remote sensing

MDPI

Article

Domain Feature Decomposition for Efficient Object Detection in Aerial Images

Ren Jin [1], Zikai Jia [1], Xingyu Yin [1], Yi Niu [2] and Yuhua Qi [3,*]

[1] Beijing Key Laboratory of UAV Autonomous Control, Beijing Institute of Technology, Beijing 100081, China; renjin@bit.edu.cn (R.J.); 3120215032@bit.edu.cn (Z.J.); 3220230098@bit.edu.cn (X.Y.)
[2] School of Artificial Intelligence, Xidian University, Xi'an 710071, China; niuyi@mail.xidian.edu.cn
[3] School of Systems Science and Engineering, Sun Yat-Sen University, Guangzhou 510006, China
* Correspondence: qiyh8@mail.sysu.edu.cn

Abstract: Object detection in UAV aerial images faces domain-adaptive challenges, such as changes in shooting height, viewing angle, and weather. These changes constitute a large number of fine-grained domains that place greater demands on the network's generalizability. To tackle these challenges, we initially decompose image features into domain-invariant and domain-specific features using practical imaging condition parameters. The composite feature can improve domain generalization and single-domain accuracy compared to the conventional fine-grained domain-detection method. Then, to solve the problem of the overfitting of high-frequency imaging condition parameters, we mixed images from different imaging conditions in a balanced sampling manner as input for the training of the detection network. The data-augmentation method improves the robustness of training and reduces the overfitting of high-frequency imaging parameters. The proposed algorithm is compared with state-of-the-art fine-grained domain detectors on the UAVDT and VisDrone datasets. The results show that it achieves an average detection precision improvement of 5.7 and 2.4, respectively. The airborne experiments validate that the algorithm achieves a 20 Hz processing performance for 720P images on an onboard computer with Nvidia Jetson Xavier NX.

Keywords: aerial image; object detection; imaging condition; feature decomposition

Citation: Jin, R.; Jia, Z.; Yin, X.; Niu, Y.; Qi, Y. Domain Feature Decomposition for Efficient Object Detection in Aerial Images. *Remote Sens.* **2024**, *16*, 1626. https://doi.org/10.3390/rs16091626

Academic Editor: Sander Oude Elberink

Received: 1 March 2024
Revised: 23 April 2024
Accepted: 30 April 2024
Published: 2 May 2024

1. Introduction

Aerial image target-detection technology can quickly and efficiently extract ground feature information and extend the scene understanding capability of UAVs. It spawns diverse application scenarios [1–4] and is one of the current popular and cutting-edge research directions. The large-scale mobility of the UAV's airborne imaging equipment brings more changes, and Figure 1 shows the imaging diagram of the optical camera mounted on the UAV platform at different flight heights and different perspectives of the gimbal. These factors and some external conditions, such as weather and illumination, are regarded as imaging condition parameters, which means the aerial image data are divided into many fine-grained domains (see Figure 2). Suppose the detector is directly trained on imbalanced fine-grained domain data; the model will overfit the domain with high frequency, underfit the domain with low frequency, and lack the generalization performance for different fine-grained domains.

In aerial imagery, the imaging condition can be clearly obtained from the sensors of the UAV. The shooting height can be obtained from the onboard GPS or barometer. The viewing angle can be obtained from the pitch angle in the gimbal. Also, the shooting time can be obtained from the onboard computer's clock module. Therefore, the efficient use of these free data is the focus of this paper. So far, there has been a lot of research work on domain adaptation [5–10]. However, they all explicitly assume one or more source domains and a target domain with no label information. Then, they use a transfer learning algorithm

to adapt the model trained in the source domain to the target domain. The feasibility of generalizing these methods to handle many fine-grained domains is questionable [11].

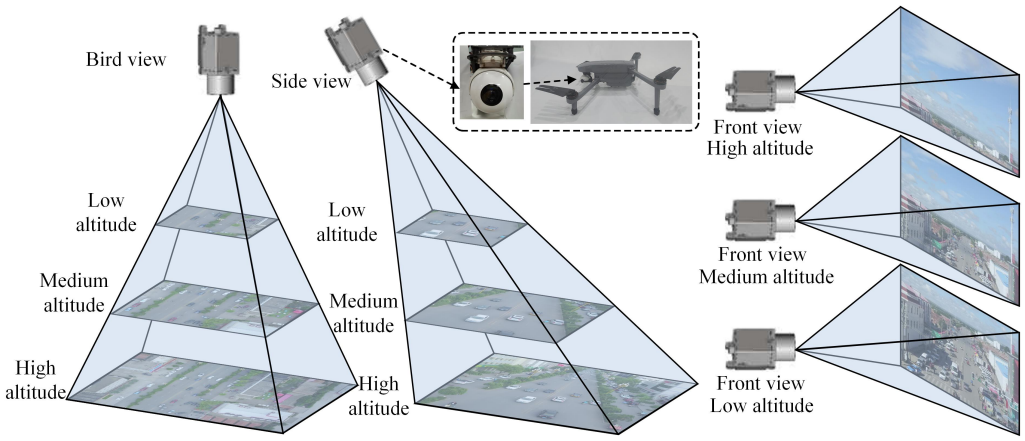

Figure 1. The imaging diagram of the imaging equipment mounted on the UAV platform at different flight heights and different pan/tilt angles; these altitude and angle data can be read out by sensors on the UAV.

Figure 2. Fine-grained domains in aerial images under different imaging conditions.

There are also some studies on fine-grained domain detectors for aerial images; Wu et al. [11] proposed an adversarial training framework dubbed Nuisance-Disentangled Feature Transform (NDFT) to learn fine-grained domain-invariant features, thereby obtaining more robust detection effects in multiple fine-grained domains. Lee et al. [12] improved the problem of the slow convergence of NDFT and used feature replay and slow learner techniques to speed up the learning of domain-invariant features. In addition to the idea of learning domain-invariant features, Kiefer et al. [13] believed that domain-specific features have a positive effect on fine-grained domain detection and the unbalanced domain distribution leads to a decrease in detection performance. Therefore, they proposed to add independent detection sub-networks for each domain to extract domain-specific features and reduce the problem of imbalance in domain distribution.

Furthermore, recent advancements in object detection have introduced novel techniques such as DETRs with hybrid matching [14], Adaptive Rotated Convolution for Rotated Object Detection [15], Rank-DETR for High-Quality Object Detection [16], and V-DETR for 3D object detection [17]. These approaches have pushed the boundaries of object detection, offering promising avenues for improving detection performance in various domains. Moreover, AdaDet, an Adaptive Object-Detection System based on Early-Exit Neural Networks [18], has demonstrated its efficacy in handling diverse object-detection

tasks. However, the application of these techniques to fine-grained domain detection in aerial imagery remains unexplored.

The above observations can be summarized as follows. Firstly, both domain-invariant and domain-specific features can improve fine-grained domain-detection performance. Second, adding a separate detection sub-network for each fine-grained domain will add more parameters. Moreover, the network structure needs to be changed when new fine-grained domains are introduced. Lastly, the above algorithms are all modified for two-stage detectors, which are not conducive to micro and small UAV deployment. Therefore, we propose a fine-grained feature-disentanglement network to learn domain-specific and domain-invariant features using airborne metadata simultaneously. Furthermore, we propose a fine-grained domain mix augmentation algorithm to better combine the fine-grained domain learning module with the YOLOv5 [19] framework and achieve better detection performance than NDFT (two-stage detector) on a single-stage detector.

