electronics

MDPI

Article

F^{2}SOD: A Federated Few-Shot Object Detection

Peng Li [1,2], Tianyu Zhang [3], Chen Qing [1,2] and Shuzhuang Zhang [3,*]

[1] National Key Laboratory of Complex Aviation System Simulation, Chengdu 610036, China; lipeng27@cetc.com.cn (P.L.); qingchen_1@cetc.com.cn (C.Q.)
[2] Southwest China Institute of Electronic Technology, Chengdu 610036, China
[3] The School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876, China; zty0756@bupt.edu.cn
* Correspondence: zhangshuzhuang@bupt.edu.cn

Abstract: With the popularity of edge computation, object detection applications face challenges of limited data volume and data privacy. To address these, we propose a federated few-shot object detection framework, F^{2}SOD. It involves three steps: collaborative base model training with base class data, novel data augmentation via an improved diffusion model, and collaborative base model fine-tuning for novel model using augmented data. Specifically, we present a data augmentation method based on diffusion models with a twofold-tag prompt construction and object location embedding. In addition, we present distributed framework for training base and novel models, where the base model integrates the Squeeze-and-Excitation attention mechanism into the feature re-weighting module. Experiments on public datasets demonstrate that F^{2}SOD achieves efficient few-shot object detection, outperforming State-of-the-Art methods in both accuracy and efficiency.

Keywords: federated learning; few shot; object detection

1. Introduction

Object detection, which entails the assignment of class labels to an object and the delineation of bounding box around each object of interest, has emerged prominently as a crucial and deeply investigated topic within the domain of computer vision [1,2]. There is a diverse array of objects detection applications across various industries, such as autonomous driving [3,4], robot vision [5], and environmental monitoring [6].

In the past few years, deep learning advancements have greatly accelerated object detection progress, leading to remarkable breakthroughs [7]. Compared to traditional methods, models with deeper architectures operate on more complex representations. However, the increased depth of these models comes at a cost: their performance is heavily dependent on large amounts of data. Moreover, collecting such extensive datasets can be costly and may even be both costly and impractical in certain scenarios, such as the diagnosis of rare diseases and the surveillance of endangered wildlife. As a result, training these models on limited datasets often leads to overfitting issues.

To effectively operate within a limited data regime, few-shot object detection (FSOD) has emerged as a crucial area of research. FSOD alleviates the reliance on large amount of labeled training data, primarily emphasizing the improvement of performance in novel classes rather than maintaining performance in base classes [8,9]. In the field of FSOD, the goal is to detect specific objects using only a limited number of annotated samples, thereby eliminating the need for extensive annotated data. The models employed in FSOD undergo pre-training using a substantial amount of data from known categories, referred

check for updates

Academic Editors: Javid Taheri and Stefano Scanzio

Received: 11 March 2025
Revised: 13 April 2025
Accepted: 17 April 2025
Published: 19 April 2025

Citation: Li, P.; Zhang, T.; Qing, C.; Zhang, S. F^{2}SOD: A Federated Few-Shot Object Detection. *Electronics* **2025**, *14*, 1651. https://doi.org/10.3390/electronics14081651

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

to as base classes, to understand object knowledge such as features, shapes, and patterns. Subsequently, the knowledge acquired from base class data equips the model to recognize new classes, denoted as novel classes, with limited annotated data.

FSOD holds great significance in diverse practical applications. To date, FSOD approaches can be classified into four groups: those based on meta-learning, metric learning, transfer learning, and data augmentation. Meta-learning-based algorithms employ feature reweighting mechanisms or analogous strategies to fuse query-support feature representations, addressing FSOD challenges through adaptive feature integration.Metric-learning-based methods focus on learning discriminative feature representations with limited data. Transfer learning-based techniques enable models to leverage previously acquired knowledge from base classes for the recognition and detection of novel classes. Data-augmentation-based approaches expand the limited novel data through various transformation techniques. These methods effectively extract, learn, and transfer the knowledge of base class data from different perspectives to expand the information or data volume of novel class data, achieving high-performance object detection on novel class data.

With the widespread popularity of edge computation, data generated by terminal devices serves as the critical source for object detection, which aim to provide personalized and accurate services. On the one hand, as mentioned above, the limited data volume in some scenarios fails to meet the requirements of effective object detection. On the other hand, this data carries user privacy information. Strict data protection laws and regulations restrict the free usage of this data, hindering its use in training models for object detection. Faced with limit and privacy data from intelligent terminals, how to achieve object detection with profound performance to make it suitable for real-world applications.

To address the above problem, we propose a framework for federated few-shot object detection, called as F^2SOD. F^2SOD includes the three steps. Firstly, clients collaboratively train the base model of object detection by their base class data based on the framework of federated learning. Secondly, each client augments its novel data via improved diffusion model. Finally, clients utilize their augmented data to collaboratively fine-tune the base model. Our overall contributions are summarized as follows.

1. In order to solve the challenges of data privacy and limited sample sizes faced in object detection, we present the framework for federated few-shot object detection. This framework not only facilitates efficient object detection tasks with a scarce amount of data but also guarantees that the original data stays on local devices.
2. In order to enhance the understanding of base data, we integrate the Squeeze-and-Excitation (SE) attention mechanism into the feature re-weighting module for each client. Moreover, we reconstruct the localization loss term in the loss function to boost the learning capability for small objects.
3. To improve the adaptability of models to novel data in few-shot scenarios, we propose a data augmentation method based on diffusion models. This method fine-tunes the diffusion model using a twofold-tag prompt and location information embedding. It is devised to generate a wide variety of data, effectively expanding the scale and improving the quality of the novel dataset.
4. The experiments conducted on the public datasets demonstrate that F^2SOD can realize more efficient few-shot object detection as the number of participants increases or the effective volume of small samples from clients goes up. Moreover, when compared with the State-of-the-Art approaches, F^2SOD outperforms them in terms of both accuracy and efficiency.

2. Related Works

In this section, we provide a comprehensive review of the advancements in object detection and few-shot object detection (FSOD).

2.1. Object Detection

Object detection fundamentally operates as a dual-task framework requiring precise localization of object instances through bounding box regression coupled with accurate category recognition. In this review, we primarily focus on the advancements achieved in CNN-based object detectors.

The CNN-based object detectors can be classified into two categories: two-stage detectors and single-stage detectors. The workflow of two-stage detectors [10] initiates with feature extraction via backbone networks, producing multi-scale hierarchical representations. These feature maps subsequently undergo region proposal generation through dedicated subnetworks that output preliminary bounding box coordinates. The secondary phase ingests these region candidates, simultaneously executing categorical classification and bounding box refinement through dedicated regression layers. The primary two-stage approaches include R-CNN [11], SPPNet [12], Fast R-CNN [13], and Faster R-CNN [14].

Single-stage CNN-based object detectors emerge to circumvent computational bottlenecks inherent in two-stage systems, prioritizing computational efficiency for real-time deployment. Through hierarchical feature extraction, these networks deploy parallel detection heads that concurrently estimate categorical probabilities and spatial coordinates across multiple scales. The YOLO (You Only Look Once) framework [15] pioneered grid-based detection through spatial grid discretization, where each unit simultaneously predicts bounding box parameters and class confidence scores. Subsequent methodological refinements of this architecture, including YOLOv7 [16], have systematically enhanced multi-scale recognition capabilities while preserving real-time processing constraints.

Here, our approach is categorized within the technical framework of single-stage object detection.

2.2. Few-Shot Object Detection

FSOD approaches are classified into four families.

Meta-learning-based approaches. Meta-learning-based approaches helps the model grasp general patterns, facilitating rapid adaptation and generalization to novel tasks. Kang et al. [17] proposed YOLO-FR based on a single-stage YoLo-v2 object detection framework, which utilizes meta-training to distill transferable meta-features from base class annotations. Wang et al. [18] is a two-stage training scheme, which trains a Faster R-CNN with the addition of a instance-level feature normalization. Wang et al. [19] builds on the YOLOv3 framework for remote sensing object detection with minimal annotated samples. Wang et al. [20] introduced a Meta-Det approach through parameter decoupling strategy, which isolates category-agnostic components from category-sensitive parameters in meta-optimized detectors. Yan et al. [8] constructed Meta-RCNN as unified meta-learning framework with joint optimization scheme for detection and segmentation. Lee et al. [21] introduced the concept of Attending to Per-Sample-Prototype (APSP), which preserves per-sample embedding distinctiveness through dynamic prototype aggregation. Han et al. [22] introduced Meta Faster R-CNN, which enhances knowledge transfer pathways by replacing the conventional Region Proposal Network (RPN) classifier with a feature reweighting module through architectural modifications and employing a metric-based classifier. Zhang et al. [23] formulated support-query mutual guidance (SQMGH) through bidirectional cross-modal guidance mechanism employing cross-attention feature alignment with multi-scale proposal correlation. Li et al. [24] introduced the Meta RetinaNet

methodby integrating task-conditioned feature recalibration layers with meta-anchors for scale-aware few-shot detection.

