewed in Section 2 introduces few-shot object detection and summarizes the progress made in the field of remote sensing. In Section 3, we introduce the task definition and present an analysis, the framework of the proposed method, and the specific composition details. In Section 4, we present experimental data, the result of comparisons with other methods, and ablation experiment results. In Section 5, we discuss our comparison of visualizations and evaluate the performance of different methods. Finally, we draw conclusions and look ahead to the future in Section 6.

2. Related Work

2.1. Few-Shot Object Detection

Current few-shot object detection methods mainly include meta-learning, metric learning, and fine-tuning approaches [30–40].

Unlike convolutional methods that take a specific task as the training objective, meta-learning methods aim to obtain the initialization parameters of models that perform well in different tasks. In this way, the model only needs a few iterations to adapt quickly

and achieve good performance with a new task. Kang et al. [36] introduced a meta-feature learner and a lightweight feature re-weighting module based on YOLOv2. By training with a base class dataset, the meta-features that could be generalized to new objects were extracted, which enabled the detector to quickly adapt to unseen object classes. Wang et al. [41] treated the last layer of a detection network as category-related, and the parameters corresponding to the base class could not be directly applied to the novel class. In view of this, the parameterization of a prediction meta-model was designed so that the weight could be adjusted between the base and novel classes. Yan et al. [42] proposed Meta R-CNN, which uses a support branch to obtain the category attention of the object, and then fuses it with the features in a region of interest (RoI) for detection or segmentation. Perez-Rua et al. [28] proposed a meta-learning method for center point prediction based on CenterNet [41]. Quan et al. [43] propose a cross-attention redistribution module combining the training mode of meta-learning with the supervised contrastive learning framework to further improve the stability and accuracy of a few-shot object detector. Cheng et al. [44] proposed a new meta-learning-based incremental few-shot object detection method, which took CenterNet as its fundamental framework and redesigned it by introducing a novel meta-learning method, thereby adapting the model to unseen knowledge while overcoming the problem of forgetting to a great extent. After adding the novel class, the model could learn incrementally without accessing base class data. However, despite these methods, in practice, it remains difficult to design effective meta-learning strategies due to, for example, the propensity of non-convergence in the iterative training process [45,46].

Metric learning methods are used to map the features of potential objects, support image features in the same embedding space, and then classify the regions by measuring the distance similarity to detect objects. Karlinsky et al. [47] proposed a few-shot object detection method based on distance. In their method, the category of object was represented by a mixture model of different modalities, and the center of each modality was used as the category representation vector. Zhang et al. [48] proposed a few-shot object detection framework based on contrast that can directly detect new classes of objects without adjusting the parameters after the model has been trained. Hsieh et al. [49] designed a margin-based ranking loss function to measure the similarity of features between region proposals and query images. Lu et al. [50] proposed a novel decoupled metric network for single-stage few-shot object detection by designing a decoupled representation transformation and image-level distance metric learning approach to solve the few-shot detection problem. Since metric learning focuses on similarities among categories, and because the ability to localize objects mainly relies on the region proposal network, the performance of this method needs to be further verified.

Methods based on fine-tuning use an abundant base class dataset to pre-train an existing network, and then use several novel class samples to fine-tune some of the parameters to detect novel class objects. Wang et al. [31] found that fine-tuning methods exceeded the performance of meta-learning methods in few-shot object detection experiments. In view of this, researchers have used Faster R-CNN as a detection framework and first pre-trained their models with large amounts of of base class data. Then, the model parameters of the feature extractor and regional proposal network (RPN) in the network are frozen. Finally, a few support samples are used to fine-tune the classification and regression branches, which provides good generalization performance. Sun et al. [33] found that although fine-tuning the classification and regression branches of the network achieved high recall, it was easy to confuse the categories. Therefore, they also fine-tuned the FPN and RPN. In addition, a contrastive loss was introduced to balance the differences between object categories. The method Sun et al. proposed achieved better performance with a public benchmark dataset. Fine-tuning is an efficient few-shot object detection strategy which can achieve relatively good performance, especially under with a limited number of samples. However, the difficulties of fine-tuning methods, i.e., determining the category-related weights and setting the hyper-parameters, must be overcome.

2.2. Few-Shot Object Detection in Remote Sensing

At present, few-shot object detection in remote sensing is still in the initial stages of exploration, with researchers focusing on how to improve general methods in order to better adapt to the characteristics of remote sensing images with multi-scale objects and complex backgrounds. Based on meta-learning, Li et al. [37] fused extracted meta-features at multiple scales based on the method proposed by Kang et al. [36], so that the network could adapt to the scale diversity of the objects in remote sensing images. Cheng et al. [51] improved Meta R-CNN [42] to convert support images into class-aware prototypes. Their guided RPN could generate more effective object proposals from the complex backgrounds of remote sensing images. Among the fine-tuning methods, Huang et al. [38] designed a shared attention module according to the diverse scales of objects and shared the multi-attention map as a priori knowledge with the feature extractor to improve the module's adaptability. Similarly, Zhao et al. [39] added a channel aggregation module to the method proposed by Wang et al. [31] to shorten the path of the bottom-up propagation process of the network and to use the location information of low-level features. Zhou et al. [40] aggregated the convolutional and contextually aware features of different scales to make a network adapt to the scale diversity of objects. Based on the framework of two-stage fine-tuning, Wang et al. [52] designed dilated convolutions and dense connections to obtain rich contextual information from different receptive fields, thereby improving the detection performance with multi-scale and complex backgrounds.

The characteristics of remote sensing images and satellite video are quite different in terms of the size and distinguishability of objects. In addition, unlike remote sensing image datasets, in which novel classes are selected from their own interiors, satellite video must rely on external data as the base class to pre-training. In view of the abovementioned differences, a few-shot object detection method for remote sensing images may not suitable for the aims of the present research.

3. Method

3.1. Problem Definition and Analysis

According to the definition of few-shot object detection [53], we define D_b as the base class set and D_n as the novel class set, while the corresponding classes of objects are C_b and C_n, respectively. For D_n, only K-labeled samples can be used for training, and the classes of the two sets do not overlap, expressed as $C_b \cap C_n = \varnothing$. For dataset $T = \{(X_i, Y_i), i = 1, 2, \cdots, N\}$, X and Y represent the images and annotations of the objects, respectively. The goal here is to detect the novel objects in D_n.

According to the problem definition, in this paper, we take the DIOR dataset [18], which is a benchmark dataset for remote sensing object detection, as D_b. Sequence frame images from satellite video with aircraft annotations are labeled as D_n. Only K objects of D_n are randomly selected for fine-tuning; the remaining samples are used for validation. In order to be consistent with the definition and setup in related work [37–40], we set K = 3, 5, 10 in the experiments described in this paper. It should be emphasized that to ensure that the base and novel classes have no overlap, we removed all images and annotations that contained aircrafts in the DIOR dataset.

Owing to different observation modes and sensor indicators, the spatial resolution of the DIOR dataset can reach sub-meter levels, with clear spatial structures and rich visual features of the objects. In contrast, satellite videos have a lower spatial resolution, i.e., approximately meter-level, and the background information is complex with small and blurred objects. Figure 1 shows a sample comparison between the DIOR dataset and a satellite video frame with the same image size (800 × 800 pixels). There are clear differences in aircraft size, object distinguishability, and overall visual characteristics of the two images.

Figure 1. Comparison of samples in the DIOR dataset (**left**) and satellite video (**right**).