In summary, our contributions are as follows:

1. We propose a fine-grained feature-disentanglement network. It uses airborne metadata as supervision information to disentangle domain-invariant and domain-specific features. These decomposed features improve cross-domain generalization and single-domain detection accuracy.
2. We propose a domain mix augmentation algorithm. It alleviates the problem of fine-grained domain-distribution imbalance and solves the problem that Mosaic augmentation cannot be used for airborne metadata label learning.
3. The proposed method achieves state-of-the-art performance on both VisDrone and UAVDT datasets. Meanwhile, it is able to process 720P images at 20 Hz on an Nvidia Jetson Xavier NX airborne computer (Nvidia Corporation, Santa Clara, CA, USA).

2. Related Work

2.1. Domain Shift of Remote Sensing

Domain shift is a common problem in aerial remote-sensing images. The model of aerial images is easily affected by various imaging conditions, such as viewpoint geometry, atmospheric effect, sensor properties, and temporal variability. Studies such as [11,13] showed that there are obvious differences in detection performance under different perspectives on the VisDrone and UAVDT datasets. The same is true for similar satellite remote-sensing images. Weir et al. [20] found that the existing satellite image datasets are taken from the vertical top view. When the detector trained on these datasets encounters the input image under a certain offset view, its detection performance is significantly reduced. Tasar et al. [10] also observed that even when the same satellite is used to sense different regions, the model performance deteriorates due to changes in color distribution.

In the context of this problem becoming more prominent, many domain-adaptation algorithms involving remote sensing orientations have been proposed [5–10]. Nevertheless, these studies all assume one or more source (ideal) domains and one target (non-ideal) domain, which requires explicit retraining whenever a new target domain emerges. On the contrary, this paper focuses on building a robust detector for fine-grained domains.

2.2. Fine-Grained Domain Object Detection

Most off-the-shelf detectors are typically trained with less variable, field-limited data. In contrast, a large number of external imaging conditions specific to UAVs (such as altitude changes, viewpoint changes, and weather changes) cause UAV-based detection models to run in a large number of different fine-grained domains. To the best of our knowledge, NDFT [11] is a pioneering work to demonstrate the effectiveness of fine-grained domain-invariant learning in UAV images. They added NDFT to Faster R-CNN, obtained a 2% mAP increase on the UAVDT dataset, and only used metadata recorded by drones without additional annotation work. Since then, Lee et al. [12] proposed feature replay and slow learner techniques to improve the problem of slow NDFT training further. Kiefer et al. [13] believed that domain imbalance is an essential factor leading to the

performance degradation of fine-grained domain detection. They proposed using different heads to learn domain-specific features so that the detector will not be affected by domain bias caused by fine-grained domain imbalance.

In comparison, our proposed algorithm utilizes metadata to disentangle invariant features and specific features in different fine-grained domains. It makes full use of them to obtain a more robust detection, taking into account the different ideas from the above literature. In addition, this paper proposes a fine-grained domain-mix-augmentation algorithm, which achieves higher accuracy and speed on the one-stage detection framework. The proposed method will be verified in the subsequent experimental part. As a key indicator to evaluate the performance of the object detection model, mAP represents the area under the average precision and recall curve of all categories, which we will use for evaluation. It has significant advantages when deployed on UAVs.

2.3. Disentangled Representation Learning

As an efficient feature-decomposition mechanism, Disentangled Representation Learning (DRL) is effective in many tasks, such as image style transfer [21] and few-shot learning [22]. Lee et al. [21] used DRL to decompose image features into one domain-invariant content space and another domain-specific attribute space. These are used to capture cross-domain shared information and improve the diversity of image style transfer, respectively. Peng et al. [23] proposed to decompose category-invariant features, domain-invariant features, and domain-specific features to achieve domain adaptive classification tasks. However, this work only considers the classification representation at the overall image level.

Wu et al. [24] used DRL for domain-invariant feature learning in object detection to achieve domain-adaptive object detection, but this paper is different from them in two respects. The first is that their framework has a clear definition of the source domain and target domain, which cannot be directly applied to fine-grained domain object detection. Second, they pay more attention to domain-invariant feature learning while ignoring the contribution of domain-specific features to the detector.

3. Approach

3.1. Problem Definition

We define an aerial image object detection training dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}$, where $\mathbf{x}_i$ is an input image and the label $\mathbf{y}_i$ includes objects' category and bounding box coordinates. The images in the dataset $\mathcal{X}$ can be divided into many fine-grained domains $\{\mathcal{X}_{d_1}, \mathcal{X}_{d_2}, ..., \mathcal{X}_{d_n}\} \in \mathcal{X}$ according to the shooting conditions. There is a covariate shift in the distributions $p_{d_a} : \mathcal{X}_{d_a} \times \mathcal{Y} \to \mathbb{R}^+$ and $p_{d_b} : \mathcal{X}_{d_b} \times \mathcal{Y} \to \mathbb{R}^+$ of any two fine-grained domains. In other words, suppose $p_{d_a}(\mathbf{y}|\mathbf{x}) = p_{d_b}(\mathbf{y}|\mathbf{x})$, but $p_{d_a}(\mathbf{x}) = p_{d_b}(\mathbf{x})$ [25].

3.2. Framework

Considering the computational performance constraints of onboard computing devices, we use a single-stage detector as the main network framework and propose two sub-modules named Fine-grained Domain Mix Augmentation and Fine-grained Feature Disentanglement. Using these sub-modules along with the free information recorded by the airborne sensors, we achieve better detection performance than the two-stage fine-grained detector NDFT [11] on two typical aerial image datasets, UAVDT and VisDrone. The specific network structure diagram is shown in Figure 3.

The Fine-grained Domain Mix Augmentation (FDM) module mixes the fine-grained domain images and outputs the fine-grained domain mask labels corresponding to the mixed images. The Fine-grained Feature Disentanglement (FGFD) module decomposes the features of the object on the image into fine-grained domain-invariant features and fine-grained domain-specific features and simultaneously uses them to improve detection performance. For the detailed description of the above two modules, please refer to the following Sections 3.3 and 3.4, respectively.

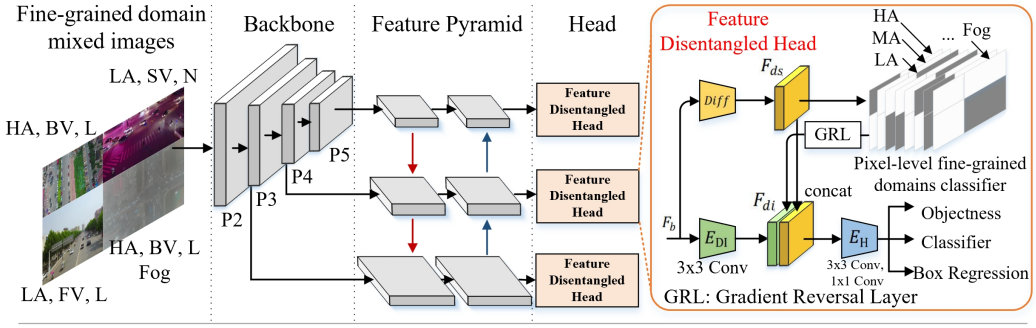

Figure 3. Illustration of the proposed main framework, including fine-grained domain mix augmentation and fine-grained domain feature disentanglement modules, and combining them as components in the YOLOv5 series. 'Diff' and 'GRL' separately indicate the difference disentanglement and Gradient Reversal Layer [26,27].