Matric-learning-based methods. Matric-learning-based methods learn the difference between similar and dissimilar objects with limited data. Karlinsky et al. [25] developed RepMet with representative-constrained metric learning that optimizes class-specific prototype distributions through distance-based clustering regularization. Lu et al. [26] proposed DMNet with unified single-stage architecture integrating dual-branch representation decoupling (DRT) and hierarchical metric alignment (IDML) for joint feature disentanglement. Li et al. [27] engineered cross-domain adaptation framework for aerial detection by embedding multi-similarity metric constraints into Faster R-CNN's region proposal mechanism. Leng et al. [28] formulated Sampling-Invariant Metric Network (SIF-Net) with density-aware prototype learning that ensures prototype invariance through distribution-aware sampling strategies. Chen et al. [29] constructed hybrid detection architecture integrating feature confusion regularization via DropBlock-enhanced generalization with deformable convolution-based geometric adaptation modules.

Transfer-learning-based methods. Transfer-learning-based methods leverage knowledge and features learned from large datasets and then transfer this knowledge to the task of detecting objects with merely a handful of training examples. Chen et al. [30] architected cross-domain knowledge distillation module that aligns base-class semantic distributions with target-domain proposals through adversarial feature alignment. Li et al. [31] devised Class Margin Equilibrium (CME) with equilibrium-driven feature space partitioning and novel class manifold regularization via margin-aware contrastive learning. Kim et al. [9] introduced a new FSOD approach that leverages knowledge transfer from base classes to detect objects with few training examples for novel classes. Li et al. [32] proposed a knowledge distillation-based FSOD approach to leverage semantic information from large-scale pre-trained models. Ma et al. [33] employed an offline ETF classifier for well-separated class centers and adaptive margins to tighten feature clusters, improving novel class generalization without compromising base class performance. Zhu et al. [34] constructed knowledge transfer framework integrating augmentation-invariant prototype learning with uncertainty-aware feature alignment. Wang et al. [35] developed Graph Knowledge Transfer (CKT) with dynamic inter-class correlation modeling through graph neural networks and intra-class diversity preservation via feature dispersion constraints.

Data-augment-based approaches. Data-augment-based approaches enlarge the novel set so that it becomes sufficiently large to offer reliable hypotheses. Sun et al. [36] developed contrastive training framework employing IoU-aware proposal sampling to construct hardness-adaptive positive pairs for representation robustness enhancement. Wu et al. [37] devised Multi-scale Positive Sample Refinement (MPSR) through hierarchical feature aggregation via Feature Pyramid Network(FPN) integration in Faster R-CNN, establishing scale-robust feature learning through pyramidal proposal matching. Wu et al. [38] constructed TD Sampler with adaptive curriculum learning that implements progressive difficulty scheduling through dynamic sample weighting. Yan et al. [39] proposed Uncertainty-aware Proposal Optimization (UNP) comprising Confusing Proposals Separation (CPS) via proposal affinity analysis using intersection-over-union (IoU) metric, coupled with Affinity-Driven Gradient Relaxation (ADGR) through affinity-adaptive gradient modulation based on prototype similarity measures. Wang et al. [40] formulated Semantic-aware Neural Implicit Data Augmentation (SNIDA) through foreground-background decomposition with disentangled neural rendering, enabling contextual diversity amplification via compositional augmentation strategies.

On the one hand, most of the few-shot object detection algorithms based on data augmentation adopt traditional data augmentation methods (such as geometric transformation,

color alterations, and mixup), model-based feature space augmentation (feature mixing, feature perturbation). These methods still have obstacles in terms of data diversity and consistency, which hinders the performance of object detection.

On the other hand, existing research on distributed object detection has not considered the few-shot scenario. For instance, Huang et al. [41] proposed a federated learning-driven cross-spatial vessel detection model (FLCSDet), which incorporates an efficient multi-scale attention module into FLCSDet. Chi et al. [42] introduced a federated cooperative learning framework that combines local cooperative perception with global federated learning through a parameter-efficient federated learning adapter (PEFLA) and a lazy communication (LazyComm) strategy, improving DL-based object detection capabilities across diverse driving scenarios. Behara et al. [43] introduced large model-assisted federated learning for object detection of autonomous vehicles in edge that organizes devices into a hierarchical structure and adopts lightweight models to mitigate prolonged training and prediction duration. Nevertheless, these methods cannot be directly transferred to the distributed few-shot object detection scenario.

To overcome these limitations, we propose F^{2}SOD: a federated few-shot object detection framework that leverages federated learning to aggregate knowledge from base class models and new class models across various clients. Furthermore, a diffusion model-driven data augmentation is used to enhance both the scale and diversity of novel datasets, thereby improving the learning capabilities of few shots.

3. F^{2}SOD: A Federated Few-Shot Object Detection

3.1. Problem Formulation

In intelligent application scenarios such as smart health, intelligent transportation and remote sensing, each intelligent terminal generates diverse data that may vary in labels and volumes, and some labels have significantly fewer instances. Furthermore, these data often contain information that is sensitive and private to clients.

Suppose that there are m intelligent terminals, denoted as $C_0, C_1, \cdots, C_{m-1}$. The data associated with terminal C_i is represented by D_i, which can be partitioned into D_i^{base} and D_i^{novel} based on the number of labels. Specifically,

$$D_i^{base} \cup D_i^{novel} = D_i, \tag{1}$$

$$D_i^{base} \cap D_i^{novel} = \Phi, \tag{2}$$

where the quantity of data associated with each novel class D_i^{novel} is small, which is characterized as few-shot, i.e.,

$$|y_{i,j} \in D_i^{novel}\}| < k, \tag{3}$$

where $y_{i,j}$ represents the j-th label of D_i^{novel} for the i-th terminal, and k is a small natural number indicating the few shot.

Although the data between the base dataset and the novel dataset for each terminal are completely different, the data among the base (novel) datasets of these terminals may overlap. This naturally raises the question: faced with few-shot and privacy data of each terminal, how can we achieve effective object detection tasks by learning and extracting the common knowledge?

3.2. F^{2}SOD Algorithm

In this paper, we propose the F^{2}SOD: federated few-shot object detection algorithm, as shown in Figure 1.

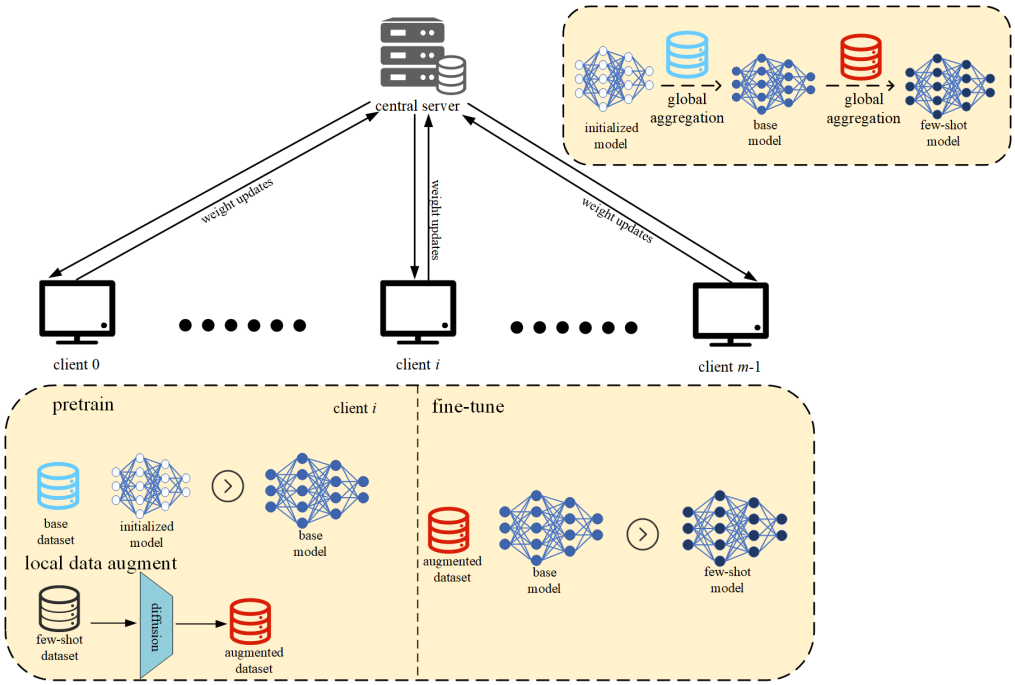


Figure 1. The framework of our F^{2}SOD.

Our F^{2}SOD includes pre-training, data augment, and fine-tuning.

Pretraining costs large computation resources but can be conducted offline. Both data augmentation and fine-tuning are online and efficient, requiring limited computational resources.