According to the COCO dataset [54] and Chen et al. [55], a small object can be defined by its absolute size (less than 32 × 32 pixels) and relative size (i.e., the proportion of the object area <0.12%). We statistically compared the size information of aircraft in the DIOR dataset and the satellite video used in this paper. In terms of absolute and relative sizes, the small aircraft in the DIOR dataset accounted for 23.03% and 17.07% of the total number of aircraft, respectively, while the corresponding values for the satellite video were 89.44% and 92.46%, respectively. Therefore, for aircraft detection in satellite video, size is a significant problem. Detecting small objects which lack visual features and texture information is the main challenging.

The influence of jitter, radiation difference, and brightness difference in of satellite video observations results in image degradation, such as low contrast and blurring [56], making the distinguishability of aircraft and backgrounds weak. Overcoming this is a significant challenge of the present detection task.

Owning to the limited amount of satellite video data and sparse objects, it is difficult to construct a base class set that meets the training needs based on the satellite video itself. This task requires external labeled remote sensing images as the base class set. The visual differences caused by domain shift also increase the difficulty of few-shot object detection.

Therefore, the method proposed in this paper mainly aims to solve the problems of the small size and poor distinguishability of aircraft, as well as domain differences between base and novel class sets.

3.2. Method Framework

To overcome the detection difficulties caused by the small size of objects, we constructed a FSSPN that makes full use of the shallow feature information by performing contextual attention, feature scale enhancement, and feature scale selection on different layers of the backbone. To improve the poor satellite video contrast and weak distinguishability of aircraft, we added a proposal contrastive learning item to the loss function in the fine-tuning stage to achieve a more robust object representation through supervised learning. To reduce the impact of visual differences caused by domain shift between the base and novel classes, only the classification and regression branches were fine-tuned. The overall framework and algorithm flow of the proposed method are shown in Figure 2 and Algorithm 1, respectively.

Algorithm 1. Processing Flow

1: Create training set T_{train} from D_b, fine-tuning set T_{ft} from $D_b \cup D_n$ and testing set T_{test} from D_n via few-shot detection definition, $D_b \cap D_n = \varnothing$.
2: Initialize the parameters in the Backbone, FSSPN, RPN, and RoI Feature Extractor.
3: **for** each sample $(X_{train}, Y_{train}) \in T_{train}$ **do**
4: Base training.
5: **end for**
6: Frozen the parameters in the Backbone, FSSPN, RPN, and RoI feature extractor, and Reshape the bounding box head of base model.
7: **for** k instance per class $(X_{ft}, Y_{ft}) \in T_{ft}$ **do**
8: Few-shot fine-tuning.
9: **end for**
10: **for** each sample $(X_{test}, Y_{test}) \in T_{test}$ **do**
11: Generate bounding boxes and category scores of aircraft in each image.
12: Calculate the precision of the correctly detected aircrafts.
13: **end for**

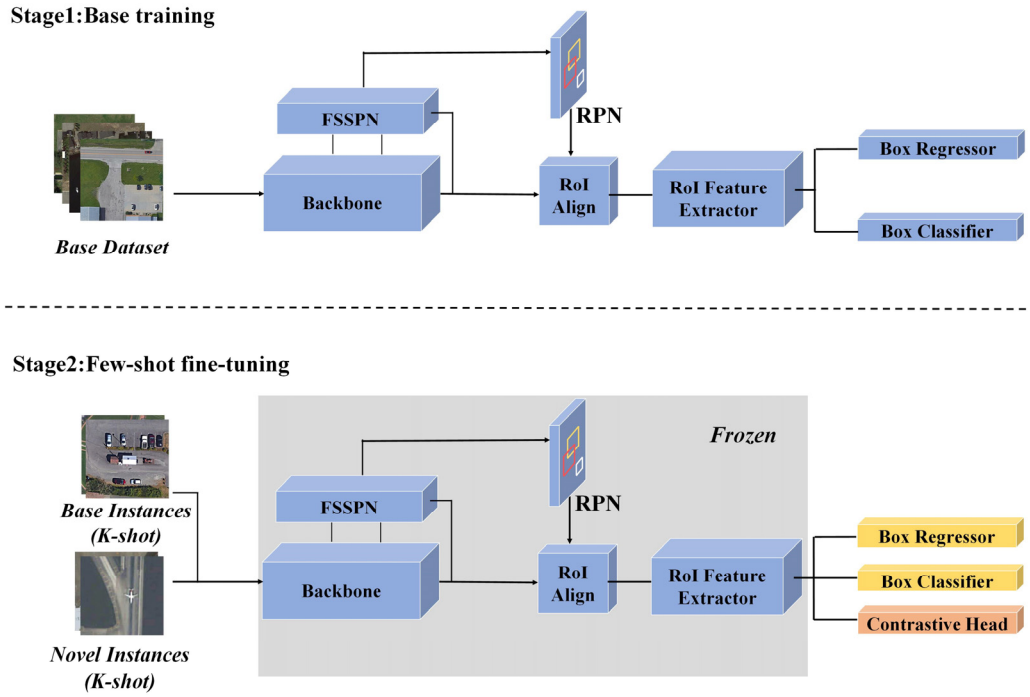

Figure 2. Framework of the proposed method.

Our detection network includes the backbone, FSSPN, RPN, RoI feature extractor, and classification and regression branches. The training process is divided into two stages. In the first stage, a base class set with sufficient annotations is used as input to complete base training. The second stage is fine-tuning, which freezes the weights of the backbone, FSSPN, RPN, and the RoI feature extractor of the network trained in the previous stage, adds a proposal contrastive learning item to the loss function, and selects K objects for each class from the base and the novel class sets to fine-tune the classification and regression branches.

3.3. Feature Scale Selection Pyramid Network

As a widely used feature fusion strategy, FPN [57] can improve the detection performance of multi-scale objects by fusing semantic information about high-level features and location information about low-level features. Since different layers cannot adaptively adjust adjacent data streams in the process of stacking the FPN, there will be inconsistent gradient computation of cross-layers for small-sized objects. This makes the deep features of the objects unable to effectively guide the learning of shallow features and increases

the training burden of the network, thereby weakening the ability to represent fusion features [58].

To improve the detection performance of aircraft in satellite videos, we constructed a FSSPN by adding contextual attention, feature scale enhancement, and feature scale selection [59–62]; these features have been shown to effectively alleviate the inconsistency of gradient computation in the feature fusion of the FPN. This makes the network more suitable for the detection of aircraft of small size in satellite video. The structure of the FSSPN is shown in Figure 3.