3.3. Fine-Grained Domain Mix Augmentation

There are many advanced single-stage detectors [28–30] using Mosaic augmentation, which effectively improves detection accuracy and robustness. However, the simultaneous application of Mosaic augmentation and imaging condition disentangled learning is complex. Because the airborne imaging condition information is for a single input image, such as an image taken at low altitude, front view, and night, Mosaic augmentation needs to mix four input images, and the position of each image is also random. This makes it difficult for single-stage detectors to directly use the training framework proposed by Wu et al. [11] to handle combinations of imaging condition. Besides that, abandoning the use of Mosaic augmentation also leads to a large drop in accuracy. Therefore, we propose a new enhancement method to deal with the above problems, called fine-grained domain mix (FDM). The main algorithm flow is shown in Algorithm 1.

Algorithm 1: FDM augmentation

Input: A batch β of B images, bounding box labels $\{\mathbf{y}_i^b\}$, and its corresponding imaging condition labels $\{\mathbf{y}_i^n\}$ from training data $\mathcal{D}$, where $y_i^n \in \{0,1\}^m$ and m is the number of conditions

Output: A batch of fine-grained domain samples $\widehat{\beta}$

$\widehat{\beta} \leftarrow \emptyset$;

for $i \leftarrow 1, B$ **do**

  $\mathcal{S} \leftarrow \{(\mathbf{x}_i, \mathbf{y}_i)\}$;

  for $j \leftarrow sample(\mathcal{D}, 3)$ **do**

    $\mathcal{S} \leftarrow \mathcal{S} \cup \{(\mathbf{x}_j, \mathbf{y}_j)\}$;

  end

  Collate crops from 4 images in $\mathcal{S}$ into $\widehat{\mathbf{x}}_i$;

  Recompute all box coordinates in $\mathcal{S}$ into $\widehat{\mathbf{y}}_i^b$;

  Create mask label $\widehat{\mathbf{y}}_i^n \in \mathbb{R}^{h \times w \times m}$, where h, w are the height and width of $\widehat{\mathbf{x}}_i$, and assign the imaging condition labels $\{0,1\}$ of the corresponding crops to the mask;

  $\widehat{\beta} \leftarrow \widehat{\beta} \cup \left\{\left(\widehat{\mathbf{x}}_i, \widehat{\mathbf{y}}_i^b, \widehat{\mathbf{y}}_i^n\right)\right\}$;

end

For each image in the batch, three images are randomly selected from the training data $\mathcal{D}$ to form a 2×2 mixed fine-grained domain collage image. The difference from Mosaic augmentation is that FDM needs to build imaging condition labels during each image mixing process and convert the imaging condition labels of each image into mask labels of

the collage images, where h, w are the height and width of the collage image, and m is the number of imaging condition labels. For example, the first condition (high altitude) of the image in the upper left area of the collage image is 1. Then its corresponding mask area is assigned the value of 1. Additionally, we build a weight-balanced sampler on sampling uniformly from each fine-grained domain.

3.4. Fine-Grained Feature Disentanglement

The right half of Figure 3 shows the Feature Disentangled Head (FDH) proposed in this paper, which is combined with the structure of a single-stage detector YOLOv5 series. E_{DI} is a convolutional layer with a kernel of 3×3, and E_{H} is the head network of YOLOv5; specifically, a 3×3 convolutional layer is followed by a 1×1 convolutional layer. The number of input channels is consistent with the output of the YOLOv5 feature fusion network, which are 128, 256, and 512, respectively. Given a sample $\widehat{\mathbf{x}}_i$ after fine-grained domain mix augmentation, the detector first extracts high-level semantic features through a feature-extraction network and then obtains multi-scale object representations through a feature pyramid network. Take one of the scale feature maps as an example, and let it be F_b. FDH uses a convolution operation E_{DI} to disentangle the domain-invariant features F_{di} in F_b. Here, F_b contains both domain-invariant and domain-specific features; that is, $F_b = F_{di} + F_{ds}$, and domain-specific features F_{ds} are the difference between F_b and F_{di}:

$$F_{di} = E_{\mathrm{DI}}(F_b), F_{ds} = F_b - F_{di} \tag{1}$$

where E_{DI} represents the domain-invariant feature extractor and F_{di} and F_{ds} are the domain-invariant and domain-specific features obtained by FDH disentanglement, respectively.

Next, we discuss how to use imaging condition labels $\widehat{\mathbf{y}}_i^n$ to train F_{di} and F_{ds} and improve the accuracy and robustness of fine-grained object detector. In order to obtain more domain-specific features, we design a pixel-by-pixel classification network C_{ds} to classify the imaging condition labels of mixed images. C_{ds} consists of a 3×3 kernel convolutional layer and a pixel-wise cross-entropy loss layer. The output channels of the convolutional layer are consistent with the number of imaging condition labels. A value of 0 means that the imaging condition label of the image to which the current pixel belongs is false, and 1 means the opposite. Then, the loss function of training F_{ds} is as follows:

$$\mathcal{L}_{ds} = -[\widehat{\mathbf{y}}_i^n \log \overline{\mathbf{y}}_i^n + (1 - \widehat{\mathbf{y}}_i^n) \log(1 - \overline{\mathbf{y}}_i^n)] \tag{2}$$

where $\overline{\mathbf{y}}_i^n = C_{ds}(F_{ds})$, $\mathcal{L}_{ds}$ is used to generate fine-grained domain-specific features. Meanwhile, $\mathcal{L}_{di} = -[\widehat{\mathbf{y}}_i^n \log \widetilde{\mathbf{y}}_i^n + (1 - \widehat{\mathbf{y}}_i^n) \log(1 - \widetilde{\mathbf{y}}_i^n)]$, and a GRL (Gradient Reversal Layer) [26,27] module generates fine-grained domain-invariant features, where $\widetilde{\mathbf{y}}_i^n = \mathrm{GRL}(C_{ds}(F_{di}))$.

Furthermore, the key to feature disentanglement is maintaining the independence of each component. Here, based on vector-decomposition theory, the orthogonality of the disentanglement components can effectively improve their independence. On this basis, we add an additional orthogonal loss $\mathcal{L}_{\perp}$ to F_{di} and F_{ds}. Considering that the detector is more concerned with the independence of object features, the orthogonal loss can be expressed as

$$\mathcal{L}_{\perp} = \frac{1}{b} \sum_{i=1}^{b} \left| \sum_{j=1}^{c} [\left\| \mathbb{1}^{\mathrm{obj}}(F_{di}) \right\|_2^2 \odot \left\| \mathbb{1}^{\mathrm{obj}}(F_{ds}) \right\|_2^2]_{i,j} \right| \tag{3}$$

where $\mathbb{1}^{\mathrm{obj}}(F_{di}) \in \mathbb{R}^{b \times c}$ indicates that the corresponding anchor in the F_{di} feature map can cover the spatial position of an object and b and c represent the number of selected spatial locations and the number of feature channels of F_{di}, respectively. $\odot$ represents the element-wise product, and $|\cdot|$ and $\|\cdot\|_2^2$ represent the absolute value operation and L2 normalization, respectively.