- **Pre-training.** This is the training stage of the base model for object detection. Clients collaboratively train the base model by their base sets, and get a global base model M_{base}.
- **Data Augment.** Before performing fine-tuning, each client augments its few-shot novel set based on the diffusion model. As a result, the augmented datasets $\tilde{D}^{novel}$ are obtained in this stage.
- **Fine-tuning.** Clients feed their locally augmented datasets $\tilde{D}^{novel}$ into the base model M_{base} for cooperative fine-tuning, resulting in the final fine-tuned model M_{novel}.

Pre-training Here, the aim of pre-training is to develop a base model of object detection in distributed style. The network architecture of base model is derived from FSODM [19], which provides a solution to the few-shot object detection task in the central setting. The network architecture of our base model also comprises three components: a meta-feature extractor, a feature reweighting module and a prediction module, as illustrated in Figure 2.

In order to meet the deployment requirements of resource-constrained clients, we adopt CSP-Darknet 53 [44] as the backbone in the meta-feature extractor module, replacing the Darknet-53 used in FSODM. The meta-feature extractor generates features for the input image at three different scales: $(h/32 \times w/32 \times 1024)$, $(h/16 \times w/16 \times 512)$, and $(h/8 \times w/8 \times 256)$, where h and w represent the height and width of the input image, respectively.

For the feature reweighting module, a CNN-based feature reweighting network is used in FSODM, which focuses on both the background and objects simultaneously. However, by making the feature reweighting network focus more closely on objects, it is possible to enhance the understanding of essential features. In this study, we integrate the squeeze and excitation (SE) attention mechanism [45] into the feature reweighting module.

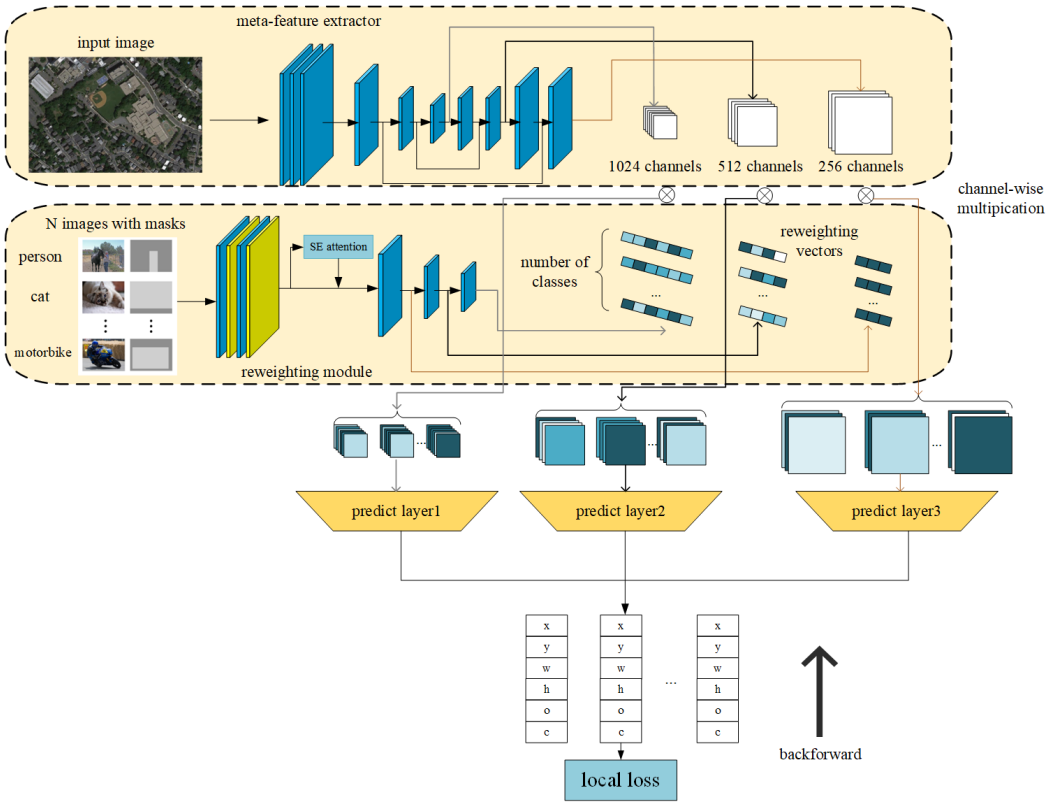


Figure 2. The network architecture of the base model for object detection.

The module processes a concatenated input comprising an RGB image (512 × 512 × 3) and its corresponding binary mask (512 × 512 × 1), forming a 4-channel tensor. Through sequential convolutional and pooling operations, this input is transformed into a 32 × 32 × 256 feature map. Subsequently, the Squeeze-and-Excitation (SE) attention mechanism is applied: In the squeeze phase, global average pooling compresses each channel's spatial dimensions (32 × 32) into a scalar, yielding a 1 × 1 × 256 channel descriptor. The excitation phase then employs two fully connected layers with bottleneck architecture to generate channel-wise attention weights (1 × 1 × 256), which are normalized via sigmoid activation. These weights perform element-wise multiplication with the original feature map, producing precise features that emphasize object-relevant channels while suppressing less informative ones.

For the prediction module, the reweighted feature maps are fed into the prediction module $\mathcal{P}$ to regress detection information, which includes the objectness score o, classification score C, and bounding box location (x, y, w, h).

After obtaining the prediction results, a loss function is employed to measure the difference between the ground truth and the prediction one, providing guidance for next iteration.

Unlike the traditional loss function used in object detection, which includes objectness loss, classification loss, and CIOU localization loss [46], we propose an improved localization loss, called Scale-IOU localization loss. The formulation of the loss function $\mathcal{L}$ is described as follows:

$$\mathcal{L} = l_{obj} + l_c + l_{Scale-IOU}, \tag{4}$$

where l_{obj}, l_c and $l_{Scale-IOU}$, represent objectness loss, classification loss and Scale-IOU localization loss, respectively. Following the weighting schemes adopted in Meta-YOLO [17], TFA [18] and FSCE [36], the weights assigned to these loss components are uniformly set to 1, as Equation (4).

In details, the objectness loss l_{obj} measures the probability that the prediction box contains an object, which is formulated as:

$$l_{obj} = -[\frac{w_o}{N_{pos}} \sum_{pos} \log(P_o) + \frac{w_{no}}{N_{neg}} \sum_{neg} \log(1 - P_o)], \tag{5}$$

where P_o denotes the predicted objectness probability, N_{pos} and N_{neg} represent the numbers of positive and negative samples, w_o and w_{no} denote the weights for objectness and non-objectness loss.

The classification loss l_c measures the probability that the object belongs to a specific class. which is formulated as:

$$l_c = -\frac{1}{N_{pos}} \sum_{pos} \log(\frac{e^{c_{pt}}}{\sum_{j=1}^{N} e^{c_{pj}}}), \tag{6}$$

where c_{pt} and c_{pj} represent the classification scores for the true class and the predicted class j ($1 \leq j \leq N$).

As is well known, objects in an image often vary in size. The localization loss for large objects is consistently higher than that for small objects, even when the IOU of large and small objects is the same. This phenomenon causes the optimization of the localization loss function to focus more on larger objects, thereby degrading the localization accuracy of smaller objects. Therefore, we give the improved localization loss, formulated as Scale-IOU loss function, $l_{Scale-IOU}$.

The Scale-IOU loss function considers the distance of the center point between the predicted bounding box and the ground truth. And the aspect ratio (width-to-height ratio) of these two bounding boxes is also taken into account. Specifically, $l_{Scale-IOU}$ is formulated as

$$l_{Scale-IOU} = 1 - IOU + \frac{d}{\sigma} + \alpha\phi, \tag{7}$$

where d denotes the Euclidean distance between the centers of the predicted bounding box and the ground truth, σ represents the length of the diagonal of the smallest rectangle that encompasses both the predicted bounding box and the ground truth. Additionally, α is the weighting factor, and ϕ is the correction factor, which are defined as

$$\alpha = \frac{\phi}{(1 - IOU) + \phi},$$

$$\phi = \frac{\arctan \delta_t - \arctan \delta_p}{\pi},$$

where δ_t and δp denote the width-to-height ratios of the ground truth and the predicted bounding box, respectively.

Finally, the process of pre-training in the federated learning framework is as follows (see Algorithm 1 for details).

1. **Initialization.** The Server initializes the parameters of base model ω_{base}^0 and broadcast them to clients.
2. **Local training.** Each client i trains the parameters of base model ω_{base}^t based on his local base dateset D_i^{base}.
3. **Upload.** Clients upload their parameters of local base models $(\omega_{base}^t)_i$.
4. **Global aggregation.** The server performs a global aggregation as utilized by [47] through the received local base models, i.e.,

$$\omega_{base}^{t+1} = \sum_{i=0}^{m-1} \frac{|D_i^{base}|}{|D|} (\omega_{base}^t)_i, \tag{8}$$

and then broadcast the aggregated model ω_{base}^{t+1} to clients. Meanwhile, ur framework is also compatible with other aggregation algorithms, such as Krum [48], Trimmed Mean [49], RFA [50] and BRFA [51].

Repeat steps 2–4 until the global model converges.