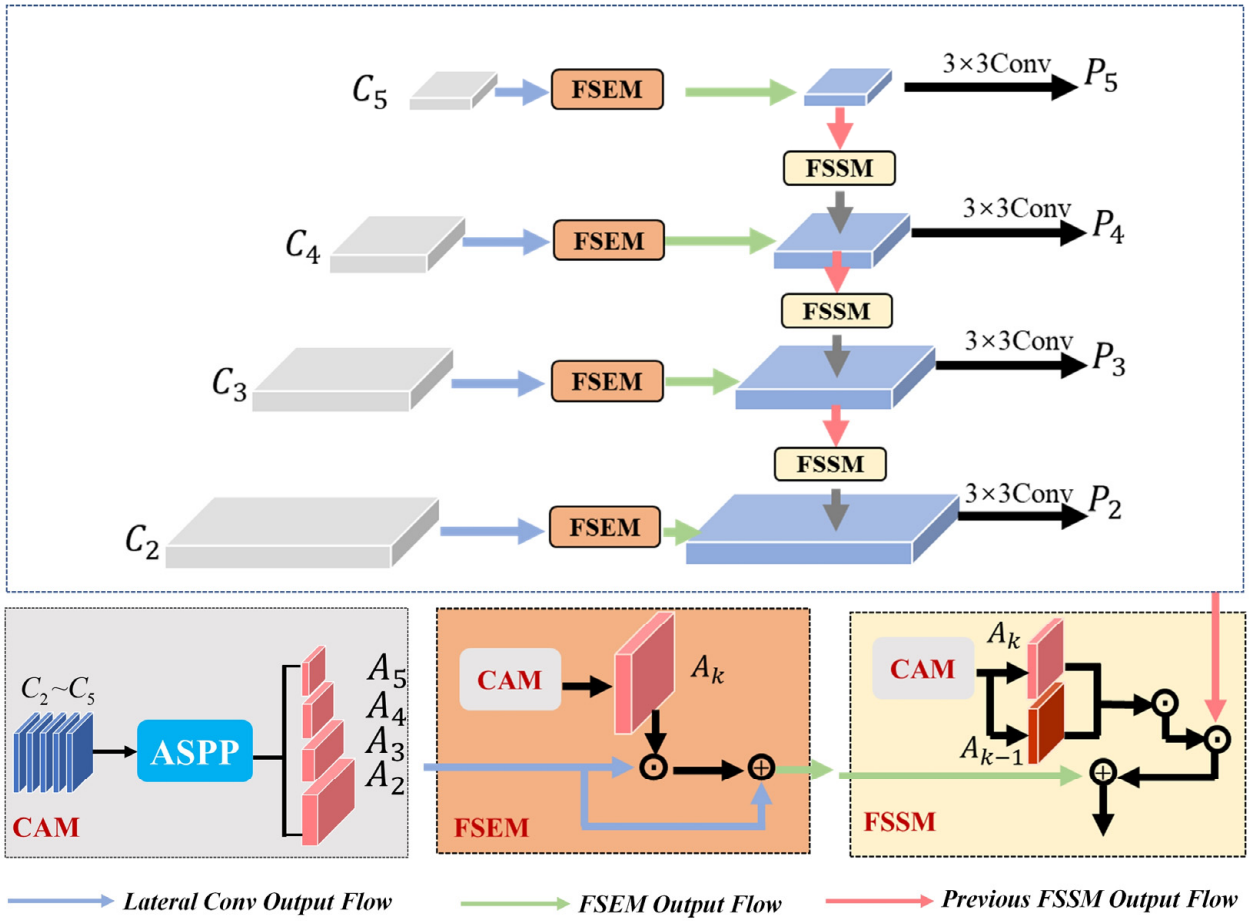

Figure 3. The FSSPN framework.

In the process of feature fusion from different layers, inconsistency of gradient computation often exists in the cross-layers, because the same object will always be considered as a positive sample in the adjacent layers for optimization. In view of this, we constructed the contextual attention module (CAM) to generate hierarchical attention heatmaps in order to indicate the objects in different layers [63–65]. Under the guidance of these hierarchical attention heatmaps, the detector focuses on the specific scales of objects in each layer through the feature scale enhancement module (FSEM). The feature scale selection module (FSSM) takes the intersection of the hierarchical attention heatmaps as guiding information with which to propagate more appropriate features from deep to shallow in the cross-layers gradient computation process. Specifically, we up-sampled the features of the different stages extracted from the backbone to further concatenate in the same size, and then extracted multi-scale features through atrous spatial pyramid pooling (ASPP) [66] to find clues about the objects. The context-aware features generated by ASPP are propagated into

an activation gate, which consists of multiple 3 × 3 convolutions and a sigmoid activation function. We then performed calculations to obtain attention heatmap A_k:

$$A_k = \sigma(\Phi_k(F_c, w, s)) \tag{1}$$

where σ is the sigmoid activation function, Φ_k represents the operation with a 3 × 3 convolution of the kth layer, $\omega \in R^{C_{F\times 1\times 3\times 3}}$ are the convolutional parameters, F_c is the context-aware feature generated by ASPP, and $s = 2^{k-2}$ is the convolutional stride.

Through association with objects matched with multiple hierarchical anchors, the contextual attention at different scales can be generated, thereby highlighting objects at specific scales and avoiding becoming overwhelmed by the background.

To mine the clues of objects at specific scales, scale perception was achieved with different contextual attention to generate scale perception features:

$$F_k^o = (1 + A_k) \odot F_k^i \tag{2}$$

where F_k^i and F_k^o are the input feature map and the output scale perception feature, respectively, and A_k is the contextual attention of the kth layer; $\odot$ represents the element-wise multiplication operation.

By selecting and guiding suitable scale features of objects to propagate to the next layer, the inconsistencies of gradient computation in the cross-layers are alleviated. In addition, when objects can be detected in the adjacent layer, the deeper layer will provide more semantic features and will be optimized simultaneously with the next layer:

$$P'_{k-1} = (A_{k-1} \odot f_{nu}(A_k)) \odot f_{nu}\left(P'_k\right) + C_{k-1} \tag{3}$$

where f_{nu} is the up-sampling operation, P'_k is the merging map of the kth layer, and C_{k-1} is the output of the $(k-1)$th residual block.

By guiding the deep layer to provide suitable features to the shallow layer, the object features from different layers are guaranteed to follow the same gradient direction optimization, thereby improving the detection performance for small aircraft in satellite video.

3.4. Proposal Contrastive Learning

In Faster R-CNN, it is difficult to obtain robust representations of objects in few-shot cases by directly classifying and regressing the features extracted from the RoI, especially for aircraft with poor distinguishability in satellite videos. Inspired by Sun et al. [33], Khosla et al. [67], and Sun et al. [68], we added a proposal contrastive learning item to the loss function in the fine-tuning stage and modeled the intra-class similarity and inter-class difference of the embedded vector from the proposals through batch contrastive learning in order to achieve more robust representation features of aircraft in satellite video. The structure is shown in Figure 4.

Specifically, we first encoded the features extracted from the RoI into a contrastive feature space $z \in R^{D_c}$ through a single-layer perceptron. Next, the similarity score between the encoded RoI features and region proposal representation was measured and the objective to be optimized was added to the loss function. By encoding features into a cosine contrast space, the feature embeddings could be formed into tighter clusters and the distance between different clusters could be enlarged, thereby increasing the generalization of the detector under few-shot conditions. The cosine similarity measure formula is expressed as

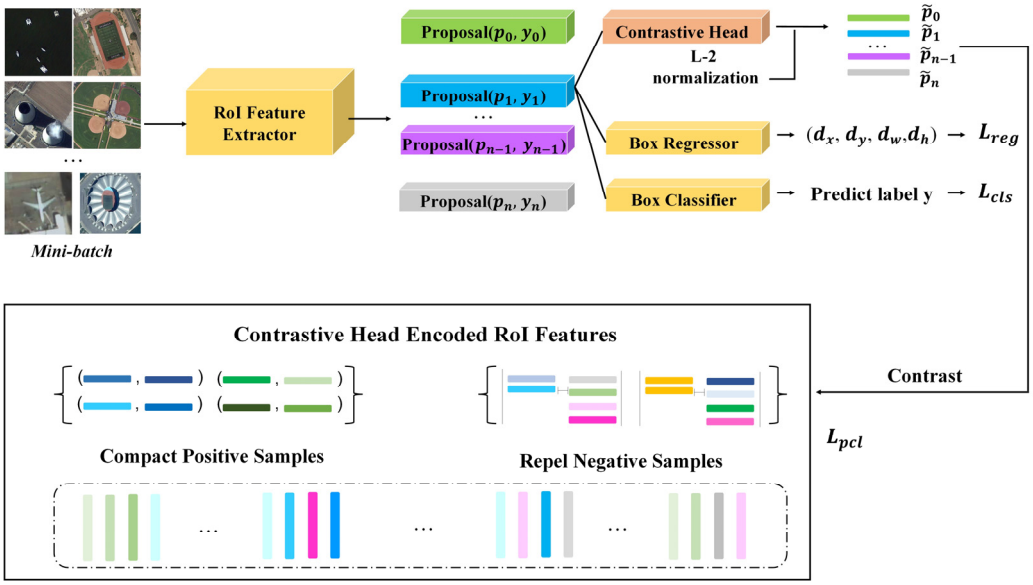

Figure 4. The proposal contrastive learning structure of the network.

