Generalized UAV Object Detection via Frequency Domain Disentanglement

Kunyu Wang, Xueyang Fu, Yukun Huang, Chengzhi Cao, Gege Shi, Zheng-Jun Zha*
University of Science and Technology of China, China
{kunyuwang@mail., xyfu@, kevinh@mail., chengzhicao@mail., shigg@mail., zhazj@}ustc.edu.cn

Abstract

When deploying the Unmanned Aerial Vehicles object detection (UAV-OD) network to complex and unseen real-world scenarios, the generalization ability is usually reduced due to the domain shift. To address this issue, this paper proposes a novel frequency domain disentanglement method to improve the UAV-OD generalization. Specifically, we first verified that the spectrum of different bands in the image has different effects to the UAV-OD generalization. Based on this conclusion, we design two learnable filters to extract domain-invariant spectrum and domain-specific spectrum, respectively. The former can be used to train the UAV-OD network and improve its capacity for generalization. In addition, we design a new instance-level contrastive loss to guide the network training. This loss enables the network to concentrate on extracting domain-invariant spectrum and domain-specific spectrum, so as to achieve better disentangling results. Experimental results on three unseen target domains demonstrate that our method has better generalization ability than both the baseline method and state-of-the-art methods.

1. Introduction

Unmanned Aerial Vehicles (UAV) equipped with cameras have been exploited in a wide variety of applications, opening up a new frontier for computer vision applications [7, 11, 22, 28]. As one of the fundamental functions for the UAV-based applications, UAV object detection (UAV-OD) has garnered considerable interest [23, 31, 38]. However, the large mobility of UAV-mounted cameras leads to an unpredictable operating environment. The domain shift that occurs when applying a UAV-OD network that has been trained on a given dataset (*i.e.*, source domain) to unseen real-world data (*i.e.*, target domain) typically results in inadequate performance. In particular, unseen real-world data consists of unexpected and unknown samples, such as images taken in various scene structures, diverse illumination conditions, and adverse weather conditions. Therefore, it is crucial to improve the generalization ability of UAV-OD.

(a) Baseline (b) Ours

Figure 1. Detection results on unseen target domains. UAV-OD network is trained on daylight images and tested on images with various scene structures (1st row), diverse illumination conditions (2nd row), and adverse weather conditions (3rd row). Green rectangular boxes denote new correct detections beyond the baseline.

To alleviate the domain shift impact, existing methods broadly come in two flavors: Domain Adaptation (DA) [3, 5, 8, 16, 17, 26, 37] and Domain Generalization (DG) [19, 20, 27, 30, 40]. In general, DA aims to tackle the domain shift problem by learning domain-invariant/aligned features between the source and target domains. However, DA methods cannot be readily employed when it is hard to guarantee the accessibility of the target data. The requirement to access both source and target data restricts the applicability of DA approaches.

Recently, considerable attention has been drawn to the field of DG. The goal of DG is to learn a model using data from a single or multiple related but distinct source domains so that the model can generalize well under distri-

*Corresponding author. This work was supported by the National Key R&D Program of China under Grant 2020AAA0105702, the National Natural Science Foundation of China (NSFC) under Grants 62225207, 62276243 and U19B2038, the University Synergy Innovation Program of Anhui Province under Grant GXXT-2019-025.

Reject band	Various Scene			Diverse Illumination			Adverse Weather			Average		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Null (full band)	66.0	37.6	36.7	11.1	3.4	4.8	42.3	14.9	19.6	39.8	18.6	20.4
$\alpha = 0, \beta = 0.01$	60.0	30.2	32.6	6.4	1.9	2.75	39.5	16.1	19.0	35.3	16.1	18.1
$\alpha = 0.01, \beta = 0.1$	61.4	30.3	32.8	**39.1**	**15.9**	**19.8**	**42.6**	**18.6**	**20.8**	**47.7**	**21.6**	**24.5**
$\alpha = 0.1, \beta = 1$	**70.2**	**35.1**	**37.1**	29.4	10.2	13.6	38.2	10.6	16.7	45.9	18.6	22.5

Table 1. We conduct preliminary experiments to explore whether different spectral bands contribute equally to the UAV-OD network's generalization ability. The specified bands of source domain images are filtered out for training according to the reject band. For testing, the generalization performance of the UAV-OD network is evaluated on three unseen target domains. We adopt the evaluation protocols AP_{50}, AP_{75}, and AP. "Average" refers to the average generalization performance across three unseen target domains. We can conclude that eliminating various bands has distinct effects on the generalization of unseen target domains for UAV-OD network.

bution shifts [43]. Most existing DG methods [19, 30, 40] focus on decoupling object-related features from global features via spatial vanilla convolution. However, unlike generic object detection scenarios based on surveillance or other ground-based cameras, the rapid movement of UAV-mounted cameras leads to severe changes in the global appearance. For UAV-OD scenarios where the global appearance changes, it is essential to explore global dependency for better disentanglement. The spatial vanilla convolution, which only emphasizes local pixel attention, cannot fully explore global dependency, leading to suboptimal disentanglement and generalization results.