Algorithm 1 The pre-training stage of F^{2}SOD.

Require: the number of clients m,
the number of pre-train communication rounds T_{base},
the learning rate η
Ensure: Aggregated base model weight ω_{base}^{T}
Server executes:
Initialize ω_{base}^{0}
for $t = 0, 1, ..., T_{base} - 1$ **do**
  for $i = 0, 1, ..., m - 1$ in parallel **do**
    send the global base model ω_{base}^{t} to i
    $(\omega_{base}^{t})_i \leftarrow$**LocalPre-Train**($i$,$w_{base}^{t}$)
  end for
  $\omega_{base}^{t+1} \leftarrow \sum_{i=0}^{m-1} \frac{|D_i|}{|D|}(\omega_{base}^{t})_i$
end for
return w_{base}^{t}
LocalPre-Train(i,w_{base}^{t})**:**
Initialize base detection network based on w_{base}^{t}
for each training episode $\in D_{base}$ **do**
  Local model pre-training and minimize loss l from Equation (4).
end for
return $(\omega_{base}^{t})_i$

3.3. Data Augment

Before fine-tuning the base model, we propose the Diffusion-driven augmentation framework simultaneously harness sample diversity while mitigating overfitting through three components: twofold-tag prompt, location information embedding, and the construction of a fusion loss function based on the Stable Diffusion 1.5 [52,53] for each client, as shown in Figure 3. The specific design is as follows:

Twofold-tag prompt. A typical prompt describes the characteristics of the objects in the image, such as their pose, behavior. To obtain a diverse array of generated samples that incorporate key features, we design twofold-tag prompts. Firstly, the input images from the novel dataset are cropped so that each cropped image contains only one object, which isolates individual characteristics and minimizes interference from surrounding elements. Then, for each object, we construct a twofold-tag prompt that includes both a broad-category tag and a background tag. The broad-category label acts as a general description based on specific object labels. For example, if an object's specific label is "helicopter", its corresponding broad-category label might be be labeled as "airplane" to introduce higher-level semantic abstraction. The background tag provides a concise description of the scene or environmental context, ensuring that the diffusion model captures both the object and its surroundings. This two-tag approach facilitates more controlled and diverse generation, as the broad-category tag allows the model to explore various semantic differences while the background tag ensures coherence with the contextual setting, ultimately resulting in generated samples that are rich in detail and aligned with the intended object features.

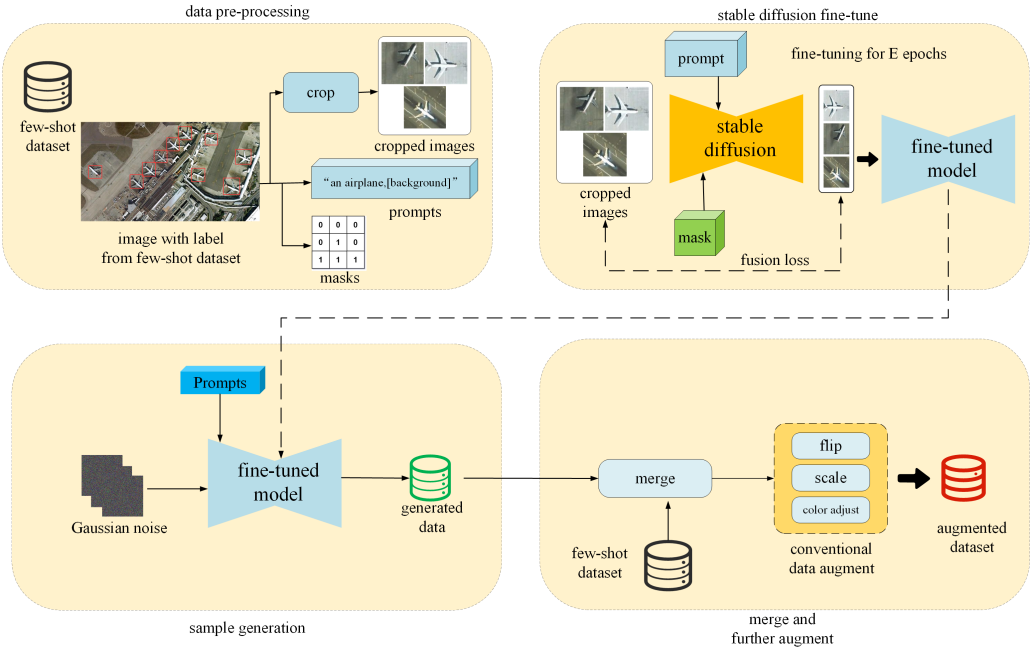


Figure 3. The process of data augment.

Location information embedding. The location information embedding is formulated as a structured positional encoding matrix derived from bounding box coordinates, establishing geometrically-aware priors for spatial feature learning. A complementary binary activation mask $M \in \{0,1\}^{h \times w}$ is from ground truth annotations, maintaining dimension with input image while encoding object presence probabilities. This mask serves as a conditional input for the diffusion model to further guide the generation process. Specifically, it is defined as:

$$M(x,y) = \begin{cases} 1, & \text{if } (x,y) \text{ is inside any bounding box } B_i, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

where $B_i = (x_{min}, y_{min}, x_{max}, y_{max})$ represents the bounding box of an object. Pixels within the target area are set to 1, while the background is set to 0. In cases where multiple bounding boxes overlap, the mask maintains a value of 1 to ensure proper inclusion of all object regions.

Loss function. To encourage the model generating diverse outputs that align with both class and background characteristics, we design a fusion loss function considering prompt and location information embedding. The specific design of the loss function is shown below:

$$\begin{aligned} l_{loss} =& \mathbb{E}_{x,c,\epsilon,\epsilon^t,t}[w_t \| \hat{x}_\theta(\alpha_t x + \sigma_t \epsilon, c) - x \|_2^2] + \\ & \lambda M \odot [w_{t'} \| \hat{x}_\theta(\alpha_{t'} x_{pr} + \sigma_{t'} \epsilon', c_{pr}, c_b) - x_{pr} \|_2^2], \end{aligned} \tag{10}$$

where x refers to the ground-truth image, c refers to conditioning vectors, and α_t, σ_t, w_t refer to terms that control the noise schedule and sample quality, M refers to the matrix of location information embedding, $\odot$ refers to element-by-element multiplication, the second term is the prompt and the location fusion term and λ controls for the relative weight of this term.

Here, we describe the steps of data augment and the details are shown in the Algorithm 2.

1. **Data pre-processing.** The images from novel dataset are cropped so that the cropped image contains one object per image. Subsequently, for each image, a prompt and a location information embedding are constructed according to its label.
2. **Diffusion fine-tuning.** Each input image contains only one object and its prompt with "a [broad-category], [background]". Use cropped images with their prompts and masks to fine-tune stable diffusion guided by the proposed loss function until the specified number of epoch.
3. **Sample generation.** Use fine-tuned stable diffusion to generate new samples.
4. **Merge and further augment.** The real samples and the generated samples are merged and then conventional data augmentation techniques, including flipping, scaling, and color adjustment are applied to make further data augment.

Algorithm 2 The process of data augment.

Require: Diffusion model M_D, few-shot samples $D_{novel} = \{(x_i, y_i)\}_{i=1}^{K}$ from novel classes, text-to-image prompt template T, generation samples quantity n_s, and standard augmentation functions A .
Ensure: Augmented dataset D_{aug}.
Fine-Tune Diffusion Model:
for each novel class c in D_{novel} **do**
Extract samples $D_c \subseteq D_{novel}$ for class c
Generate prompts $\mathcal{P}_c$ and masks$\mathcal{M}$
Adjust M_D on D_c with Fusion Loss
end for
Generate New Samples:
Initialize $D_{gen} = \emptyset$
for each sample $(x_i, y_i) \in D_{novel}$ **do**
Create prompt $p_i = T(y_i)$
for j = 1 to n_s **do**
Generate new image $x_{ij} = M(p_i)$
$Add(x_{ij}, y_i)$ to D_{gen}
end for
end for
Combine and Augment:
Merge datasets: $D_{aug} = D_{novel} \cup D_{gen}$
Apply augmentations: $D_{aug} = A(D_{aug})$
Return: D_{aug}

3.4. Fine-Tuning

To facilitate the base model's ability to detect novel classes, we fine-tune the base class model utilizing augmented data as shown in Figure 4.

Except for the same distributed framework, the fine-tuning process has two differences: (1) the server takes the global base class model obtained from pre-training as the initialized global model. (2) clients use their augmented novel class dataset as the local training sets. The fine-tuning process is shown in Algorithm 3.