$$\log it_{\{i,j\}} = \alpha \frac{x_i^T w_j}{\| x_i \| \cdot \| w_j \|} \tag{4}$$

where w_j is the class weight and α is the scaling factor used to amplify the gradient.

3.5. Loss Function

The few-shot satellite video aircraft detection network design reported in this paper is optimized by a joint loss function, including RPN loss, classification and regression loss of the detection head, attention loss in the FSSPN, and proposal contrastive learning loss:

$$L = L_{RPN} + L_{Head} + L_A + \lambda L_{pcl} \tag{5}$$

where

$$L_{RPN} = \frac{1}{N_{bce}} \sum_{i=1} L_{bce}(rc_i, rc_i^*) + \frac{1}{N_{reg}} \sum_{i=1} L_{reg}(rt_i, rt_i^*) \tag{6}$$

$$L_{Head} = \frac{1}{N_{ce}} \sum_{i=1} L_{ce}(hc_i, hc_i^*) + \frac{1}{N_{reg}} \sum_{i=1} L_{reg}(ht_i, ht_i^*) \tag{7}$$

$$L_A = \alpha L_A^b + \beta L_A^d \tag{8}$$

$$L_{pcl} = \frac{1}{N} \sum_{i=1}^{N} f(u_i) \cdot \frac{-1}{N_{y_i} - 1} \sum_{j=1, j \neq i}^{N} 1\{y_i = y_j\} \cdot \log \frac{\exp(\widetilde{z}_i \cdot \widetilde{z}_j / \tau)}{\sum_{k=1}^{N} 1_{k \neq i} \cdot \exp(\widetilde{z}_i \cdot \widetilde{z}_k / \tau)} \tag{9}$$

For the RPN and head parts of the network, smooth L1 loss is used for bounding box regression, while, for the classification part, binary cross-entropy loss (BCE) is used in L_{RPN} and cross-entropy loss is used in L_{Head}. For L_{RPN}, i represents the ith bounding box in a mini-batch, rc_i and rc_i^* represent the probability distribution of the predicted class and the ground-truth value, respectively, and rt_i and rt_i^* represent the predicted and labeled bounding boxes, respectively.

L_A represents the attention loss, where α and β represent the hyper-parameters of diss loss L_A^b and BCE loss L_A^d, respectively; that is, we use BCE loss to learn poorly classified features.

L_{pcl} represents the proposal contrastive learning loss. For the RoI features of a mini-batch N $\{z_i, u_i, y_i\}_{i=1}^N$, where z_i is the contrast representation embedding of the ith proposal, u_i is the intersection over union (IoU) score, and y_i is the label value. N_{y_i} is the number

of proposals, $\widetilde{z}_i = \frac{z_i}{\|z_i\|}$ denotes the normalized features, and hence, $\widetilde{z}_i \cdot \widetilde{z}_j$ represents the cosine similarity between the ith and jth proposal in the projected hypersphere, where 1 is the indicator function and τ is the regular term.

4. Experiments and Results

We compared the aircraft detection results of the proposed method with those of other advanced few-shot object detection methods. The results of ablation experiments of the proposed method are also presented.

4.1. Experimental Setup

4.1.1. Experimental Data

Base class sets: The DIOR dataset has a total of 22,169 images after excluding the aircraft category, containing 182,524 instances of 19 classes. Object categories include airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, golf course, ground track field, harbor, overpasses, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill. All images measure 800 × 800 pixels in size and have a spatial resolution of 0.5–30 m. The scales of various objects in the dataset are quite different; small objects are most prevalent in the vehicle and ship categories.

Novel class sets: The satellite videos used were all collected from the Jilin-1 satellite. Forty-six videos making up a total of 7086 frames including aircraft scenes were selected, with a framerate of 10 f/s and an average duration of approximately 15 s. For consistency with the images in the DIOR dataset, all videos were cropped to 800 × 800 pixels. According to the definition of video object detection, we fully annotated the aircraft in each frame, for a total of 42,563 instances. In addition to the sufficient labels, the novel class sets also have advantages in terms of diversity and richness. The video scenes included various airports around the world under various levels of illumination, e.g., Beijing Daxing Airport, Italy's Fiumicino Airport, India's Indira Gandhi International Airport, Apolis Sao Paulo International Airport, and Tunisia's Carthage International Airport. The aircraft also exhibited large differences in terms of visual appearance and structure, e.g., civil aircraft, business aircraft, and military aircraft. The videos covered most motion states of aircraft, including high-speed flight, slow taxiing, turning, and stationary. As far as we know, no comparable satellite video aircraft detection labeling dataset exists. We also plan to open source the data for researchers in the remote sensing community as soon as possible after our annotations are further expanded.

4.1.2. Evaluation

After testing the reproducibility of various methods, we selected representative and advanced methods, i.e., Attention RPN [30], FSDetView [32], FSCE [33] and Meta R-CNN [42] for the sake of comparison using the same configuration. In addition, only the precision of aircraft under different few-shot settings (K = 3, 5, 10) were evaluated using all methods; the performance of other categories in the DIOR dataset was not considered. It should be noted that, for different K values, we randomly selected K satellite videos and only selected one aircraft label from them for training; all other labels were used for validation. As mentioned above, due to the small size of the aircraft in the satellite videos, a small offset of the predicted bounding box would have led to large fluctuations in the IoU score [21]; thus, we took IoU = 0.5 as the detection threshold.

4.1.3. Implementation Details

In our experiments, the ResNet-101 pre-trained in ImageNet was used as the backbone network. The optimization strategy adopts stochastic gradient descent (SGD) to perform end-to-end training with a mini-batch size of 4, momentum = 0.9, and weight decay = 0.0001. The learning rate was set to 0.005 and 0.001 during the base training and fine-tuning stages, respectively. All experiments were implemented on a server with an Intel(R) Core(TM) i9 CPU, with 128 GB of memory, and a 2 × RTX3090 GPU (2 × 24 GB).