Inspired by the spectral theorem that the frequency domain obeys the nature of global modeling, we propose to improve the UAV-OD generalization ability via frequency domain disentanglement. We first conduct preliminary experiments, i.e., exploring whether all spectrum bands contribute equally to the generalization for the UAV-OD task, to gain insight into how to implement our idea. If not, we can extract the spectrum that is conducive to generalization and use it to train the UAV-OD network to enhance its generalization. Specifically, we first convert each source domain image $x \in \mathbb{R}^{H\times W\times C}$ into frequency space through Fast Fourier Transform (FFT) [24]:

$$\mathcal{F}(x)(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} x(h,w)e^{-j2\pi\left(\frac{h}{H}u+\frac{w}{W}v\right)}. \quad (1)$$

The frequency space signal $\mathcal{F}(x)$ can be further decomposed to an amplitude spectrum $\mathcal{A}(x)$ and a phase spectrum $\mathcal{P}(x)$, which is expressed as:

$$\begin{aligned}\mathcal{A}(x)(u,v) &= \left[\mathcal{R}^2(x)(u,v) + \mathcal{I}^2(x)(u,v)\right]^{1/2}, \\ \mathcal{P}(x)(u,v) &= \arctan\left[\frac{\mathcal{I}(x)(u,v)}{\mathcal{R}(x)(u,v)}\right],\end{aligned} \quad (2)$$

where $\mathcal{R}(x)$ and $\mathcal{I}(x)$ represent the real and imaginary part of $\mathcal{F}(x)$. For each source image, we filter out the bands of the amplitude spectrum $\mathcal{A}(x)$ between a certain upper threshold α and lower threshold β ('Reject band' in Tab. 1) with a band reject filter $f_s \in \mathbb{R}^{H\times W\times C}$ and obtain the remaining amplitude spectrum $\hat{\mathcal{A}}(x)$:

$$f_s(i,j) = \begin{cases} 1, & i \in [\frac{\alpha H}{2}, \frac{\beta H}{2}] \cup [\frac{(1-\alpha)H}{2}, \frac{(1-\beta)H}{2}] \\ & j \in [\frac{\alpha W}{2}, \frac{\beta W}{2}] \cup [\frac{(1-\alpha)W}{2}, \frac{(1-\beta)W}{2}] \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

$$\mathcal{A}(x) = \hat{\mathcal{A}}(x) \otimes f_s, \quad (4)$$

where $\otimes$ denotes element-wise multiplication. $\hat{\mathcal{A}}(x)$ is then fed to Inverse Fast Fourier Transform (IFFT) with $\mathcal{P}(x)$ to generate the remaining image $\hat{x}$ which is utilized to train the UAV-OD network. After training, we apply the UAV-OD network to three unseen target domains to evaluate the generalization ability. The experimental results are presented in Tab. 1. We can observe that removing different bands has varying effects on generalization to three unseen target domains. Therefore, we can conclude that different bands contribute differently to the UAV-OD generalization.

Based on the above observation, we employ two learnable filters to identify and extract the domain-invariant and domain-specific spectrums. The former contributes positively to generalization, while the latter is the opposite. Furthermore, we design a new instance-level contrastive loss to aid in learning the learnable filters, enabling them to concentrate on disentangling the two different spectrums. By optimizing the instance-level contrastive loss, the instance features of those are encouraged to contain domain-invariant characteristics shared by target objects, and the domain-specific characteristics shared in the source domain, respectively. In this way, the UAV-OD network can generalize well on unseen target domains. For experiment settings, we focus on learning a single-domain generalized UAV-OD network, which is more challenging [30]. We further validate the network on three unseen target domains, including various scene structures, diverse illumination conditions, and adverse weather conditions, demonstrating superior generalization ability, as shown in Fig. 1.

Our main contributions are highlighted as follows:

- We provide a new perspective to improve the generalization ability of the UAV-OD network on unseen target domains. To our best knowledge, this is the first attempt to learn generalized UAV-OD via frequency domain disentanglement.
- Based on the frequency domain disentanglement, we propose a new framework that utilizes two learnable filters to extract the domain-invariant and domain-specific spectrum and design an instance-level contrastive loss to guide the disentangling process.
- Extensive experiments on three unseen target domains reveal that our method enables the UAV-OD network to achieve superior generalization performance in comparison to the baseline and state-of-the-art methods.