When the orthogonal loss is minimized, the independence of F_{di} and F_{ds} is increased, minimizing $\mathcal{L}_{di}$ and $\mathcal{L}_{ds}$ making F_{di} and F_{ds} retain more fine-grained domain-invariant and

specific features. Finally, the detection head is run on the concatenated feature maps of F_{di} and F_{ds} to better utilize both domain-invariant and domain-specific features, and ablation studies can also prove this point well. The overall training loss function is as follows:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{loc}} + \mu\mathcal{L}_{di} + \mu\mathcal{L}_{ds} + \nu\mathcal{L}_{\perp} \quad (4)$$

where $\mathcal{L}_{\text{obj}}$, $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{loc}}$ represent the objectness loss, classification loss, and localization loss, respectively; these three losses are consistent with the standard loss in YOLOv5.

4. Experiments

4.1. Datasets

We evaluate our proposed algorithm on two datasets, UAVDT and VisDrone.

UAVDT is the first dataset to explore object detection and tracking in unconstrained scene aerial images. The dataset collected 10 h of airborne raw video and extracted about 80,000 representative image frames. This dataset completely annotate the objects' bounding boxes and 14 additional imaging condition labels (such as flight altitude, camera angle of view, weather conditions) in these images, covering the three UAV-specific fine-grained domains used in the paper. UAVDT contains about 41,000 frames of images and 840,000 bounding boxes in the object-detection task. These objects include three categories, namely, cars, trucks, and buses. However, the category distribution is hugely unbalanced, and the number of trucks and buses accounts for less than 5% of the total. Therefore, referring to the original author's agreement, these three categories are combined into the vehicle category, and our quantitative evaluation is carried out on this basis.

VisDrone is a representative dataset based on a micro-UAV platform, including 263 video clips, 179,264 video frames, and 10,209 still images. Various drone platforms collected these images in 14 different urban scenes in China. In terms of the object-detection task, VisDrone contains 10,209 static images in unconstrained scenes, of which 6471 are used as the training set, 548 are used as the validation set, and 1580 are used as the test-challenge set. Reference [11] has already annotated the training set with labels of imaging conditions. Therefore, we used the same settings to conduct experiments and quantitative evaluations on the VisDrone validation set.

4.2. Implementation Details

We used YOLOv5m as the baseline model for the experimental part. The feature-disentanglement module, Equation (1), is added after the model's 23rd layer. The inputs of this module are the feature maps generated by the FPN network at three scales with 128, 256, and 512 channels, respectively; the outputs of the module are the domain-invariant and domain-specific features corresponding to the input feature maps with the same number of channels as the inputs. The output fine-grained domain mask label of FDM is the same size as the input mixed image, which is 1280 $\times$ 1280 pixels. The domain-invariant features F_{di} generate a feature map with the same dimension as the number of imaging conditions through a GRL layer and a convolutional layer. The output feature map and mask labels are calculated through pixel-wise cross-entropy loss for domain-invariant feature learning. The learning of domain-specific features F_{ds} is similar except that the GRL layer is removed.

The other training parameters mainly include the initial learning rate, which is 0.01; the final OneCycle [31] learning rate, which is 0.002; the momentum, which is 0.94; and the optimizer weight decay, which is 5×10^{-4}. There are 4 and 32 training epochs on the UAVDT and VisDrone datasets, respectively, because of the large number of similar images on UAVDT. In terms of training loss, the weight of objectness is 0.35, and the weight of classification and box regression are 0.5 and 0.05. The feature disentanglement learning loss $\mu_1 = \mu_2 = 0.05$, $\mu_3 = 0.2$, and the orthogonal loss $\nu = 0.1$. In terms of data augmentation, Mosaic or FDM is enabled by default. The range of image scale augmentation is [0.72, 1.28], the range of image translation augmentation is [−45, 45] pixels, and the probability of image horizontal flipping is 0.5. Our training and testing experiments were performed on

an Nvidia GTX 1080Ti graphics card, including the experiments testing the inference time. Some more specific implementation details can be found in the open-source code.

4.3. Evaluation Metrics

The evaluation criteria in the experiments follow the evaluation protocol in the COCO [32] dataset. The main evaluation metrics we use are AP_{50} and AP_{70}, which represent the average precision overall classes with IoU thresholds of 0.5 and 0.7, respectively. This is based on the evaluation metrics used on UAVDT and VisDrone datasets in references [11,12]. Specifically, the AP70 evaluation metric is used on the UAVDT datasets, and the AP50 evaluation metric is used on VisDrone. We also used the evaluation metric AP, which is the average of the evaluation results, with the IoU threshold ranging from 0.5 to 0.95.

4.4. Ablation Study

We carried out ablation studies of the proposed algorithm on the UAVDT dataset. In UAVDT, all image frames are manually annotated with UAV-specific imaging conditions, including flying altitude (low, medium, and high), camera views (front-view, side-view, and bird-view), and weather conditions (daylight, night). Consistent with [11], a small number of foggy conditions were ignored in the experiment. These three conditions are referred to as A, V, and W for short.

4.4.1. Influence of Imaging Conditions on Fine-Grained Feature Disentanglement

We first tested the effect of a single A, V, W fine-grained domain disentanglement on the AP_{70} by adjusting the coefficient μ in Equation (4). μ_1, μ_2, and μ_3 represent the learning coefficients of A, V, and W, respectively. Then, we gradually extended the test to verify the gain in detection performance with two and three imaging conditions. Tables 1–3 show the benefit of fine-grained domain disentanglement of independently adding fly altitude (A), camera view (V), and weather (W) conditions, respectively. For FGFD training, the corresponding condition coefficient μ is nonzero. The baseline model without FGFD has a $\mu_i = 0 (i = 1, 2, 3)$. As can be seen from Table 1, when $\mu_1 = 0.05$, an overall AP_{70} improvement of 3.0 can be obtained. Similarly, in Table 2, when $\mu_2 = 0.05$, a maximum overall AP_{70} improvement of 2.8 is obtained. Table 3 shows the effect of weather condition on AP_{70}, and it can be seen that when $\mu_3 = 0.2$, the overall AP_{70} increases by 3.4.

Table 1. Learning FGFD on altitude condition with different μ_1 values.

A / μ_1	Low	Med	High	Overall
0.0	75.5	60.4	23.2	53.5
0.02	75.4	60.2	28.4	55.9
0.05	75.4	59.6	**31.5**	**56.5**
0.1	76.7	61.2	25.6	55.6
0.2	**76.8**	**61.6**	23.8	55.4

Table 2. Learning FGFD on view angle condition with different μ_2 values.