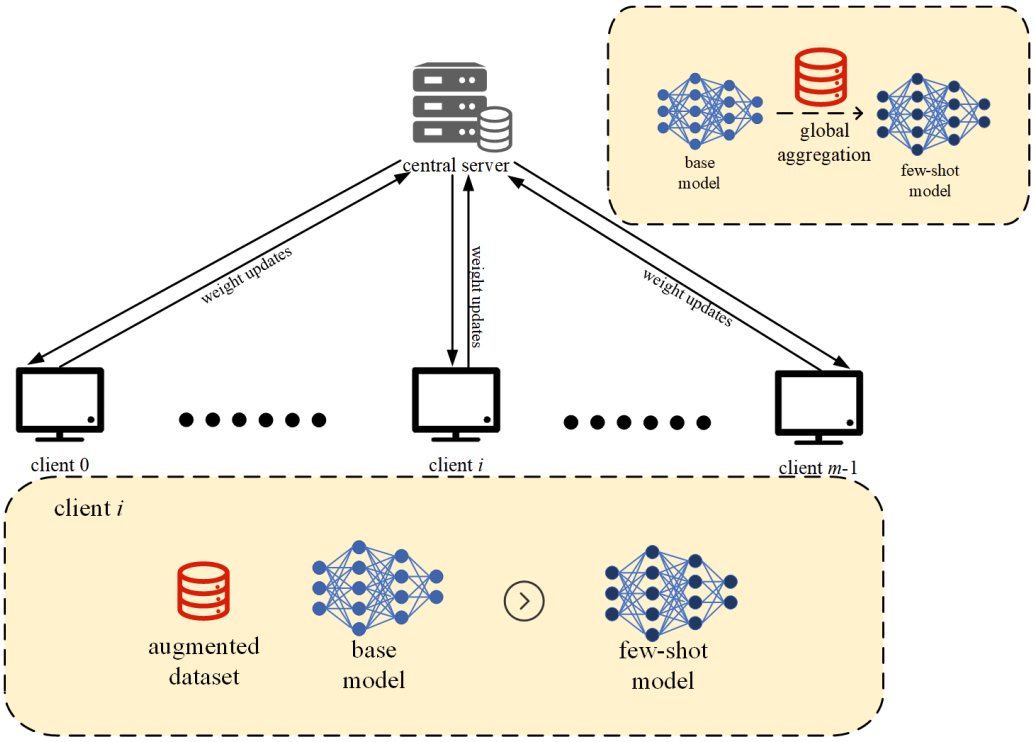


Figure 4. The process of fine−tune object detection model.

Algorithm 3 F^{2}SOD fine-tune process

Require: number of clients m,
number of fine-tuning communication rounds $T_{few-shot}$,
learning rate η
global base model ω_{base}
Ensure: Aggregated few-shot model weight $\omega^{T}_{few-shot}$

Server executes:
Initialize $\omega^{0}_{few-shot} \leftarrow \omega_{base}$
for $t = 0, 1, ..., T_{few-shot} - 1$ **do**
  for $i = 0, 1, ..., m - 1$ in parallel **do**
    send the global base model $\omega^{t}_{few-shot}$ to i
    $(\omega^{t}_{few-shot})_i \leftarrow$**LocalFine-tuning**($i$,$w^{t}_{few-shot}$)
  end for
  $\omega^{t+1}_{few-shot} \leftarrow \sum_{i=0}^{m-1} \frac{|D_i|}{|D|}(\omega^{t}_{few-shot})_i$
end for
return $w^{t}_{few-shot}$
LocalFine-tuning(i,$w^{t}_{few-shot}$)**:**
Initialize few-shot detection network based on $w^{t}_{few-shot}$
for each training episode $\in D_{aug}$ **do**
  Local model fine-tuning and minimize loss l from Equation (4).
end for
return $(\omega^{t}_{few-shot})_i$

4. Experiments

In this section, we show the performance of the method and compare it with the baselines, to evaluate the performance in the distributed few-shot object detection tasks.

4.1. Experimental Setup

Dataset. To evaluate the method, we select a public object detection dataset in remote sensing, called NWPU-VHR10 (https://github.com/Gaoshuaikun/NWPU-VHR-10, accessed on 12 October 2024), which is provided by Northwestern Polytechnical University. Specifically, the dataset offers 800 high-resolution images in 10 object categories, including airplanes, ships, storage tanks, baseball diamonds, tennis courts, basketball courts, ground track fields, harbors, bridges, and vehicles.These images are collected from Google Earth and the ISPRS Vaihingen dataset, covering diverse backgrounds, complex scenes, and various object scales. Among these, there are 150 manually annotated negative samples without objects and 650 positive samples where each one contains at least one object.

In these experiments, each client randomly chooses the seven classes of samples from NWPU-VHR10 and chosen samples are called as its own base set, and for remaining three classes of samples, this client randomly chooses the k (k is a small integer) samples from each class of samples and chosen samples are called as its own few-shot novel set. Obviously, the base set and novel set has no overlapped.

Baseline algorithms. Considering there are no works on distributed few-shot object detection, we mainly select popular and State-of-the-Art methods in the central scenario as baselines in the central scenario and transfer them into in the distributed scenario. These baseline algorithms are summarized as follows:

- Meta-YOLO (meta learning) designs a few-shot detection model that learns generalized meta features and automatically reweights the features for novel class.
- TFA is a two-stage training scheme, which trains a Faster R-CNN with the addition of a instance-level feature normalization.
- FSCE leverages object proposals with different IoU scores for contrastive training, resulting in more robust feature representations.
- FSODM builds on the YOLOv3 framework for remote sensing object detection with minimal annotated samples.
- Digeo uses an offline ETF classifier for well-separated class centers and adaptive margins to tighten feature clusters, improving novel class generalization without compromising base class performance.

Environment and hyperparameters. The experiments in this paper are conducted on a personal computer running Windows 11. The specific hardware configuration includes an Intel i9-13950HX CPU, an NVIDIA GEFORCE RTX 4080 laptop GPU with 12 GB of VRAM. The PyTorch version used is 2.1.2, and the CUDA version is 12.0.1. To simulate a distributed training environment, we set up a simulated federated learning framework on the personal computer.

Inspired by the references [16,17,27] and experimental attempts, the hyperparameters are set as follows: the input image resolution is 512 × 512, batch size is set to 4, and the learning rate for base and novel classes is set to 0.01 and 0.001, respectively. The number of epochs for training the base and novel classes is set to 800 and 10, respectively. The momentum factor is set to 0.95, and the optimizer weight decay is set to 0.0005.

Evaluation metrics. In this paper, we evaluate the performance of our F^{2}SOD using mean Average Precision (**mAP**). For object detection problem, IoU is a parameter to analyze the overlaps of the predicted bounding boxes with those provided by the ground truth. It is computed as

$$IoU = \frac{\text{the area of overlap}}{\text{the area of union}} = \frac{b_{true} \cap b_{pred}}{b_{true} \cup b_{pred}}, \tag{11}$$

where b_{true} and b_{pred} are the ground truth and predicted bounding box, respectively. Average Precision (**AP**) is calculated by averaging the recall at different IoU levels, typically

from 0.5 to 1.0. Mean Average Precision (**mAP**) is the average of AR values across all classes. In this work, we compute mAP at the IoU thresholds of 0.5 (mAP@0.5).

4.2. Our F²SOD Performances

In order to show the performance of our F²SOD, we conduct experiments with two cases: the same number of clients but different shots, the same number of shots but different clients.

Different numbers of shots for each client. In order to explore the impact of shot numbers on the accuracy of object detection, we conduct experiments under 3, 5, 10, 20 shots in 4 clients.

As shown in Figure 5, we find that when the number of clients is fixed and the number of samples per client is increased, mAP of our F²SOD increases obviously. Thus, the increasing number of shots within constraint of few samples is beneficial for model performance.

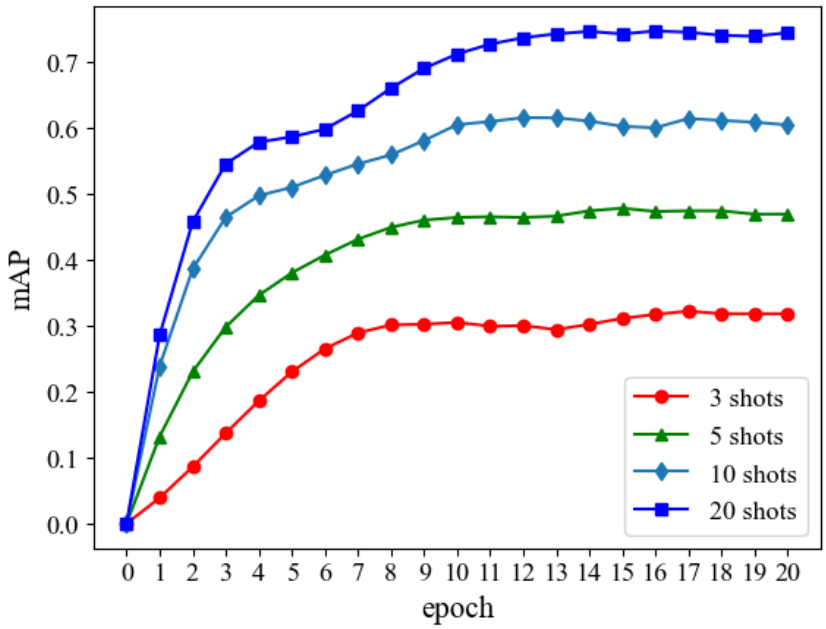


Figure 5. F²SOD performance for different number of shots with 4 clients.