2. Related work

Domain Adaptive Object Detection. Various domain adaptive object detection methods [1, 3, 5, 8, 26, 33, 34, 37, 41, 42, 45] have been proposed to eliminate domain shifts during training and testing. Typically, they achieved alignment at the feature or pixel level by utilizing the target data distribution. For example, Chen *et al*. [5] utilized adversarial training to align global feature distributions of the source and target domains. Zhuang *et al*. [45] aligned feature distributions at both image and instance levels. However, these methods cannot be reliably applied in situations where the accessibility of the target data cannot be guaranteed. The need to acquire both source and target data limits their applicability. Therefore, we focus on domain generalization.

Domain Generalized Object Detection. Recently, domain generalized object detection [15, 19, 20, 27, 30, 32, 40] has garnered increasing interest as it does not require data from the target domain, and its performance exceeds that of domain adaptive object detection. For example, Lin *et al*. [19] proposed a Domain-Invariant Disentangled Network for learning a universal object detector by decoupling image-level and instance-level disentanglement across multiple source domains. Zhang *et al*. [40] proposed a Region Aware Proposal Reweighting method to learn weights for proposals to eliminate the statistical dependence between features and disentangle relevant features and irrelevant features to improve detectors' generalization under distribution shifts. Wu *et al*. [30] focused on single-domain generalized object detection and proposed a Cyclic-Disentangle Self-Distillation method to disentangle domain-invariant representations for object detection. Wu *et al*. [32] proposed to decouple domain-robust features via an adversarial training framework dubbed Nuisance Disentangled Feature Transform. However, most existing methods decouple domain-invariant features via spatial vanilla convolution, which only emphasizes local pixel attention. For UAV-OD scenarios where the global appearance varies significantly, it is essential to explore global dependency for better disentanglement. Compared to existing methods, we consider this property and propose decoupling in the frequency domain that obeys the nature of global modeling, achieving superior generalization results and providing a novel perspective for learning generalized UAV-OD.

Frequency-based Domain Generalization. Most existing frequency-based DG methods decouple invariant and specific components based on frequency prior knowledge [21, 29, 35, 36]. However, these methods do not incorporate specific task peculiarities. As the most relevant method to our work, FSDR realizes decoupling via dynamic spectrum learning. However, FSDR [12] is designed for segmentation, so the entropy-based loss is used for frequency disentanglement. As the detection includes classification and localization stages, we design a new instance-level contrastive loss to enable two learnable filters to extract invariant and specific spectrums for UAV-OD.

3. Methodology

We propose a novel framework for enhancing the generalization ability via frequency domain disentanglement. We begin by illustrating the problem definition. Next, we introduce frequency-based learnable filtering. The instance-level contrastive loss is then clarified, allowing the learnable filters to concentrate on extracting the domain-invariant spectrum. The strategy for training is presented in the final section. Fig. 2 provides an overview of the framework.

3.1. Problem Definition

Let $X_s \subset \mathbb{R}^{H\times W\times C}$ denote source domain data with height H, width W, and number of channel C, $Y_s \subset \mathbb{R}$ denote the category labels of X_s, $B_s \subset \mathbb{R}^4$ denote the bounding boxes of X_s. The source domain can be fomulated as $D_s = \left\{x_s^i, \{y_s^{ij}, b_s^{ij}\}_{j=1}^{N_i}\right\}_{i=1}^{N}$, which includes N images and each image has N_i pairs of category labels $y_s^{ij} \in Y_s$ and bounding boxes $b_s^{ij} \in B_s$. Let $D_t = \{D_t^1, \cdots, D_t^M\}$ denote M unseen target domains. Our goal is to learn a network trained on source domain D_s that generalize well on unseen target domains D_t.

3.2. Frequency-based Learnable Filtering

Given a souce domain image $x_s \in \mathbb{R}^{H\times W\times C}$, we can obtain the frequency space signal of x_s through Eq.1:

$$x_s^{\mathcal{F}} = \mathcal{F}(x_s). \tag{5}$$

The frequency signal $x_s^{\mathcal{F}}$ can be further decomposed to an amplitude spectrum $x_s^{\mathcal{A}} \in \mathbb{R}^{H\times W\times C}$ and a phase spectrum $x_s^{\mathcal{P}} \in \mathbb{R}^{H\times W\times C}$ using (Eq.2), which is expressed as:

$$x_s^{\mathcal{A}} = \mathcal{A}(x_s^{\mathcal{F}}), \quad x_s^{\mathcal{P}} = \mathcal{P}(x_s^{\mathcal{F}}). \tag{6}$$