V / μ_2	Front	Side	Bird	Overall
0.0	59.8	69.1	34.6	53.5
0.02	**61.0**	69.3	38.1	55.8
0.05	60.8	69.6	**39.9**	**56.3**
0.1	60.0	**70.2**	36.0	55.6
0.2	60.3	69.3	33.3	54.4

Table 3. Learning FGFD on weather condition with different μ_3 values.

W / μ_3	Day	Night	Overall
0.0	63.8	72.1	53.5
0.02	64.9	72.8	54.8
0.05	64.7	72.6	54.6
0.1	64.4	**73.7**	55.2
0.2	**66.0**	65.4	**56.9**

Next, we further tested the performance of the combination of two or three conditions in FGFD learning. Table 4 shows the full results of this experiment, where A+V means that both flying altitude and camera view conditions are used for FGFD training. A+V+W stands for simultaneously using flying altitude, camera view, and weather conditions. To obtain the maximum AP_{70} performance and refer to the experimental results shown in Table 1–3, for altitude condition, we set $\mu_1 = 0.05$; at the same time, for the view angle and weather condition, we set $\mu_2 = 0.05$ and $\mu_3 = 0.2$, respectively. It can be observed that as more conditions are introduced, the AP_{70} also increases. The final model, A+V+W, achieves better performance in each fine-grained domain and improves the AP_{70} by 4.2 compared to the baseline model. Improving the object detection performance of UAVs at a high altitude, from a bird's-eye view, and in night scenes can also further improve the reliability and robustness of the model in potentially harsh scenarios.

Table 4. Learning FGFD with multiple conditions on the UAVDT dataset.

	Baseline	**A**	**V**	**W**	**A+V+W**
			Flying Altitude		
Low	75.5	77.3	75.2	75.4	**77.6**
Med	59.6	61.6	59.5	59.3	**61.9**
High	23.2	26.3	23.4	23.1	**28.1**
			Camera View		
Front	59.8	59.7	60.8	59.6	**61.0**
Side	69.1	68.9	69.6	69.0	**70.3**
Bird	34.6	34.9	**39.9**	34.5	38.8
			Weather Condition		
Day	62.8	62.5	62.7	66.0	**66.9**
Night	70.1	70.5	70.3	73.7	**73.9**
Overall	53.5	$56.1_{\uparrow 2.6}$	$55.2_{\uparrow 1.7}$	$56.3_{\uparrow 2.8}$	$\mathbf{57.7}_{\uparrow 4.2}$

4.4.2. Effectiveness of the Fine-Grained Domain Mix

Compared with the two-stage fine-grained domain detector NDFT proposed by Wu et al. [11], whose paper reported the highest AP_{70} of 52.03 on the UAVDT dataset, our best model achieves an AP_{70} improvement of 5.7 to an AP_{70} of 57.7. Using a single-stage detector to achieve such an AP_{70} is inseparable from the introduction of the fine-grained domain mix data augmentation. To verify this, we performed the following ablation experiments. YOLOv5m is used as the baseline model in this comparison experiment, where '-' indicates that the Mosaic augmentation is not used, 'Mosaic' means that Mosaic augmentation is applied, and the other parameters remain the same. As can be seen in Table 5, the baseline model YOLOv5m has an AP_{70} improvement of 2.6 after the use of Mosaic augmentation. FGFD has an AP_{70} improvement of 4.6 after the use of FDM, showing the contribution of the proposed FDM in this paper.

4.4.3. Effectiveness of Invariant and Specific Features

Using both domain-invariant and domain-specific features to improve the performance of fine-grained domain object detection is an essential innovation of this paper. As shown in Figure 3, we used a fine-grained domain mask segmentation loss to learn domain-specific

features F_{ds}; meanwhile, a gradient-reversal layer was used to learn fine-grained domain-invariant features F_{di}, where F_{ds} and F_{di} are obtained from a common feature map F_b by feature vector decomposition, i.e., $F_b = F_{ds} + F_{di}$. Finally, the head network outputs the object-detection results based on the concatenation of the F_{ds} and F_{di} feature maps. In order to verify the contribution of F_{ds} and F_{di} to AP_{70}, their corresponding losses $\mathcal{L}_{ds}$ and $\mathcal{L}_{di}$ are, respectively, assigned to 0, other parameters are kept the same, and the effect on the overall AP_{70} is observed; the results are shown in Table 6. We can see that the learning of both domain-invariant and domain-specific features contributes to the overall AP. We also reproduced the YOLOv5m+NDFT [11] algorithm, and its AP is comparable to the result of using only domain-invariant features in this paper.

Table 5. Effects of the proposed fine-grained domain mix augmentation on the UAVDT dataset.

Model	Augmentation	AP	AP_{50}	AP_{70}
YOLOv5m	-	38.8	69.3	50.9
YOLOv5m	Mosaic	41.5	72.9	53.5
FGFD	-	40.4	71.6	53.1
FGFD	FDM	44.1	75.8	57.7

Table 6. Effects of the proposed invariant and specific feature learning on the UAVDT dataset, where ✓ indicates to activate the corresponding feature map.

Model	Invariant Features	Specific Features	AP	AP_{50}	AP_{70}
YOLOv5m+NDFT [11]	✓		42.6	75.1	56.8
FGFD	✓		42.4	74.9	56.9
FGFD		✓	43.8	75.3	57.1
FGFD	✓	✓	44.1	75.8	57.7

4.4.4. Effectiveness of Loss Functions

The independence of each feature vector after feature disentanglement is another essential factor in improving fine-grained domain detection performance. The ablation experiments in this section verify the proposed orthogonal loss $\mathcal{L}_{\perp}$, and the results are shown in Table 7, where $\mathcal{L}_{\text{obj}}$, $\mathcal{L}_{\text{cls}}$, and $\mathcal{L}_{\text{loc}}$ are the objectness loss, classification loss, and localization loss in the baseline model YOLOv5m, respectively. $\mathcal{L}_{\text{di}}$ and $\mathcal{L}_{\text{ds}}$ are used to learn fine-grained domain-invariant and specific features, respectively. $\mathcal{L}_{\perp}$ is used to increase the independence of fine-grained domain-invariant features and specific features. From Table 7, we can see that the proposed orthogonal loss helps AP_{70} improve by 1.1.

Table 7. Effects of proposed orthogonal loss on the UAVDT dataset, where ✓ indicates the use of the corresponding loss function.

Model	$\mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{loc}}$	$\mathcal{L}_{\text{di}} + \mathcal{L}_{\text{ds}}$	$\mathcal{L}_{\perp}$	AP	AP_{50}	AP_{70}
FGFD	✓			41.6	73.0	54.1
FGFD	✓	✓		43.0	74.8	56.6
FGFD	✓	✓	✓	44.1	75.8	57.7

4.5. Comparisons with the State of the Art

4.5.1. UAVDT

We compared the proposed method with the currently popular fine-grained domain object detectors on the UAVDT [33] dataset. The comparison methods include the fine-grained domain-object-detection algorithm NDFT proposed by Wu et al. [11], an improved version A-NDFT based on NDFT by Lee et al. [12], and a multi-branched fine-grained domain detector proposed by Kiefer et al. [13]. In addition, Faster RCNN [34] and YOLOv5m [19] are added to the comparison as two-stage and single-stage baseline models, respectively. Table 8 shows the quantitative evaluation results on the UAVDT dataset, where NDFT+FPN is the combined method of NDFT and FPN [35], and it is also the model with the highest

AP reported on UAVDT in the literature [11]. Note that all algorithm implementations used the same Pytorch framework and did not use acceleration techniques such as TensorRT. The experimental results show that the single-stage detector YOLOv5 and our FGFD have significant advantages over the two-stage detector, obtaining higher AP at a lower input image resolution, and the algorithm inference time is much less than the two-stage detector. Furthermore, our proposed FGFD improves AP by 5.7 compared to NDFT+FPN and by 4.2 AP compared to YOLOv5m, and the inference time is comparable to YOLOv5m. It is more conducive to deployment on existing airborne computing platforms. Figure 4 shows some representative examples of visual comparisons.