4.2. Results

4.2.1. Comparison with Other Methods

Table 1 shows the comparison aircraft detection results. The proposed method ranked first or second in terms of precision under different K values (K = 3, 5, 10) and outperformed the other methods overall. As the value of K increased, the precision of all methods improved. It is worth mentioning that when K = 3, Meta R-CNN performed best; however, when the K value was increased to 5 and 10, the improvement of Meta R-CNN was inferior to that of the proposed method. This was due to the fact that meta-learning has certain advantages when using extremely few samples in satellite video scenarios; with a larger number of samples, the precision of Meta R-CNN was approximately 10% worse than that of the proposed method. With different K values, the precision of FSDetView was always the worst. A possible reason for this is that this method simply aggregates the features by concatenating the channel-wise multiplicate and difference of features corresponding to the support and query image. The results from the COCO dataset showed that it does not perform well on small-sized objects; the evaluation results with satellite video further confirmed its limitations. The Attention-RPN method processes the features of the RoI in a purified manner based on attention and multi-relational modules and predicts objects consistent with the class of support based on similarity. From the verification results, we believe that due to the visual differences between high-resolution remote sensing images and satellite video, the domain shift between the base and novel classes increases the difficulty of similarity measurements, resulting in no obvious improvement in precision. The proposed method and the FSCE (baseline) have a similar framework, and the precision of the proposed method always performed better for different K values. On the one hand, the baseline could not propagate suitable features for small-scale objects from deep to shallow to improve the detection performance, in contrast to the method proposed in this paper. On the other hand, unlike the way that the baseline only freezes the parameters of the backbone, the proposed method also freezes the weights of modules, such as the FSSPN, RPN, and RoI feature extractor, and only fine-tunes the detection head of the classification and regression branches. In this regard, we speculate that the conclusion drawn by Sun et al. [30], i.e., that fine-tuning more modules in a network can enhance performance, is not applicable to a domain-shift scene with remote sensing images as the base class and satellite video as the novel class.

Table 1. Comparison of precision @50 of aircraft in satellite videos (%).

Model	Venue	3-Shot	5-Shot	10-Shot
Meta R-CNN	ICCV 2019	15.2	32.3	41.1
FSDetView	ECCV 2020	10.6	22.5	37.5
Attention-RPN	CVPR 2020	11.3	30.2	42.3
FSCE	CVPR 2021	13.5	34.4	47.5
Ours	-	14.7	37.1	51.3

4.2.2. Ablation Studies

To further explore the novelty of the method proposed in this paper, we conducted ablation experiments and analyzed the impact of each step on aircraft detection performance by adding FSEM, FSSE, and expanding freezing parameters (EFP) based on the baseline. The ablation experiment results are shown in Table 2.

Table 2. Ablation experiment results.

Baseline	FSEM	FSSM	FEP	3-Shot	5-Shot	10-Shot
√				13.5	34.4	47.5
√	√			13.6	35.2	48.3
√	√	√		14.4	36.2	50.5
√	√	√	√	14.7	37.1	51.3

It can be seen from the table that the addition of FSEM allowed the network to focus on the specific scales of objects, rather than the broad background; the detection performance was thus improved compared with the baseline for different K values. FSSE achieves more effective multi-level feature fusion for small-sized objects such as aircraft by propagating the appropriate features of the objects from deep to shallow and avoids the negative impact of gradient computation inconsistency in cross-layers. This module also contributes the most to detection performance; with increasing K, the precision improved significantly. By expanding the freezing parameters to adjust the fine-tuning strategy, the influence caused by the difference domain between the base and novel classes can be minimized and the precision can be further improved.

To further analyze the impact of proposal contrastive learning on the performance of the proposed method, we conducted ablation experiments for different hyper-parameters to explore the effect of different contrast head encoding dimensions and temperature τ; and the results are shown in Table 3. As mentioned in Section 3, primary RoI feature vector was truncated at zero after post-ReLU activations. To avoid the inability to meaningfully measure similarities, we encoded the RoI feature with a contrastive head in dimensions of 128 and 256. In addition, in the setting of the temperature τ, we followed the commonly used contrastive objectives [33,66,67] with 0.07, 0.2 and 0.5, respectively. It can be seen from the results that the few-shot detection performance was not sensitive to the contrastive head dimension, while for τ, a moderate value performed relatively better. However, the effect of different contrastive learning hyper-parameters on precision was limited.

Table 3. Ablation for contrastive hyper-parameters, results from 10 shot of the proposed method.

Contrast Head Dimension	**Temperature (τ)**		
	0.07	**0.2**	**0.5**
$D_C = 128$	50.9	51.3	50.8
$D_C = 256$	51.0	51.1	50.7

5. Discussion

5.1. Visualization and Analysis

In view of the baseline and the fact that the proposed method outperformed others in terms of precision, we only compared the visualization results of these two methods and focused on the details in specific scenarios where K = 10. We selected six representative scenes from the satellite videos, covering the main difficulties of the detection task. Figures 5 and 6 show a visualization comparison of the baseline and the proposed method in Figures 5a–c and 6a–c scenarios, respectively. The visualization results showed that both methods successfully detected the aircraft in various scenarios to a certain degree; however, for some difficult and challenging objects, the limitations of few-shot detection were also evident, and problems of missed and false detections were revealed in both methods. Missed detection mainly occurred for extremely small objects in Figures 5c and 6c, as well as for low-contrast objects in Figure 6a and densely distributed objects in Figure 6b. Regarding false detections, the method proposed in this paper performed slightly better than the baseline for road signs in Figure 5a and boarding bridges in Figure 5b,c. This indicates that the baseline confused the structure and shape of aircraft with similar objects, while the proposed method could accurately identify aircraft in these cases.