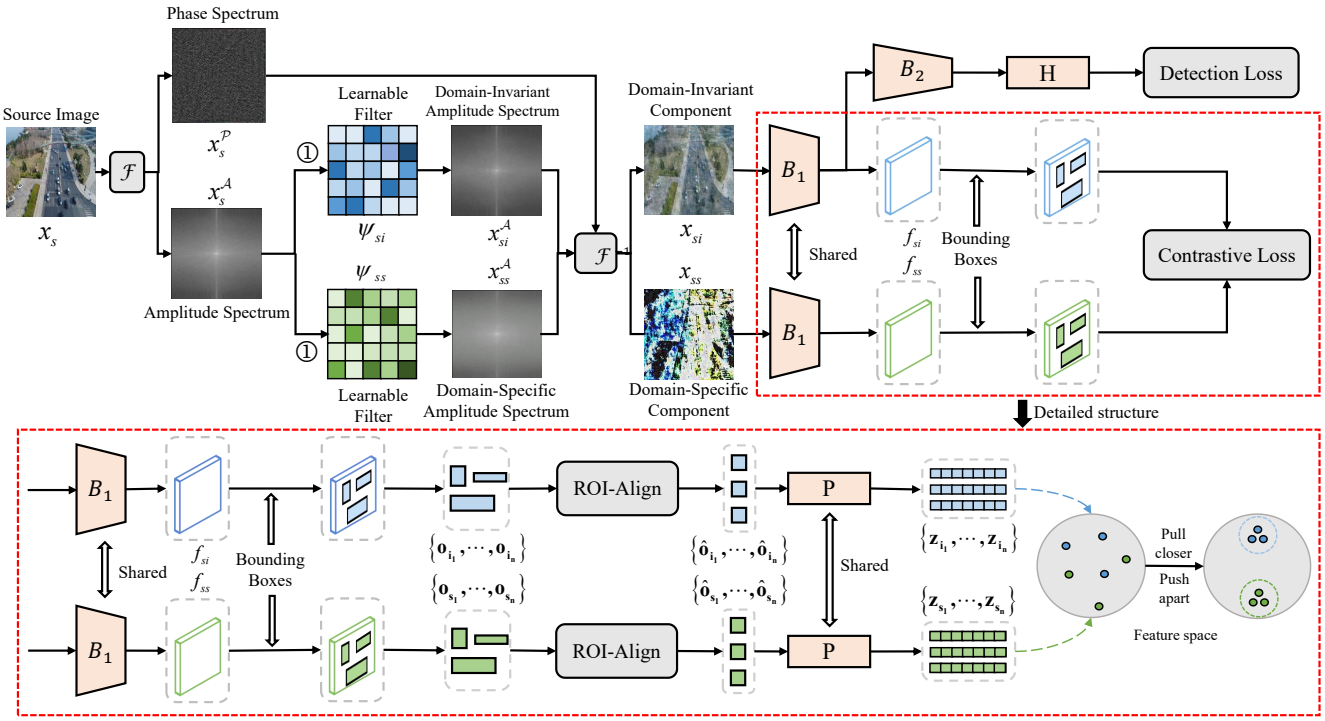


Figure 2. Overview of the proposed framework. $\mathcal{F}$ and $\mathcal{F}^{-1}$ indicate FFT and IFFT. The backbone of UAV-OD network is divided into B_1 and B_2. H represents the detection head of UAV-OD network. ROI-Align indicates ROI-Alignment operation [10], P denotes the contrastive projection head, the lines marked with ① represent element-wise multiplication, the red dashed box below shows the detailed structure of the red dashed box above. We employ two learnable filters ψ_{si} and ψ_{ss} to extract the domain-invariant spectrum $x_{si}^{\mathcal{A}}$ that contributes positively to generalization, and the domain-specific spectrum $x_{ss}^{\mathcal{A}}$ that contributes negetively to generalization from the image's amplitude spectrum $x_s^{\mathcal{A}}$. Furthermore, we design an instance-level contrastive loss to aid the training of two learnable filters, enabling it to concentrate on extracting the domain-invariant and domain-specific spectrum.

We then employ two learnable filters $\psi_{si}, \psi_{ss} \in \mathbb{R}^{H\times W\times C}$ to identify and extract the domain-invariant amplitude spectrum $x_{si}^{\mathcal{A}}$ that contribute positively to generalization, and the domain-specific amplitude spectrum $x_{ss}^{\mathcal{A}}$ that contribute negatively to generalization from the amplitude spectrum $x_s^{\mathcal{A}}$. Specifically, the learnable filters ψ_{si} and ψ_{ss} are continuous variables and each element of the learnable filters ranges from 0 to 1:

$$x_{si}^{\mathcal{A}} = x_s^{\mathcal{A}} \otimes \psi_{si}, \tag{7}$$

$$x_{ss}^{\mathcal{A}} = x_s^{\mathcal{A}} \otimes \psi_{ss}, \tag{8}$$

where $\otimes$ denotes element-wise multiplication. Then, the domain-invariant amplitude spectrum $x_{si}^{\mathcal{A}}$ and the domain-specific amplitude spectrum $x_{ss}^{\mathcal{A}}$ are fed to IFFT with $x_s^{\mathcal{P}}$ to generate the domain-invariant component x_{si} and domain-specific component x_{ss}.