Table 8. Quantitative results on the UAVDT dataset.

Model	Input Size	Backbone	AP_{70}	Avg. Time (ms)
Faster RCNN [34]	800 × 1280	ResNet-101	45.6	136.2
NDFT [11]	800 × 1280	ResNet-101	47.9	138.4
NDFT+FPN	800 × 1280	ResNet-101	52.0	106.1
A-NDFT [12]	800 × 1280	ResNet-101	48.1	138.1
Kiefer et al. [13]	800 × 1280	ResNet-101	49.4	125.7
YOLOv5m [19]	800 × 1280	CSPDarknet	53.5	28.5
FGFD (ours)	800 × 1280	CSPDarknet	**57.7**$_{\uparrow 5.7}$	29.2

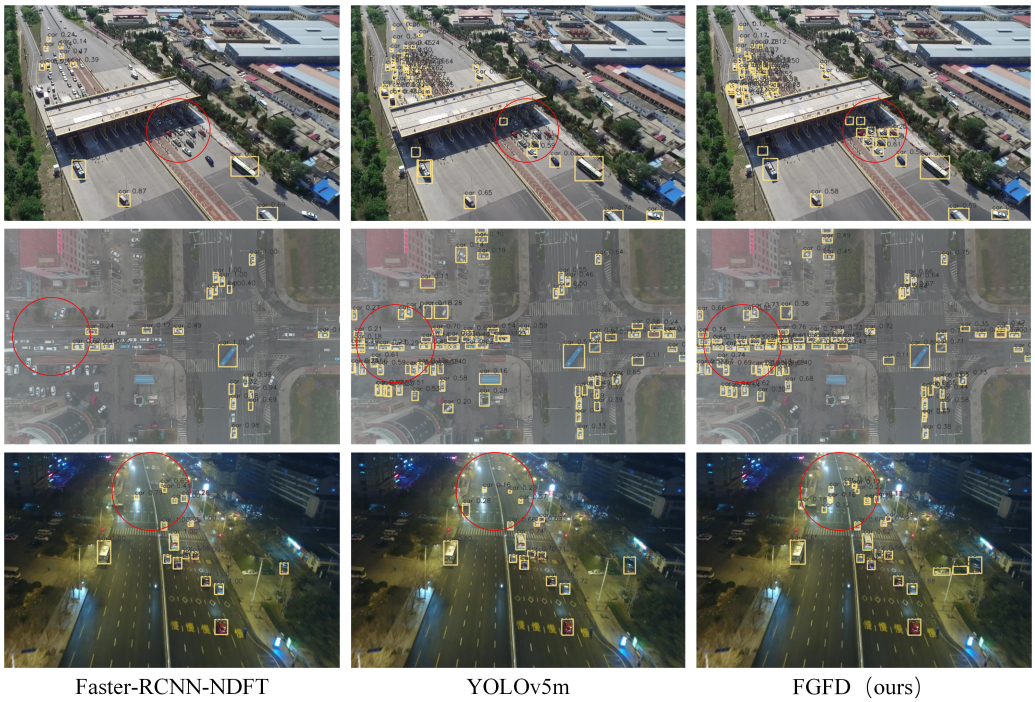

Figure 4. Several visualization results of the proposed algorithm compared with NDFT and YOLOv5m. The red circles are some representative hard-to-detect objects.

4.5.2. VisDrone

We validated the proposed FGFD on VisDrone dataset using the same experimental parameters, $\mu_1 = \mu_2 = 0.05$ and $\mu_3 = 0.2$. The comparative experimental results on VisDrone are shown in Table 9. The algorithms involved in the comparison include the two-stage baseline model Faster RCNN with FPN, and the single-stage baseline models YOLOv5s, YOLOv5m, and YOLOv5l. Moreover, the popular fine-grained domain-object-detection algorithms NDFT-DE-FPN [11] , PG-YOLO [36], ASNet [37], and Kiefer et al. [13] are also added to the comparison.

Table 9. Quantitative results on the VisDrone dataset.

Model	Input Size	Backbone	AP_{50}	Avg. Time (ms)
Faster RCNN+FPN	800 × 1280	ResNet-101	40.0	108.5
NDFT-DE-FPN [11]	800 × 1280	ResNeXt-101 64-4d	52.8	227.8
Kiefer et al. [13]	800 × 1280	ResNet-101	49.6	127.1
PG-YOLO [36]	800 × 1280	CSPDarknet	49.6	53.3
ASNet [37]	800 × 1280	ResNet-50	52.3	92.1
YOLOv5s [19]	800 × 1280	CSPDarknet	44.8	21.5
YOLOv5m [19]	800 × 1280	CSPDarknet	51.1	32.7
YOLOv5l [19]	800 × 1280	CSPDarknet	52.0	51.9
FGFD (ours)	800 × 1280	CSPDarknet	**55.2**$_{\uparrow 2.4}$	32.9

NDFT-DE-FPN is based on the best-performing single model reported in the leaderboard [38], which utilized FPN with a ResNeXt-101 64-4d backbone, and then the fine-grained domain learning module NDFT proposed by [11] is added on it. From Table 9, we can see that the proposed FGFD also leads the inference speed and AP on the VisDrone dataset. Specifically, FGFD leads to a 2.4-point improvement in mAP, indicating a substantial increase in detection accuracy. This improvement suggests that the fine-grained feature-disentanglement technique effectively enhances the model's ability to detect and classify objects due to the better handling of domain-invariant and domain-specific features.The inference speed gains are even more striking. FGFD's processing speed is 6.9 times faster than that of NDFT-DE-FPN, which underscores the efficiency of FGFD, making it particularly suitable for real-time applications where fast processing is crucial, such as drone-based object detection tasks. Additionally, comparing FGFD with a single-stage detector baseline, YOLOv5m, reveals further insights. FGFD manages to improve mAP by 4.1 points while maintaining a comparable inference speed. This balance between accuracy and speed suggests that FGFD is a valuable upgrade over existing single-stage detectors, offering a boost in accuracy without sacrificing efficiency.