Different numbers of clients. Based on the characteristics of distributed scenarios, we conduct experiments under 1, 2, 4, 8 clients with 5 shots (see Figure 6).

From Figure 6, we find that as the number of clients increases, the performance of F²SOD improves significantly in terms of mAP values. This suggests that increasing the number of clients is an effective way when dealing with few samples in distributed object detection scenarios.

The effectiveness of SE Attention. To verify the validity of SE Attention, we conduct an ablation experiment under 3, 5, 10, 20 shots with 4 clients (see Figure 7).

From Figure 7, we find that across all shot settings, F²SOD consistently outperforms F²SOD without SE Attention in terms of mAP, which indicates that incorporating SE Attention improves the model's performance. Meanwhile, the difference between the two frameworks is more noticeable at lower shot numbers. SE Attention helps compensate for limited data, making it particularly useful for few-shot learning.

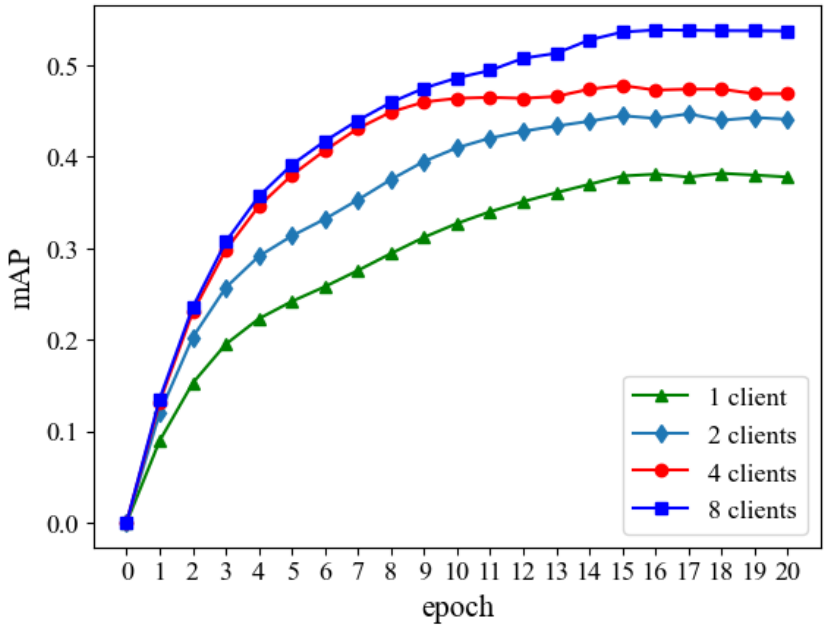


Figure 6. F^{2}SOD performance for different number of clients with 5 shots.

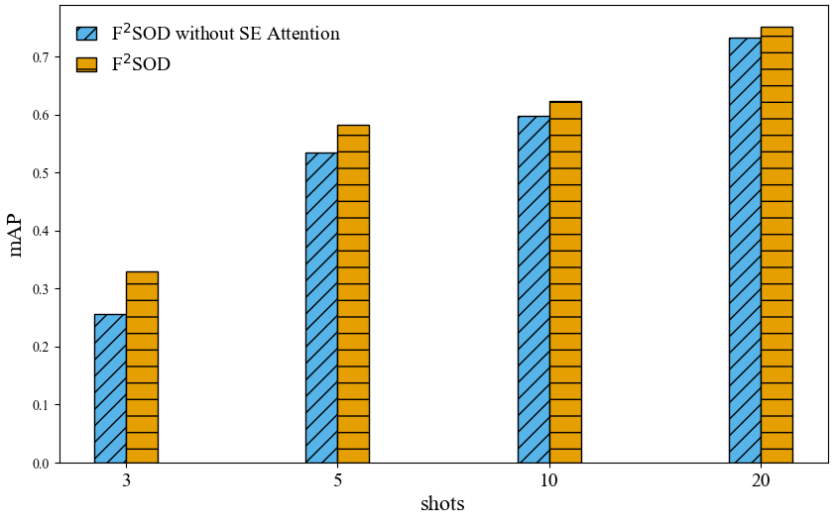


Figure 7. The impact of SE Attention on the performance of F^{2}SOD.

The effectiveness of α and ϕ in loss function. Here, we explore the impact of α and ϕ on the performance of few-shot object detection. In existing studies, α is often set as a constant between 0 and 1, and ϕ is used to describe the Euclidean distance between the predicted bounding box and the ground truth bounding box. In our work, ϕ uses the normalized arctan distance and α is calculated dynamically based on IOU and ϕ .Below, we compare the effects of α and ϕ with different choices (see Figure 8).

In our experiments, α is set to 0, 0.2, 0.4, 0.6, 0.8, 1.0 and ours, which is calculated based on IoU and ϕ. ϕ is set to 0, normalized cityblock distance, normalized Euclidean distance and arctan distance(ours).As shown in Figure 8, α and ϕ we proposed notably enhances detection performance compared to using no scale correction or conventional normalized distance metrics.

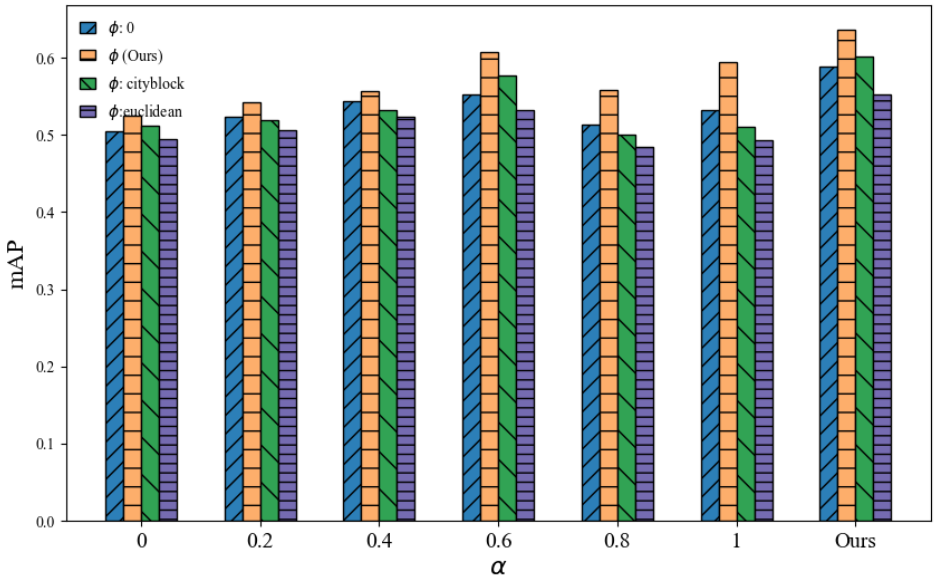


Figure 8. The impact of α and ϕ onthe performance of F^{2}SOD.

Given that the values of these two parameters are associated with the size of the object, we show the performance of our methods under different object sizes (see Figure 9).

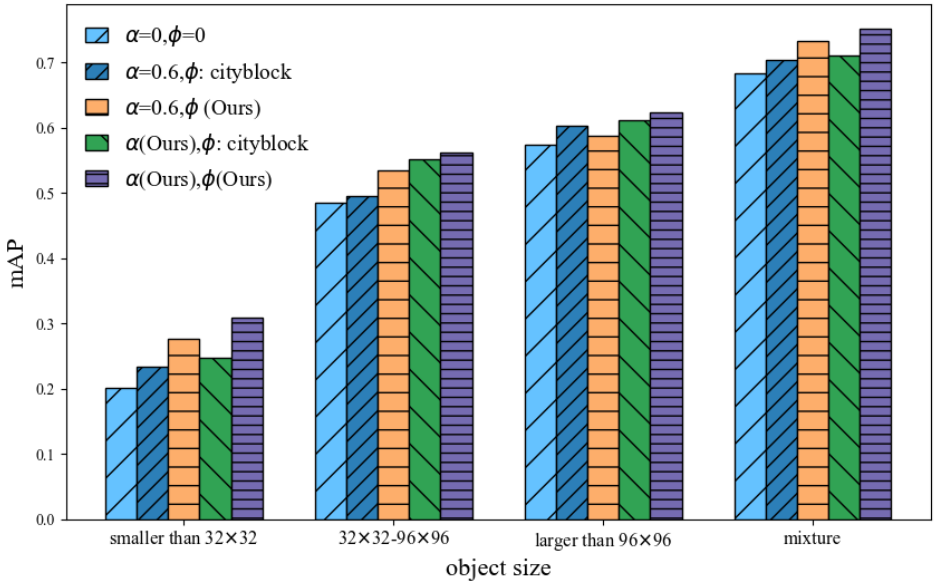


Figure 9. The relevance of α, ϕ to different size of objects.