3.3. Contrastive-based Frequency Disentanglement

We design an instance-level contrastive loss to enable the learnable filters to focus on extracting domain-invariant and domain-specific spectrums. Particularly, we adopt the efficient and accurate Yolov5 [14] as the base model for UAV-OD, which is composed of two parts: backbone and detection head. Firstly, we divide the backbone into two sections (i.e., B_1 and B_2) according to its depth and original structure. Given the domain-invariant component x_{si} and the domain-specific component x_{ss}, we use B_1 to obtain the domain-invariant feature $f_{si} \in \mathbb{R}^{h\times w\times c}$ and the domain-specific feature $f_{ss} \in \mathbb{R}^{h\times w\times c}$, where h, w, and c respectively denote the height, width and number of channels:

$$f_{si} = B_1(x_{si}), \quad f_{ss} = B_1(x_{ss}) \tag{9}$$

Furthermore, according to the localization labels b_s of x_s and the dimension scale between $\{x_{si}, x_{ss}\}$ and $\{f_{si}, f_{ss}\}$, we clip the domain-invariant instance-level features $\{o_i, \cdots, o_{i_n}\}$ from x_{si} and the domain-specific instance-level features $\{o_{s_1}, \cdots, o_{s_n}\}$ from x_{ss}, n represents the total number of the object in x_s. As different instance-level features have different spatial size, we utilize the RoI-Alignment operation [10] to align the spatial size of all instance-level features:

$$\{\hat{o}_{i_1}, \cdots, \hat{o}_{i_n}\} = \text{RoIAlign}(\{o_{i_1}, \cdots, o_{i_n}\}), \tag{10}$$

$$\{\hat{o}_{s_1}, \cdots, \hat{o}_{s_n}\} = \text{RoIAlign}(\{o_{s_1}, \cdots, o_{s_n}\}), \tag{11}$$

where $\hat{o}_{i_j}, \hat{o}_{s_j} \in \mathbb{R}^{s \times s \times c}$, $j \in \{1, \cdots, n\}$, s indicates the output size of RoI-Alignment. To map $\{\hat{o}_{i_1}, \cdots, \hat{o}_{i_n}\}$ and $\{\hat{o}_{s_1}, \cdots, \hat{o}_{s_n}\}$ to the space where the contrastive loss is applied, we adopt a projection head [4] P, which consists of an MLP with two hidden layers:

$$\{z_{i_1}, \cdots, z_{i_n}\} = P(\{\hat{o}_{i_1}, \cdots, \hat{o}_{i_n}\}), \tag{12}$$

$$\{z_{s_1}, \cdots, z_{s_n}\} = P(\{\hat{o}_{s_1}, \cdots, \hat{o}_{s_n}\}). \tag{13}$$

To ensure the frequency domain disentanglement via the learnable filters, we define an instance-level contrastive loss. Specifically, let $\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$ denote the dot product between l_2 normalized u and v. The contrastive loss is calculated as follows:

$$\begin{aligned} \mathcal{L}_{\text{con}} = & \sum_{z_{i_j} \in Z_i} \frac{-1}{|\hat{Z}_i|} \sum_{\substack{z_{i_k} \in \hat{Z}_i \\ y_{i_j} = y_{i_k}}} \log \frac{\exp\left(\text{sim}\left(z_{i_j}, z_{i_k}\right)/\tau\right)}{\sum_{z_a \in Z_a} \exp\left(\text{sim}\left(z_{i_j}, z_a\right)\right)} \\ & + \sum_{z_{s_j} \in Z_s} \frac{-1}{|\hat{Z}_s|} \sum_{\substack{z_{s_k} \in \hat{Z}_s \\ y_{s_j} = y_{s_k}}} \log \frac{\exp\left(\text{sim}\left(z_{s_j}, z_{s_k}\right)/\tau\right)}{\sum_{z_b \in Z_b} \exp\left(\text{sim}\left(z_{s_j}, z_b\right)\right)} \end{aligned} \tag{14}$$

where $Z_i = \{z_{i_1}, \cdots, z_{i_n}\}$, $Z_s = \{z_{s_1}, \cdots, z_{s_n}\}$, $\hat{Z}_i = Z_i - \{z_{i_j}\}$, $\hat{Z}_s = Z_s - \{z_{s_j}\}$, $Z_a = \hat{Z}_i \cup Z_s$, $Z_b = Z_i \cup \hat{Z}_s$, $|\hat{Z}_i|$ and $|\hat{Z}_s|$ represent the cardinality of $\hat{Z}_i$ and $\hat{Z}_s$, τ denotes the temperature hyper-parameter [9]. The invariant instance-level features Z_i from the same category are set as positive samples, while all others are negative. The same settings are also applied to the specific instance-level features Z_s. By optimizing $\mathcal{L}_{\text{con}}$, we pull close positive pairs and push away negative pairs, which enables two learnable filters to decouple invariant and specific spectrums.