Hanging flight experiment: We further validated the performance of FGFD in real environments by deploying the model learned on the VisDrone training dataset to our flight platform (see Figure 5). The flight platform we use is a self-developed 5 kg class quadrotor drone equipped with the Pixhawk open-source flight controller and an optical pod. The gimbal pitch angle, flight altitude, and time information when taking pictures can be obtained in real time during the flight. To ensure data consistency, we marked images taken below 20 m as low altitude, images taken from 20 m to 60 m as medium altitude, and images taken above 60 m as high altitude. Meanwhile, images taken with a pitch angle of -10° to 10° were marked as front view, images taken with a pitch angle of 10° to 80° were marked as side view, and images taken with a pitch angle of 80° to 100° were marked as bird view. The effectiveness of FGFD is verified by the hanging flying experiment, and some representative results are shown in Figure 6. For example, when the drone is flying at a low altitude, the FGFD can stably detect larger objects, and the false detection rate of small objects is low. In high-altitude flight, FGFD can stably detect dense small objects and greatly reduce the false-detection rate for large objects, see the comparison within the red circle. During the deployment phase, we accelerated the detection network with TensorRT technology and used INT8 quantization for inference. Some speed comparisons are shown in Table 10. On an onboard computer with Nvidia Jetson Xavier NX, the proposed algorithm can process images with 720P (1280 × 720) resolution at a rate of 20 ($\pm$2) Hz.

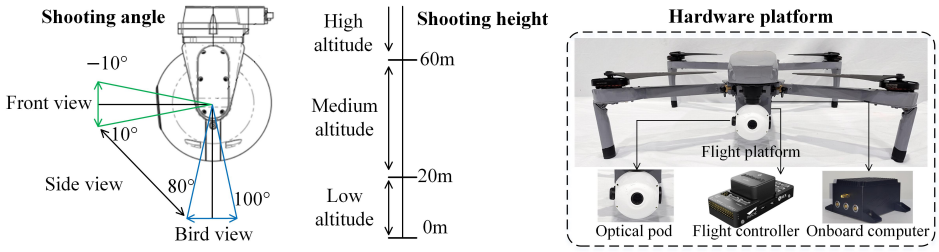

Figure 5. Experimental flight platform hardware and definition of flight height and gimbal angle conditions.

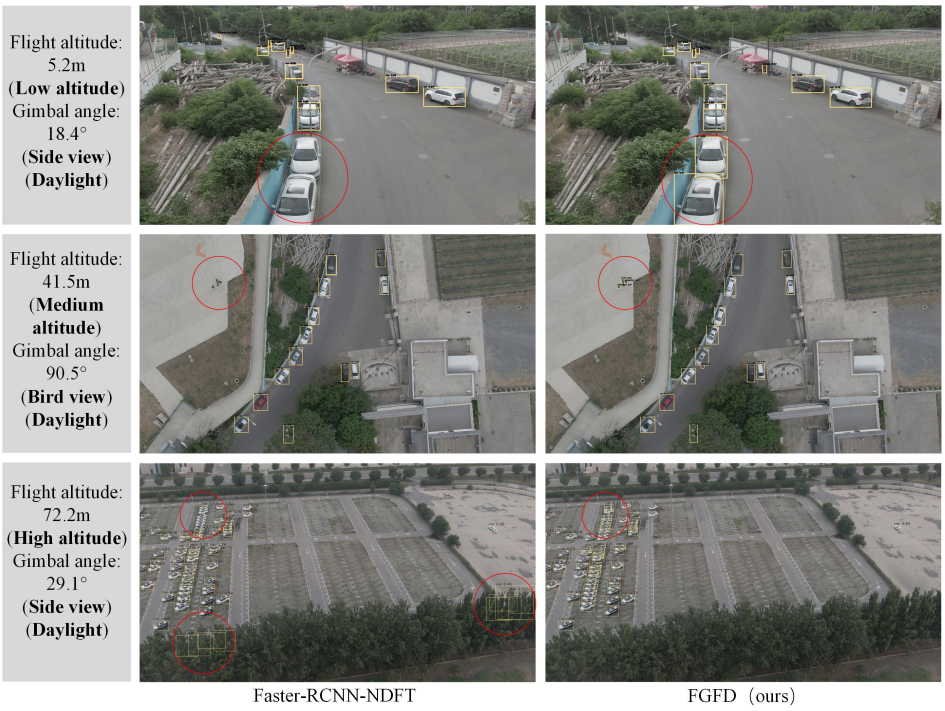

Figure 6. Several visualization results of the proposed algorithm compared to the NDFT in our hanging flight experiment.

Table 10. Comparison of the inference speed of detection models on the Nvidia Jetson Xavier NX onboard computer, where ✓ indicates whether to use TensorRT and INT8 acceleration.

Model	Input Size	Backbone	TensorRT	INT8	FPS
Faster RCNN+FPN	720 × 1280	ResNet-101	✓	✓	$4_{\pm 1}$
YOLOv5m	720 × 1280	CSPDarknet	✓	✓	$21_{\pm 2}$
FGFD (ours)	720 × 1280	CSPDarknet	✓	✓	$20_{\pm 2}$

5. Conclusions

This paper proposes a new fine-grained domain-object-detection algorithm, FGFD. The algorithm effectively improves the performance of fine-grained domain object detection by learning both domain-invariant and domain-specific features using imaging conditional decomposition. Meanwhile, the fine-grained domain mix augmentation proposed in this paper combines the advantages of domain-invariant features and domain-specific features and enables single-stage detectors to use airborne sensors' "free" data to improve object detection AP while ensuring data augmentation performance.

Whether flying at low or high altitudes, FGFD can effectively respond to changes in imaging conditions such as altitude and weather, thereby improving generalization and accuracy. The results show that it has good performance and fast inference speed. It is more convenient to deploy the algorithm on airborne platforms with limited computing power.

Author Contributions: Conceptualization, R.J., Y.Q. and Y.N.; methodology and software, R.J. and Z.J.; investigation, X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China grant number 62206020 and Civilian Aircraft Research (MJG5-1N21).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: VisDrone: https://github.com/VisDrone/VisDrone-Dataset (accessed on 1th May 2024); UAVDT: https://sites.google.com/view/grli-uavdt/ (accessed on 20 April 2024).

Acknowledgments: We would like to thank the Beijing Key Laboratory of UAV Autonomous Control for providing the hardware equipment and venue for practical flying.