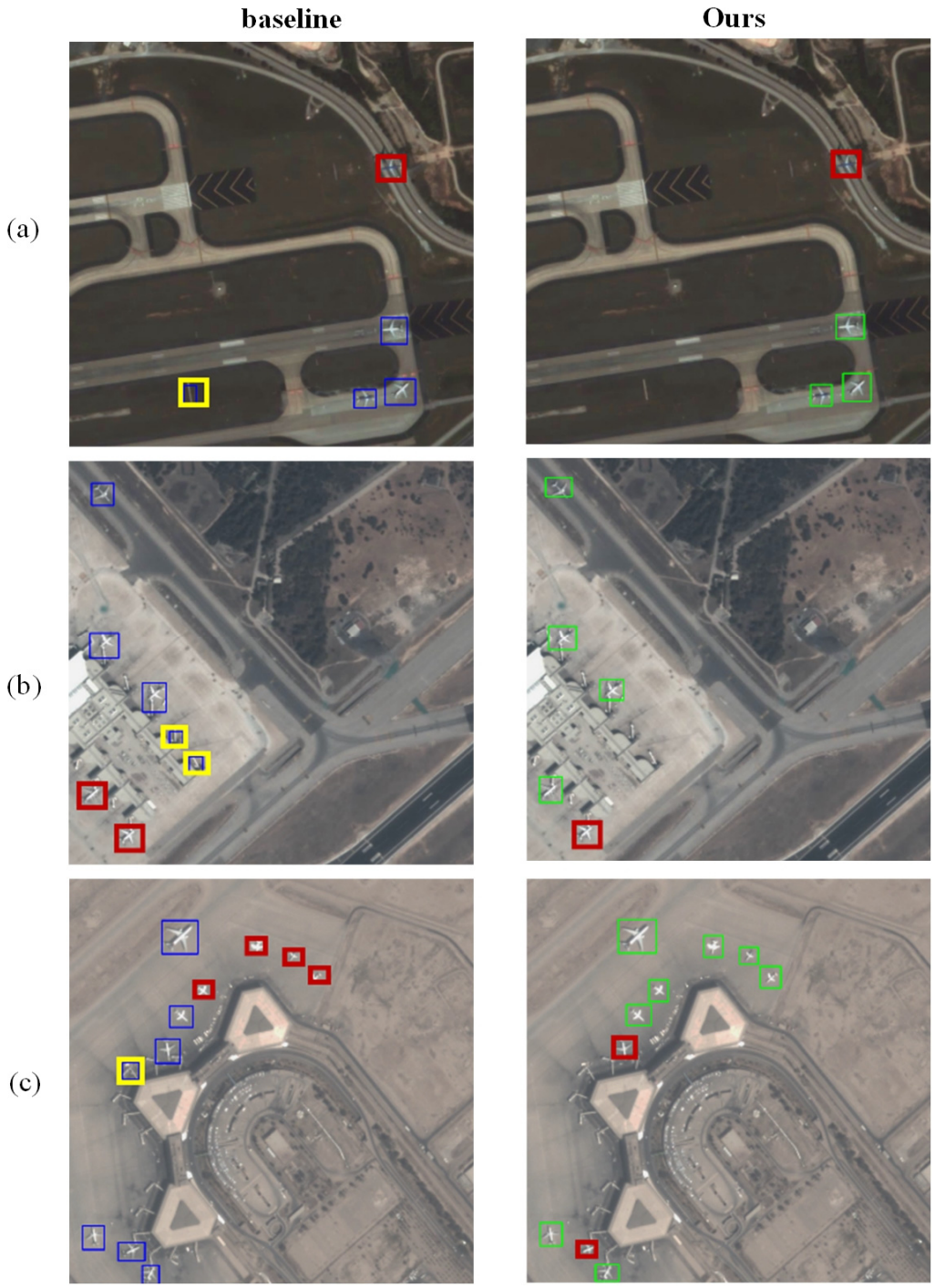

Figure 5. Visualization comparison between the proposed method and the baseline in (**a**) high-speed movement, (**b**) similar object interference, and (**c**) small size scenarios (green, blue, red, and yellow represent the results of the proposed method, the results of the baseline, missed detection, and false detection, respectively).

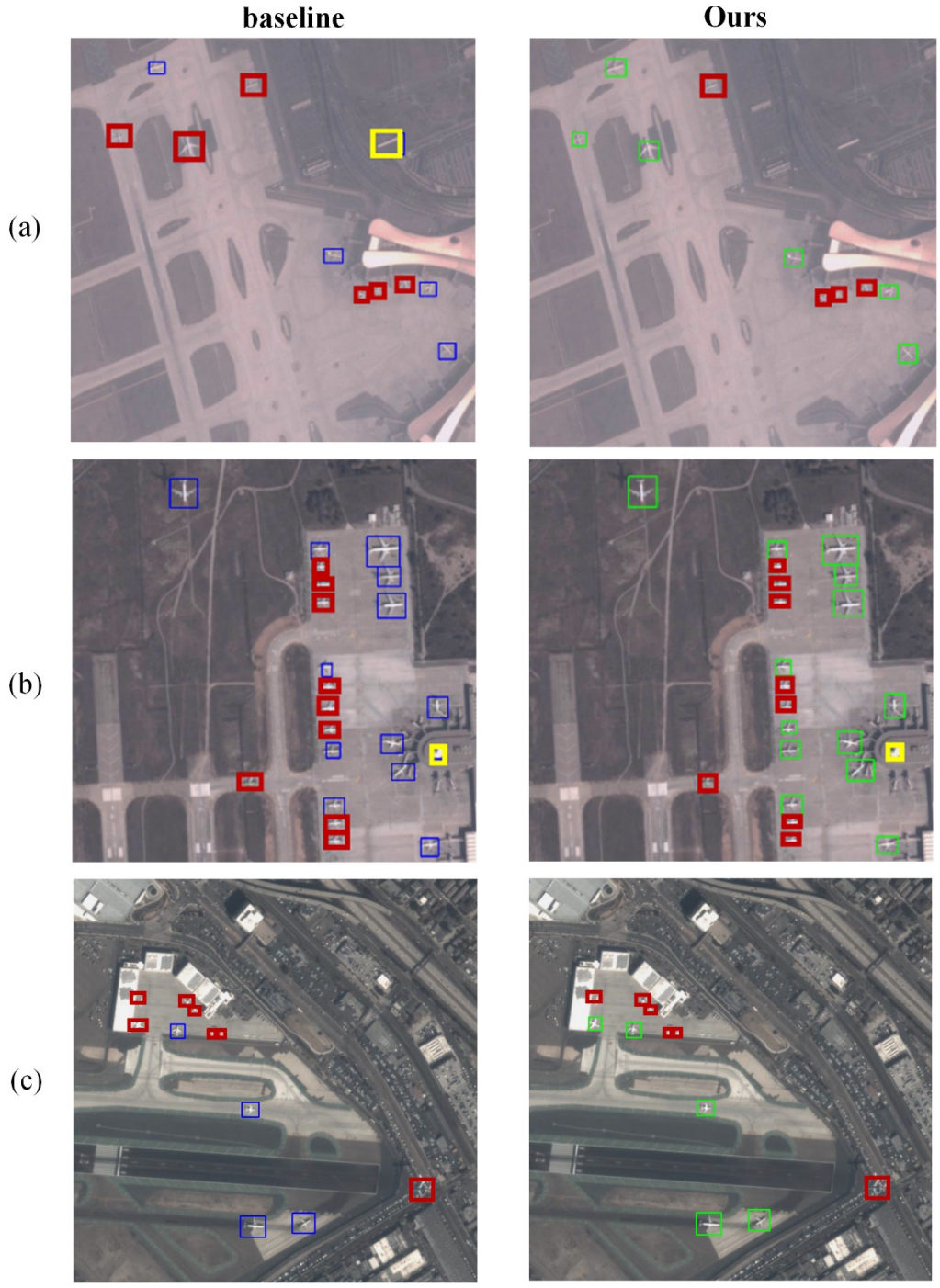

Figure 6. Visualization comparison between the proposed method and the baseline in (**a**) low contrast, (**b**) dense distribution, and (**c**) complex background scenarios (green, blue, red, and yellow represent the results of the proposed method, the results of the baseline, missed detection, and false detection, respectively).

The proposed method was also better in terms of missed detection, especially for the extremely small-sized objects in the Figure 6a–c scenarios, which confirmed the effectiveness of the optimization we implemented for small-sized objects. Although the conventional FPN structure in the baseline method can alleviate the limitation in terms of detecting multi-scale and small-sized objects, its performance is still limited for small objects that are more common in satellite videos.

5.2. Performance between Frames of Video

Since the target task was detection in satellite video, the performance of both the proposed and baseline methods between frames in the same video also needed to be evaluated. To this end, for the high-speed aircraft in scenario Figure 5a, we generated statistics of the detection scores of both methods frame by frame in case of K = 10, as shown in Figure 7.

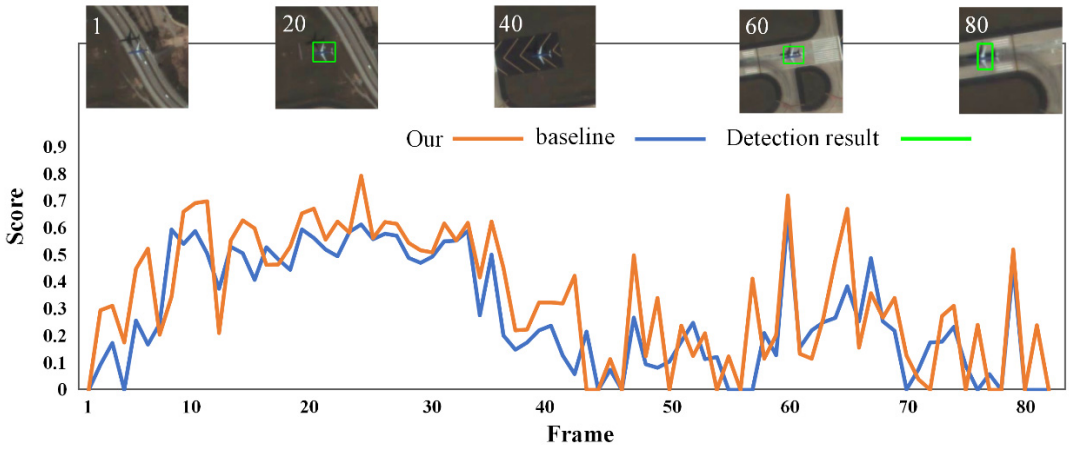

Figure 7. Statistical comparison of detection scores for high-speed aircraft inter-frame.