3.4. Training With the Alternating Optimization

First, let us denote the detection loss of the UAV-OD network as $\mathcal{L}_{\text{det}}$, which contains the regression loss $\mathcal{L}_{\text{reg}}$ and the classification loss $\mathcal{L}_{\text{cls}}$:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{cls}}. \tag{15}$$

As shown in Fig. 2, the whole learnable parameters consist of two learnable filters ψ_{si} and ψ_{ss}, backbone B_1 and B_2, detection head H and projection head P. For training, we adopt the alternating strategy, which fixes one set of parameters and solving for the other set. Specifically, we divide the whole learnable parameters into two groups:

$$\theta = \{\psi_{si}, \psi_{ss}, P\}, \tag{16}$$

$$\eta = \{B_1, B_2, H\}. \tag{17}$$

At first step, we fix η and optimize θ using $\mathcal{L}_{\text{con}}$:

$$\theta^t \leftarrow \arg\min_{\theta} \lambda \mathcal{L}_{\text{con}}\left(\theta, \eta^{t-1}\right). \tag{18}$$

At second step, we fix θ and optimize η using $\mathcal{L}_{\text{det}}$:

$$\eta^t \leftarrow \arg\min_{\eta} \mathcal{L}_{\text{det}}\left(\theta^t, \eta\right). \tag{19}$$

Here λ is the hyper-parameter for balancing $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{det}}$, t is the index of alternation and $\leftarrow$ means assigning. The purpose of alternating optimization is to avoid frequency domain disentanglement and UAV-OD conflicts. Therefore, we divide the entire learnable parameters into two groups: θ for frequency domain disentanglement and η for UAV-OD.

4. Experiments

This section presents evaluations of our method, including datasets, implementation details, domain generalization results, ablation analysis, statistic analysis on learnable filters, and visualization analysis. The supplementary file contains additional experimental results, including the ablation analysis of the contrastive loss, comparisons of the training time, more visualizations, etc.

4.1. Datasets

In the real scenario, it is straightforward to collect and label the data from the daylight scene. Thus, we train our model on the daylight scene (*i.e.*, the source domain) and evaluate its generalization ability to unseen target domains (*i.e.*, various scene structures, diverse illumination conditions, and adverse weather conditions). Since UAVDT [6] includes weather annotations (daylight, nighttime, and fog), we conduct extensive experiments on UAVDT. UAVDT is comprised of 41k frames with 840k bounding boxes and is divided into three categories: car, truck, and bus. As the class distribution of UAVDT is highly unbalanced, the latter two classes fill fewer than 5 % of bounding boxes, we merged the three classes into a single class, following the authors' convention in [6]. UAVDT can be separated into three sections based on the weather annotations: 23741 daylight images, 11489 nighttime images, and 2492 foggy images. We choose the nighttime portion to replicate the diverse illumination scenario, the foggy portion to simulate the adverse weather scenario, and 2850 daylight images with different scene structures compared with the remaining daylight images. The remaining daylight images serve as the UAV-OD network's training set.

In addition, to validate the capacity to generalize across datasets, we further train the UAV-OD network with VisDrone2019-VID dataset [44]. VisDrone2019-VID dataset consists of 24201 training images captured by drone platforms in different places at different heights. Images are manually labeled with bounding boxes and ten predefined classes (*i.e.*, car, van, bus, truck, etc.). We arbitrarily

Method	Various Scene			Diverse Illumination			Adverse Weather			Average		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Baseline	66.0	37.6	36.7	11.1	3.4	4.8	42.3	14.9	19.6	39.8	18.6	20.4
JiGen [2]	61.3	26.4	35.8	33.5	12.9	15.9	45.5	19.5	**22.7**	46.8	19.6	24.8
RSC [13]	73.3	48.7	44.3	14.6	6.2	7.3	47.1	16.6	21.2	45.0	23.8	24.3
StableNet [39]	75.0	48.8	44.9	18.6	9.0	9.5	47.5	17.0	21.1	47.0	24.9	25.2
Single-DGOD [30]	73.7	49.3	43.6	27.5	11.9	13.8	47.3	18.7	22.8	49.5	26.6	26.7
Ours	**75.1**	**49.7**	**45.3**	**39.0**	**18.5**	**20.7**	**48.0**	**17.2**	22.3	**54.0**	**28.4**	**29.4**

Table 2. Comparisons of the domain generalization results. All methods are trained on daylight images from UAVDT [6] and tested on daylight images with various scene structures, nighttime images simulating diverse illumination conditions, and foggy images simulating adverse weather conditions from UAVDT. The average generalization performance across three unseen target domains is "average".