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

References

1. Jin, R.; Owais, H.M.; Lin, D.; Song, T.; Yuan, Y. Ellipse proposal and convolutional neural network discriminant for autonomous landing marker detection. *J. Field Robot.* **2019**, *36*, 6–16. [CrossRef]
2. Shao, W.; Kawakami, R.; Yoshihashi, R.; You, S.; Kawase, H.; Naemura, T. Cattle detection and counting in UAV images based on convolutional neural networks. *Int. J. Remote Sens.* **2020**, *41*, 31–52. [CrossRef]
3. Tijtgat, N.; Van Ranst, W.; Goedeme, T.; Volckaert, B.; De Turck, F. Embedded real-time object detection for a UAV warning system. In Proceedings of the ICCVW, Venice, Italy, 22–29 October 2017; pp. 2110–2118.
4. Zhou, Y.; Rui, T.; Li, Y.; Zuo, X. A UAV patrol system using panoramic stitching and object detection. *Comput. Electr. Eng.* **2019**, *80*, 106473. [CrossRef]
5. Song, S.; Yu, H.; Miao, Z.; Zhang, Q.; Lin, Y.; Wang, S. Domain adaptation for convolutional neural networks-based remote sensing scene classification. *IEEE Geosci. Remote Sens.* **2019**, *16*, 1324–1328. [CrossRef]
6. Lu, X.; Gong, T.; Zheng, X. Multisource compensation network for remote sensing cross-domain scene classification. *IEEE Trans. Geosci. Remote* **2019**, *58*, 2504–2515. [CrossRef]
7. Deng, X.; Yang, H.L.; Makkar, N.; Lunga, D. Large scale unsupervised domain adaptation of segmentation networks with adversarial learning. In Proceedings of the IGARSS, Yokohama, Japan, 28 July–2 August 2019; pp. 4955–4958.
8. Koga, Y.; Miyazaki, H.; Shibasaki, R. A method for vehicle detection in high-resolution satellite images that uses a region-based object detector and unsupervised domain adaptation. *Remote Sens.* **2020**, *12*, 575. [CrossRef]
9. Tasar, O.; Giros, A.; Tarabalka, Y.; Alliez, P.; Clerc, S. Daugnet: Unsupervised, multisource, multitarget, and life-long domain adaptation for semantic segmentation of satellite images. *IEEE Trans. Geosci. Remote* **2020**, *59*, 1067–1081. [CrossRef]
10. Tasar, O.; Tarabalka, Y.; Giros, A.; Alliez, P.; Clerc, S. Standardgan: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization. In Proceedings of the CVPRW, Virtual, 14–19 June 2020; pp. 192–193.
11. Wu, Z.; Suresh, K.; Narayanan, P.; Xu, H.; Kwon, H.; Wang, Z. Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach. In Proceedings of the ICCV, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1201–1210.
12. Lee, C.; Seo, J.; Jung, H. Training Domain-invariant Object Detector Faster with Feature Replay and Slow Learner. In Proceedings of the CVPR, Virtual, 19–25 June 2021; pp. 1172–1181.
13. Kiefer, B.; Messmer, M.; Zell, A. Diminishing Domain Bias by Leveraging Domain Labels in Object Detection on UAVs. In Proceedings of the ICAR, Ljubljana, Slovenia, 6–10 December 2021; pp. 523–530.
14. Jia, D.; Yuan, Y.; He, H.; Wu, X.; Yu, H.; Lin, W.; Sun, L.; Zhang, C.; Hu, H. DETRs with Hybrid Matching. *arXiv* **2023**, arXiv:2207.13080.
15. Pu, Y.; Wang, Y.; Xia, Z.; Han, Y.; Wang, Y.; Gan, W.; Wang, Z.; Song, S.; Huang, G. Adaptive Rotated Convolution for Rotated Object Detection. *arXiv* **2023**, arXiv:2303.07820.
16. Pu, Y.; Liang, W.; Hao, Y.; Yuan, Y.; Yang, Y.; Zhang, C.; Hu, H.; Huang, G. Rank-DETR for High Quality Object Detection. *arXiv* **2023**, arXiv:2310.08854.
17. Shen, Y.; Geng, Z.; Yuan, Y.; Lin, Y.; Liu, Z.; Wang, C.; Hu, H.; Zheng, N.; Guo, B. V-DETR: DETR with Vertex Relative Position Encoding for 3D Object Detection. *arXiv* **2023**, arXiv:2308.04409.
18. Yang, L.; Zheng, Z.; Wang, J.; Song, S.; Huang, G.; Li, F. AdaDet: An Adaptive Object Detection System Based on Early-Exit Neural Networks. *IEEE Trans. Cogn. Dev. Syst.* **2024**, *16*, 332–345. [CrossRef]
19. Glenn, J. YOLOv5 in PyTorch. 2022. Available online: https://github.com/ultralytics/yolov5 (accessed on 20 April 2022).
20. Weir, N.; Lindenbaum, D.; Bastidas, A.; Etten, A.V.; McPherson, S.; Shermeyer, J.; Kumar, V.; Tang, H. Spacenet mvoi: A multi-view overhead imagery dataset. In Proceedings of the ICCV, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 992–1001.
21. Lee, H.Y.; Tseng, H.Y.; Huang, J.B.; Singh, M.; Yang, M.H. Diverse image-to-image translation via disentangled representations. In Proceedings of the ECCV, Munich, Germany, 8–14 September 2018; pp. 35–51.

22. Ridgeway, K.; Mozer, M.C. Learning deep disentangled embeddings with the f-statistic loss. In Proceedings of the NeurIPS, Montreal, QC, Canada, 3–8 December 2018; pp. 1–10.
23. Peng, X.; Huang, Z.; Sun, X.; Saenko, K. Domain agnostic learning with disentangled representations. In Proceedings of the ICML, Long Beach, CA, USA, 9–15 June 2019; pp. 5102–5112.
24. Wu, A.; Liu, R.; Han, Y.; Zhu, L.; Yang, Y. Vector-Decomposed Disentanglement for Domain-Invariant Object Detection. In Proceedings of the ICCV, Virtual, 11–17 October 2021; pp. 9342–9351.
25. Sugiyama, M.; Kawanabe, M. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*; MIT Press: Cambridge, MA, USA , 2012.
26. Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; Van Gool, L. Domain adaptive faster r-cnn for object detection in the wild. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3339–3348.
27. Saito, K.; Ushiku, Y.; Harada, T.; Saenko, K. Strong-weak distribution alignment for adaptive object detection. In Proceedings of the CVPR, Long Beach, CA, USA, 15–20 June 2019; pp. 6956–6965.
28. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
29. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
30. Ramamonjison, R.; Banitalebi-Dehkordi, A.; Kang, X.; Bai, X.; Zhang, Y. Simrod: A simple adaptation method for robust object detection. In Proceedings of the ICCV, Virtual, 11–17 October 2021; pp. 3570–3579.
31. Smith, L.N.; Topin, N. Super-convergence: Very fast training of neural networks using large learning rates. In Proceedings of the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. International Society for Optics and Photonics, Baltimore, MD, USA, 14–18 April 2019; Volume 11006, pp. 1–612.
32. 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 ECCV, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland , 2014; pp. 740–755.
33. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the ECCV, Munich, Germany, 8–14 September 2018; pp. 370–386.
34. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the NeurIPS, Montreal, QC, Canada, 7–12 December 2015; pp. 1–9.
35. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
36. Dong, C.; Pang, C.; Li, Z.; Zeng, X.; Hu, X. PG-YOLO: A Novel Lightweight Object Detection Method for Edge Devices in Industrial Internet of Things. *IEEE Access* **2022**, *10*, 123736–123745. [CrossRef]
37. Froehlich, S.; Klemmer, L.; Große, D.; Drechsler, R. ASNet: Introducing Approximate Hardware to High-Level Synthesis of Neural Networks. In Proceedings of the 2020 IEEE 50th International Symposium on Multiple-Valued Logic (ISMVL), Miyazaki, Japan, 9–11 November 2020; pp. 64–69. [CrossRef]
38. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Ling, H.; Hu, Q.; Nie, Q.; Cheng, H.; Liu, C.; Liu, X.; et al. Visdrone-det2018: The vision meets drone object detection in image challenge results. In Proceedings of the ECCVW, Munich, Germany, 8–14 September 2018; pp. 1–30.