In details, the small scale is smaller than 32 $\times$ 32, the large scale is larger than 96 $\times$ 96 and others are medium size [54]. From the Figure 9, we find that our proposed α and ϕ can be well adapted to objects of different sizes.

Comparison of different data augment-based approaches. Here, we conduct a comparison between ours with other data augment-based methods: basic image manipulations, GAN [55] and DDPM [56] (see Figure 10).

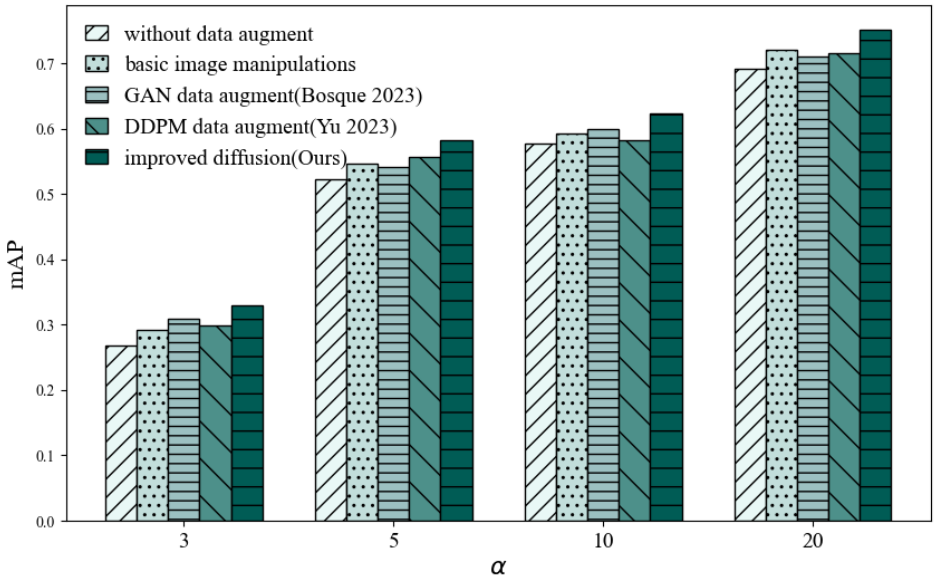


Figure 10. Comparison of different data augment−based approaches [55,56].

From Figure 10, Compared with the method without data augmentation, all different data augment-based methods are helpful to improve the performance of few-shot object detection. Particularly, our method performs the best.

4.3. The Comparison with Baselines

To validate the advantage of our F^{2}SOD, we compare it with the baselines in distributed scenarios with three aspects: accuracy, loss and visualization.

Accuracy. Figure 11 compares the mAP performance of our model against several distributed baselines with four clients under five shots, including FL_Meta-YOLO and FL_FSODM. These results demonstrate the effectiveness of our approach in distributed few-shot learning scenarios. Our model achieved an obvious improvement in accuracy than FL_FSODM, proving the effectiveness of the diffusion model-based data augmentation method in enhancing model performance.

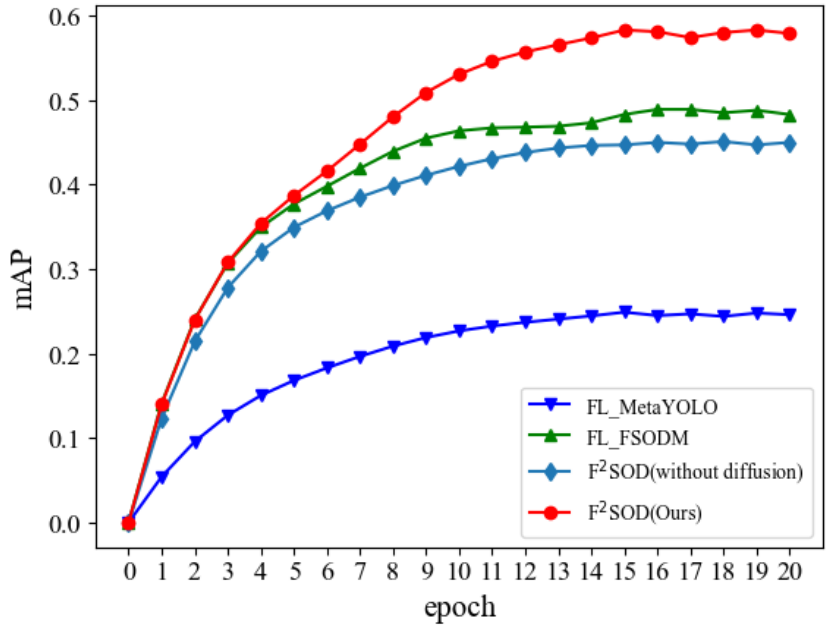


Figure 11. Comparison with baselines in the distributed scenarios.

Convergence speed. From Figure 12, it can be observed that as the number of training epochs increases, the detection accuracy gradually stabilizes and the detection performance improves. Our F^{2}SOD model shows a relatively fast convergence speed and a stable overall convergence rate, which demonstrates that our model requires less training iterations to obtain a effective detection model.

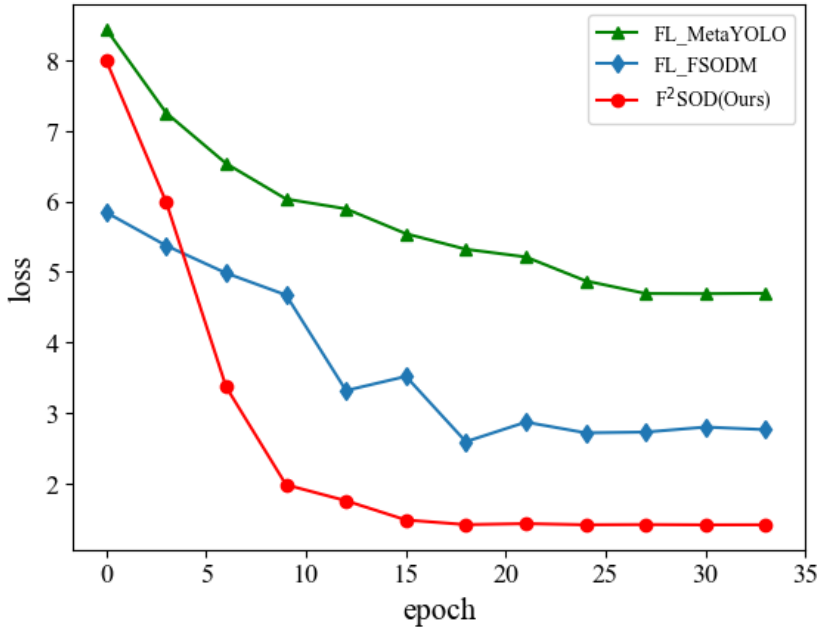


Figure 12. Comparison of the convergence speed.

Visualization. Figure 13 shows some examples of the few-shot detection results of our F^{2}SOD model, baselines on the NWPU VHR-10 dataset and their ground truth. The first column shows ground truth bounding boxes, while the second to fifth columns display detection results from FL_Meta-YOLO, FL_FSODM, and F^{2}SOD (without diffusion) and F^{2}SOD, respectively. As shown in Figure 13, F^{2}SOD can detect all of novel class in the dataset successfully. Compared with the ground truth, F^{2}SOD detects closely while other methods show some detection accuracy but lack precision with missed or incorrect detections especially in small scale objects. This proves the effectiveness of our F^{2}SOD in few-shot object detection area.

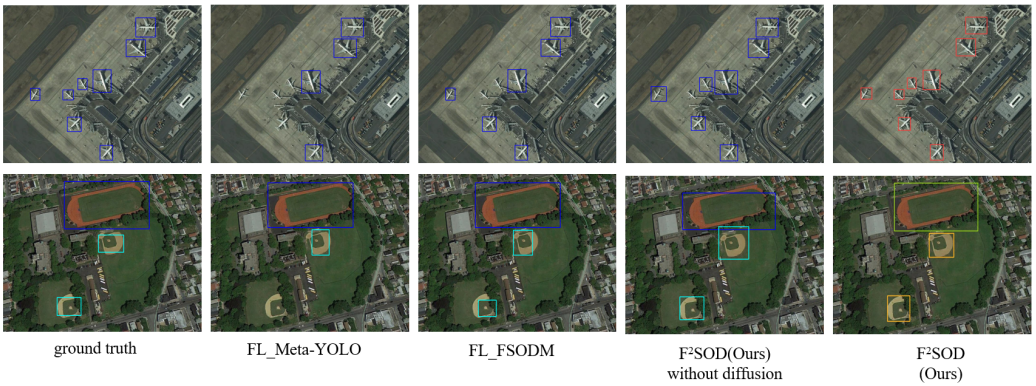


Figure 13. The visual comparison of distributed object detection methods.