Method	Various Scene			Diverse Illumination			Adverse Weather			Average		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Baseline	53.4	21.8	26.0	31.0	13.6	15.4	37.6	7.6	14.5	40.7	14.3	18.6
JiGen [2]	62.8	28.6	32.3	36.5	10.1	15.8	39.5	**11.8**	17.4	46.3	16.8	21.8
RSC [13]	65.3	25.5	31.3	33.5	11.6	15.3	39.3	9.0	16.3	46.0	15.4	21.0
StableNet [39]	64.3	25	30.6	31.7	15.2	16.3	**40.7**	9.9	16.8	45.6	16.7	21.2
Single-DGOD [30]	62.9	26.3	30.7	**36.8**	16.7	18.5	34.8	7.3	13.9	44.8	16.8	21.0
Ours	**65.8**	**33.9**	**35.4**	36.4	**19.5**	**19.3**	40.6	11.3	**18.4**	**47.6**	**21.6**	**24.4**

Table 3. Comparisons of the domain generalization results. All methods are trained on daylight images from VisDrone2019-VID [44] and tested on daylight images with various scene structures, nighttime images simulating diverse illumination conditions, and foggy images simulating adverse weather conditions, from UAVDT.

select 16238 daylight images for the training set and evaluate its cross-dataset generalization performance on the three above-mentioned single unseen target domains.

4.2. Implementation Details

Our approach is implemented in Pytorch with eight NVIDIA 1080ti GPUs. We train the framework for 300 epochs with a batch size of 128. The UAV-OD network is optimized using Adam [18] with a learning rate of 0.001, a momentum of 0.9, lambda learning rate decay, and linear warmup for the first five epochs. The projection head is optimized using SGD [25] with a learning rate of 0.05 and weight decay of 10-4. The linear warmup is used for the first five epochs and decays the learning rate with the step decay schedule. The learnable filters are optimized using SGD with a learning rate of 0.001, weight decay of 10-4, and step learning rate decay. The temperature parameter τ is set to 0.7. The hyper-parameter λ is set to 0.15. For evaluation protocol, we evaluate detectors with the widely accepted criterion, AP, AP_{50}, and AP_{75}.

4.3. Domain Generalization Results

As the majority of current DG methods are intended for image classification and there are few works [30] focusing on single domain generalized object detection, we further select and develop various model-agnostic DG methods [2, 13, 39] to learn a single domain generalized object detection network as comparative approaches besides [30]. As [30] is not yet open-source, we reproduce its code according to the paper [30]. Jigen [2] is a strategy for representative representation enhancement of DG in a self-supervised manner. As proposed in the paper [2], we add a jigsaw classifier to Yolov5 and minimize the image-level jigsaw loss. RSC [13] is a dropout-based DG approach that discards dominating features triggered on the training data iteratively. StableNet [39] proposes sample reweighting to enhance generalization under distribution shifts. We directly compute RFF for image representations and implement image-wise reweighting.

Tab. 2 shows the generalization results of various methods trained on daylight images from UAVDT. Eleven out of the entire twelve performance metrics demonstrate that our strategy delivers optimal performance. In terms of average generalization ability, our method outperforms the baseline method by 14.2%, 9.8%, and 9.0%, and the runner-up by 4.5%, 1.8%, and 2.7% on AP_{50}, AP_{75}, and AP, respectively. Baseline, JiGen, RSC, StableNet, and Single-DGOD, which are less generalizable than our method, tend to overfit the source domain, suffering from performance degradation on the unseen target domains due to the large domain shift.

Disentangle	Various Scene			Diverse Illumination			Adverse Weather			Average		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Baseline	66.0	37.6	36.7	11.1	3.4	4.8	42.3	14.9	19.6	39.8	18.6	20.4
Spatial	66.7	41.1	39.2	38.1	16.2	17.8	47.9	16.5	21.9	50.9	24.6	26.3
Frequency	**75.1**	**49.7**	**45.3**	**39.0**	**18.5**	**20.7**	**48.0**	**17.2**	**22.3**	**54.0**	**28.4**	**29.4**

Table 4. Ablation analysis of the frequency domain disentanglement. "Spatial" indicates spatial domain disentanglement via two convolution blocks while keeping the loss function and the training strategy fixed.