As the background and the clarity of the high-speed aircraft changed over the course of the video sequence, the scores varied between frames, especially when the background was similarly structured. In cases of similar background textures, e.g., those including roads, both methods failed, but the detection scores of the proposed method exceeded those of the baseline in more than 70% of the sequence frames, indicating that the proposed method offers distinct advantages when it is applied to video streaming data.

The efficiency of satellite video object detection methods needs to be evaluated. Algorithms must be fast enough to detect objects in real-time. We compared the efficiency of the proposed method and the baseline under the same conditions, with framerates (FPS) of 22.9 and 38.1, respectively. Since the proposed method introduces a more complicated computational process in the feature scale selection pyramid, its efficiency was lower than that of the baseline. However, the FPS of both was still much higher than the framerate of Jilin-1 satellite videos, and both achieved the efficiency required for real-time detection.

6. Conclusions

To the best of our knowledge, this paper presents the first application of a few-shot object detection method to satellite video. First, we analyzed the difficulties regarding satellite video aircraft detection under the few-shot condition, and identified the shortcomings of the conventional two-stage fine-tuning framework under the constraints of small size, poor distinguishability, and domain differences in the base and novel classes. We then detailed the construction of an FSSPN to improve Faster R-CNN, which optimized the detection performance of small-scale objects through contextual attention, feature scale enhancement, and feature scale selection in the process of feature fusion. Furthermore, a contrastive learning item was added to the loss function to enhance the ability to identify objects with poor distinguishability. At the same time, to reduce the training difficulties caused by the domain differences between the base and novel classes, we adjusted the parameter freezing strategy and only fine-tuned the detection head of the classification and regression branches. Finally, the proposed method was compared with various advanced methods. Our experiments showed that the proposed method achieved the best performance with satellite video. Currently, this method only considers visual information in satellite videos.

In the future, we plan to use timing information to further improve the performance of few-shot object detection in satellite videos.

Author Contributions: Conceptualization, Z.Z. and S.L.; methodology, Z.Z. and S.L.; software, W.G.; validation, Z.Z. and W.G.; formal analysis, Z.Z.; investigation, W.G.; resources, S.L.; data curation, Z.Z. and W.G.; writing—original draft preparation, Z.Z.; writing—review and editing, S.L. and Y.G.; visualization, Z.Z.; supervision, S.L.; project administration, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Director's Foundation of Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences under Grant CSU-JJKT-2020-09.

Data Availability Statement: The DIOR data set used as base class is publicly open data and the access manner of the data can refer to the corresponding published papers. The satellite videos with annotations will be released as soon as possible after it is further annotated, which are used for visual tasks of satellite videos such as object detection, instance segmentation, and object tracking.

Acknowledgments: The authors would like to thank all the researcher who kindly shared the data and codes.

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

Abbreviations

The following abbreviations are used in this manuscript:

FSSPN	Feature scale selection pyramid network
FPN	Feature pyramid network
RPN	Region proposal network
RoI	Region of interest
CAM	Contextual attention module
FSEM	Feature scale enhancement module
FSSM	Feature scale selection module
ASPP	Atrous spatial pyramid pooling
IoU	Intersection over union
SGD	Stochastic gradient descent
EFP	Expanding freezing parameters
FPS	Frames per second

References

1. Xuan, S.; Li, S.; Han, M.; Wan, X.; Xia, G.-S. Object Tracking in Satellite Videos by Improved Correlation Filters with Motion Estimations. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 1074–1086. [CrossRef]
2. Gu, Y.; Wang, T.; Jin, X.; Gao, G. Detection of Event of Interest for Satellite Video Understanding. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7860–7871. [CrossRef]
3. Shao, J.; Du, B.; Wu, C.; Zhang, L. Can We Track Targets from Space? A Hybrid Kernel Correlation Filter Tracker for Satellite Video. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8719–8731. [CrossRef]
4. 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]
5. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE international conference on computer vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
6. 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.
7. Gong, H.; Mu, T.; Li, Q.; Dai, H.; Li, C.; He, Z.; Wang, W.; Han, F.; Tuniyazi, A.; Li, H. Swin-Transformer-Enabled YOLOv5 with Attention Mechanism for Small Object Detection on Satellite Images. *Remote Sens.* **2022**, *14*, 2861. [CrossRef]
8. Li, Z.; Wang, Y.; Zhang, N.; Zhang, Y.; Zhao, Z.; Xu, D.; Ben, G.; Gao, Y. Deep Learning-Based Object Detection Techniques for Remote Sensing Images: A Survey. *Remote Sens.* **2022**, *14*, 2385. [CrossRef]
9. Wu, X.; Hong, D.; Tian, J.; Chanussot, J.; Li, W.; Tao, R. ORSIm Detector: A Novel Object Detection Framework in Optical Remote Sensing Imagery Using Spatial-Frequency Channel Features. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5146–5158. [CrossRef]
10. Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; Sun, X. Rotation-Aware and Multi-Scale Convolutional Neural Network for Object Detection in Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 294–308. [CrossRef]
11. Cheng, G.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]