5. Conclusions

In this work, we propose F^{2}SOD, a federated few-shot object detection framework tailored for edge computing scenarios constrained by data scarcity and privacy preservation requirements. F^{2}SOD integrates three key components: collaborative base model training, diffusion-based novel data augmentation with twofold tag prompting and object location embedding, collaborative base model fine-tuning for novel model. Experimental validation on public benchmarks confirms that F^{2}SOD significantly outperforms existing methods in both detection accuracy and efficiency.

Although our method performs well in the remote-sense dataset, the cross-domain adaptability is not taken into consideration. Furthermore, with our distributed architecture, the process of clients uploading local models and downloading globally updated models inevitably consumes communication resources. Future work will extend this framework to cross-domain adaptation scenarios and optimize its communication efficiency for large-scale edge networks.

Author Contributions: P.L.: conceptualization, methodology, software, validation, simulation, investigation, writing—original draft preparation. T.Z.: software, validation, simulation, visualization, writing—original draft preparation. C.Q.: writing—review and editing. S.Z.: conceptualization, methodology, supervision, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: There is no external funding for this study.

Data Availability Statement: The dataset used in this study is publicly available and can be accessed at the following link: https://github.com/Gaoshuaikun/NWPU-VHR-10 (accessed on 12 October 2024).

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

References

1. Cazzato, D.; Cimarelli, C.; Sanchez-Lopez, J.L.; Voos, H.; Leo, M. A survey of computer vision methods for 2d object detection from unmanned aerial vehicles. *J. Imaging* **2020**, *6*, 78. [CrossRef] [PubMed]
2. Xin, Z.; Chen, S.; Wu, T.; Shao, Y.; Ding, W.; You, X. Few-shot object detection: Research advances and challenges. *Inf. Fusion* **2024**, *107*, 102307. [CrossRef]
3. Song, Z.; Liu, L.; Jia, F.; Luo, Y.; Jia, C.; Zhang, G.; Yang, L.; Wang, L. Robustness-aware 3d object detection in autonomous driving: A review and outlook. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 15407–15436. [CrossRef]
4. Liang, M.; Su, J.C.; Schulter, S.; Garg, S.; Zhao, S.; Wu, Y.; Chandraker, M. Aide: An automatic data engine for object detection in autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 14695–14706.
5. Van Eden, B.; Rosman, B. An overview of robot vision. In Proceedings of the 2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA), Bloemfontein, South Africa, 28–30 January 2019; pp. 98–104.
6. Himeur, Y.; Rimal, B.; Tiwary, A.; Amira, A. Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives. *Inf. Fusion* **2022**, *86*, 44–75. [CrossRef]
7. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
8. 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.
9. Kim, G.; Jung, H.G.; Lee, S.W. Few-shot object detection via knowledge transfer. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 11–14 October 2020; pp. 3564–3569.
10. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, Nevada, USA, 3–6 December 2012; Volume 25.

11. 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.
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
13. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
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. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
17. 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.
18. Wang, X.; Huang, T.E.; Darrell, T.; Gonzalez, J.E.; Yu, F. Frustratingly simple few-shot object detection. *arXiv* **2020**, arXiv:2003.06957.
19. Li, X.; Deng, J.; Fang, Y. Few-shot object detection on remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
20. Wang, Y.X.; Ramanan, D.; Hebert, M. Meta-learning to detect rare objects. In Proceedings of the the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9925–9934.
21. 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.
22. Han, G.; Huang, S.; Ma, J.; He, Y.; Chang, S.F. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 780–789.
23. Zhang, L.; Zhou, S.; Guan, J.; Zhang, J. Accurate few-shot object detection with support-query mutual guidance and hybrid loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 14424–14432.
24. Li, S.; Song, W.; Li, S.; Hao, A.; Qin, H. Meta-RetinaNet for few-shot object detection. In Proceedings of the British Machine Vision Virtual Conference, BMVC, Virtual, 7–10 September 2020.
25. 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, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5197–5206.
26. Lu, Y.; Chen, X.; Wu, Z.; Yu, J. Decoupled metric network for single-stage few-shot object detection. *IEEE Trans. Cybern.* **2022**, *53*, 514–525. [CrossRef]
27. Li, W.z.; Zhou, J.w.; Li, X.; Cao, Y.; Jin, G. Few-shot object detection on aerial imagery via deep metric learning and knowledge inheritance. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *122*, 103397. [CrossRef]
28. Leng, J.; Chen, T.; Gao, X.; Mo, M.; Yu, Y.; Zhang, Y. Sampling-invariant fully metric learning for few-shot object detection. *Neurocomputing* **2022**, *511*, 54–66. [CrossRef]
29. Chen, Q.; Ke, X. Few-shot object detection based on generalized features. In Proceedings of the 2023 2nd International Conference on Artificial Intelligence and Intelligent Information Processing (AIIIP), Hangzhou, China, 27–29 October 2023; pp. 80–84. [CrossRef]
30. 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.
31. Li, B.; Yang, B.; Liu, C.; Liu, F.; Ji, R.; Ye, Q. Beyond max-margin: Class margin equilibrium 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. 7363–7372.
32. Li, J.; Zhang, Y.; Qiang, W.; Si, L.; Jiao, C.; Hu, X.; Zheng, C.; Sun, F. Disentangle and remerge: Interventional knowledge distillation for few-shot object detection from a conditional causal perspective. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 1323–1333.

33. Ma, J.; Niu, Y.; Xu, J.; Huang, S.; Han, G.; Chang, S.F. Digeo: Discriminative geometry-aware learning for generalized few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3208–3218.
34. Zhu, S.; Zhang, K. Few-shot object detection via data augmentation and distribution calibration. *Mach. Vis. Appl.* **2024**, *35*, 11. [CrossRef]
35. Wang, M.; Wang, Y.; Liu, H. Explicit knowledge transfer of graph-based correlation distillation and diversity data hallucination for few-shot object detection. *Image Vis. Comput.* **2024**, *143*, 104958. [CrossRef]
36. 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, 19–25 June 2021; pp. 7352–7362.
37. 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; Springer: Berlin/Heidelberg, Germany, 2020; pp. 456–472.
38. Wu, C.; Wang, B.; Liu, S.; Liu, X.; Wu, P. TD-sampler: Learning a training difficulty based sampling strategy for few-shot object detection. In Proceedings of the 2022 7th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 22–24 April 2022; pp. 275–279.
39. Yan, B.; Lang, C.; Cheng, G.; Han, J. Understanding negative proposals in generic few-shot object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 5818–5829. [CrossRef]
40. Wang, Y.; Zou, X.; Yan, L.; Zhong, S.; Zhou, J. Snida: Unlocking few-shot object detection with non-linear semantic decoupling augmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 12544–12553.
41. Huang, Y.; Liu, W.; Lin, Y.; Kang, J.; Zhu, F.; Wang, F.Y. FLCSDet: Federated learning-driven cross-spatial vessel detection for maritime surveillance with privacy preservation. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 1177–1192. [CrossRef]
42. Chi, F.; Wang, Y.; Nasiopoulos, P.; Leung, V.C. Parameter-efficient federated cooperative learning for 3D object detection in autonomous driving. *IEEE Internet Things J.* 2025, *early access*. [CrossRef]
43. Behera, S.; Adhikari, M.; Menon, V.G.; Khan, M.A. Large model-assisted federated learning for object detection of autonomous vehicles in edge. *IEEE Trans. Veh. Technol.* **2025**, *74*, 1839–1848. [CrossRef]
44. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
45. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
46. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans. Cybern.* **2022**, *52*, 8574–8586. [CrossRef]
47. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the Artificial Intelligence and Statistics, PMLR, Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
48. Blanchard, P.; El Mhamdi, E.M.; Guerraoui, R.; Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017; NIPS'17, pp. 118–128.
49. Yin, D.; Chen, Y.; Kannan, R.; Bartlett, P. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Dy, J., Krause, A., Eds.; PMLR, Cambridge, MA, USA 2018; Volume 80, pp. 5650–5659.
50. Pillutla, K.; Kakade, S.M.; Harchaoui, Z. Robust Aggregation for Federated Learning. *IEEE Trans. Signal Process.* **2022**, *70*, 1142–1154. [CrossRef]
51. Li, S.; Ngai, E.; Voigt, T. Byzantine-Robust Aggregation in Federated Learning Empowered Industrial IoT. *IEEE Trans. Ind. Inform.* **2023**, *19*, 1165–1175. [CrossRef]
52. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
53. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 10684–10695.
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 Computer Vision–ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Cham, Switzerland, 2014; pp. 740–755.

55. Bosquet, B.; Cores, D.; Seidenari, L.; Brea, V.M.; Mucientes, M.; Bimbo, A.D. A full data augmentation pipeline for small object detection based on generative adversarial networks. *Pattern Recognit.* **2023**, *133*, 108998. [CrossRef]
56. Yu, X.; Li, G.; Lou, W.; Liu, S.; Wan, X.; Chen, Y.; Li, H. Diffusion-based data augmentation for nuclei image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Vancouver, BC, Canada, 8–12 October 2023; pp. 592–602.

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.