12. Han, J.; Ding, J.; Li, J.; Xia, G.-S. Align Deep Features for Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–11. [CrossRef]
13. Lei, J.; Luo, X.; Fang, L.; Wang, M.; Gu, Y. Region-Enhanced Convolutional Neural Network for Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5693–5702. [CrossRef]
14. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-Scale Object Detection in Remote Sensing Imagery with Convolutional Neural Networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
15. Wu, B.; Shen, Y.; Guo, S.; Chen, J.; Sun, L.; Li, H.; Ao, Y. High Quality Object Detection for Multiresolution Remote Sensing Imagery Using Cascaded Multi-Stage Detectors. *Remote Sens.* **2022**, *14*, 2091. [CrossRef]
16. Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
17. Ding, J.; Xue, N.; Xia, G.-S.; Bai, X.; Yang, W.; Yang, M.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef]
18. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and a New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
19. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-Class Geospatial Object Detection and Geographic Image Classification Based on Collection of Part Detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
20. Zhao, M.; Li, S.; Xuan, S.; Kou, L.; Gong, S.; Zhou, Z. SatSOT: A Benchmark Dataset for Satellite Video Single Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5617611. [CrossRef]
21. Yin, Q.; Hu, Q.; Liu, H.; Zhang, F.; Wang, Y.; Lin, Z.; An, W.; Guo, Y. Detecting and Tracking Small and Dense Moving Objects in Satellite Videos: A Benchmark. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5612518. [CrossRef]
22. Xuan, S.; Li, S.; Zhao, Z.; Zhou, Z.; Zhang, W.; Tan, H.; Xia, G.; Gu, Y. Rotation Adaptive Correlation Filter for Moving Object Tracking in Satellite Videos. *Neurocomputing* **2021**, *438*, 94–106. [CrossRef]
23. Sun, X.; Wang, B.; Wang, Z.; Li, H.; Li, H.; Fu, K. Research Progress on Few-Shot Learning for Remote Sensing Image Interpretation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2387–2402. [CrossRef]
24. Zeng, Q.; Geng, J. Task-Specific Contrastive Learning for Few-Shot Remote Sensing Image Scene Classification. *ISPRS J. Photogramm. Remote Sens.* **2022**, *191*, 143–154. [CrossRef]
25. Zheng, X.; Gong, T.; Li, X.; Lu, X. Generalized Scene Classification from Small-Scale Datasets with Multitask Learning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5609311. [CrossRef]
26. Gong, T.; Zheng, X.; Lu, X. Meta Self-Supervised Learning for Distribution Shifted Few-Shot Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6510005. [CrossRef]
27. 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.
28. Perez-Rua, J.-M.; Zhu, X.; Hospedales, T.M.; Xiang, T. Incremental Few-Shot Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13846–13855.
29. Li, A.; Li, Z. Transformation Invariant 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. 3094–3102.
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, 14–19 June 2020; pp. 4013–4022.
31. Wang, X.; Huang, T.E.; Darrell, T.; Gonzalez, J.E.; Yu, F. Frustratingly Simple Few-Shot Object Detection. *arXiv* **2020**, arXiv:2003.06957.
32. Xiao, Y.; Marlet, R. Few-Shot Object Detection and Viewpoint Estimation for Objects in the Wild. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland; pp. 192–210.
33. 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.
34. Wu, J.; Liu, S.; Huang, D.; Wang, Y. Multi-Scale Positive Sample Refinement for Few-Shot Object Detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany; pp. 456–472.
35. Ren, X.; Zhang, W.; Wu, M.; Li, C.; Wang, X. Meta-YOLO: Meta-Learning for Few-Shot Traffic Sign Detection via Decoupling Dependencies. *Appl. Sci.* **2022**, *12*, 5543. [CrossRef]
36. 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, Korea, 21 October–5 November 2019; pp. 8420–8429.
37. Li, X.; Deng, J.; Fang, Y. Few-Shot Object Detection on Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5601614. [CrossRef]
38. Huang, X.; He, B.; Tong, M.; Wang, D.; He, C. Few-Shot Object Detection on Remote Sensing Images via Shared Attention Module and Balanced Fine-Tuning Strategy. *Remote Sens.* **2021**, *13*, 3816. [CrossRef]
39. Zhao, Z.; Tang, P.; Zhao, L.; Zhang, Z. Few-Shot Object Detection of Remote Sensing Images via Two-Stage Fine-Tuning. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]

40. Zhou, Y.; Hu, H.; Zhao, J.; Zhu, H.; Yao, R.; Du, W.-L. Few-Shot Object Detection via Context-Aware Aggregation for Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8021805. [CrossRef]
41. 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, 21 October–5 November 2019; pp. 9925–9934.
42. 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, Korea, 21 October–5 November 2019; pp. 9577–9586.
43. Quan, J.; Ge, B.; Chen, L. Cross Attention Redistribution with Contrastive Learning for Few Shot Object Detection. *Displays* **2022**, *72*, 102162. [CrossRef]
44. Cheng, M.; Wang, H.; Long, Y. Meta-Learning-Based Incremental Few-Shot Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 2158–2169. [CrossRef]
45. Zhang, G.; Luo, Z.; Cui, K.; Lu, S. Meta-Detr: Few-Shot Object Detection via Unified Image-Level Meta-Learning. *arXiv* **2021**, arXiv:2103.11731.
46. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A.A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A. Overcoming Catastrophic Forgetting in Neural Networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526. [CrossRef]
47. Karlinsky, L.; Shtok, J.; Harary, S.; Schwartz, E.; Aides, A.; Feris, R.; Giryes, R.; Bronstein, A.M. Repmet: Representative-Based Metric Learning for Classification and Few-Shot Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5197–5206.
48. Zhang, T.; Zhang, Y.; Sun, X.; Sun, H.; Yan, M.; Yang, X.; Fu, K. Comparison Network for One-Shot Conditional Object Detection. *arXiv* **2019**, arXiv:1904.02317.
49. 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*, 2725–2734.
50. Lu, Y.; Chen, X.; Wu, Z.; Yu, J. Decoupled Metric Network for Single-Stage Few-Shot Object Detection. *IEEE Trans. Cybern.* **2022**, 1–12. [CrossRef] [PubMed]
51. Cheng, G.; Yan, B.; Shi, P.; Li, K.; Yao, X.; Guo, L.; Han, J. Prototype-CNN for Few-Shot Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5604610. [CrossRef]
52. Wang, Y.; Xu, C.; Liu, C.; Li, Z. Context Information Refinement for Few-Shot Object Detection in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3255. [CrossRef]
53. Chen, H.; Wang, Y.; Wang, G.; Qiao, Y. Lstd: A Low-Shot Transfer Detector for Object Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
54. 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 European conference on computer vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany; pp. 740–755.
55. Chen, G.; Wang, H.; Chen, K.; Li, Z.; Song, Z.; Liu, Y.; Chen, W.; Knoll, A. A Survey of the Four Pillars for Small Object Detection: Multiscale Representation, Contextual Information, Super-Resolution, and Region Proposal. *IEEE Trans. Syst. man Cybern. Syst.* **2020**, *52*, 936–953. [CrossRef]
56. Xiao, A.; Wang, Z.; Wang, L.; Ren, Y. Super-Resolution for "Jilin-1" Satellite Video Imagery via a Convolutional Network. *Sensors* **2018**, *18*, 1194. [CrossRef]
57. 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.
58. Liu, S.; Huang, D.; Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. *arXiv* **2019**, arXiv:1911.09516.
59. Zhang, P.; Wang, L.; Wang, D.; Lu, H.; Shen, C. Agile Amulet: Real-Time Salient Object Detection with Contextual Attention. *arXiv* **2018**, arXiv:1802.06960.
60. Wang, T.; Anwer, R.M.; Khan, M.H.; Khan, F.S.; Pang, Y.; Shao, L.; Laaksonen, J. Deep Contextual Attention for Human-Object Interaction Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–5 November 2019; pp. 5694–5702.
61. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122.
62. Hong, M.; Li, S.; Yang, Y.; Zhu, F.; Zhao, Q.; Lu, L. SSPNet: Scale Selection Pyramid Network for Tiny Person Detection from UAV Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8018505. [CrossRef]
63. Zhang, J.; Shi, Y.; Zhang, Q.; Cui, L.; Chen, Y.; Yi, Y. Attention Guided Contextual Feature Fusion Network for Salient Object Detection. *Image Vis. Comput.* **2022**, *117*, 104337. [CrossRef]
64. Zhang, J.; Xie, C.; Xu, X.; Shi, Z.; Pan, B. A Contextual Bidirectional Enhancement Method for Remote Sensing Image Object Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4518–4531. [CrossRef]
65. Fan, B.; Shao, M.; Li, Y.; Li, C. Global Contextual Attention for Pure Regression Object Detection. *Int. J. Mach. Learn. Cybern.* **2022**, *13*, 2189–2197. [CrossRef]

66. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
67. 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.
68. Sun, Y.; Chen, Y.; Wang, X.; Tang, X. Deep Learning Face Representation by Joint Identification-Verification. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1988–1996.