Strategy	Various Scene			Diverse Illumination			Adverse Weather			Average		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Joint	71.0	45.4	42.4	**41.8**	**19.1**	**21.7**	47.1	17.1	21.9	53.3	27.2	28.7
Alternating	**75.1**	**49.7**	**45.3**	39.0	18.5	20.7	**48.0**	**17.2**	**22.3**	**54.0**	**28.4**	**29.4**

Table 5. Ablation analysis of the training strategy. "Joint" refers to the joint training strategy that utilizes $\mathcal{L}_{\text{det}}$ and $\mathcal{L}_{\text{con}}$ to jointly optimize the whole learnable parameter $\{\psi_{si}, \psi_{ss}, P, B_1, B_2, H\}$.

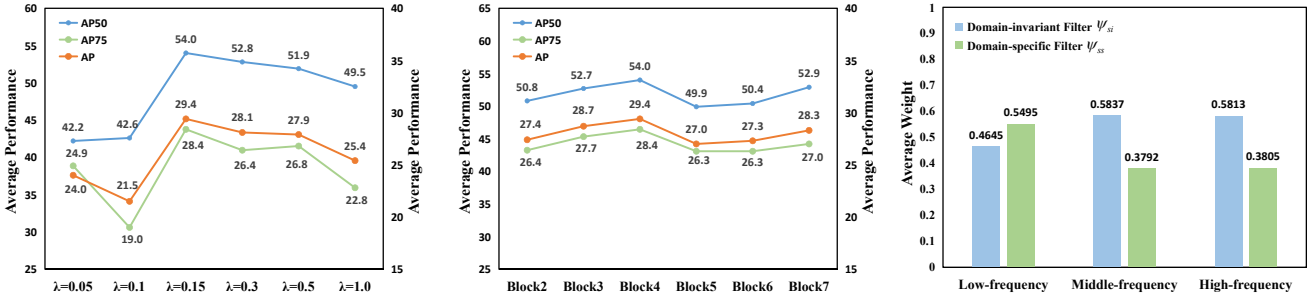


Figure 3. (a) Ablation analysis of the hyper-parameter λ of the proposed framework, which balances $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{det}}$. (b) Ablation analysis of the backbone division. The UAV-OD network's backbone is divided according to the specified block. (c) Statistic analysis of two learnable filters ψ_{si} and ψ_{ss}. The ψ_{si} and ψ_{ss} extract the domain-invariant and domain-specific spectrums from the image's spectrum. We can conclude that middle- and high-frequency components contain more domain-invariant information than the low-frequency component.

Tab. 3 further explains the broad applicability of our work, which illustrates the generalization results across datasets. Nine out of the twelve performance metrics demonstrate that our method achieves optimal performance. Our method exceeds the baseline method by 6.9%, 7.3%, and 5.8%, and the runner-up by 1.3%, 4.8%, and 2.6% on AP_{50}, AP_{75}, and AP on average generalization performance. It can be concluded that, compared to the baseline, almost all methods benefit the generalization to unseen target domains, and our method achieves decent performance, demonstrating its efficacy on generalization performance.

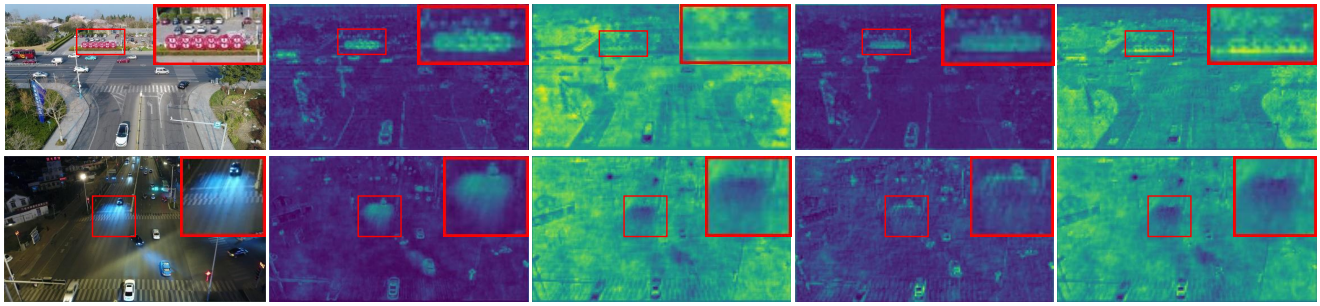
4.4. Ablation Analysis

Analysis of the frequency domain disentanglement. To verify the effectiveness of the frequency domain disentanglement, we substitute the two learnable filters in the frequency domain with two convolution blocks in the spatial domain, comprised of three convolution layers, while maintaining the same loss function and training strategy. The experimental results are shown in Tab. 4. We can observe that frequency domain disentanglement achieves better results.

Analysis of the training strategy. To verify the necessity of the alternating training strategy, we compare it with the joint training strategy in which $\mathcal{L}_{\text{det}}$ and $\mathcal{L}_{\text{con}}$ are used to optimize the entire learnable parameter $\{\psi_{si}, \psi_{ss}, P, B_1, B_2, H\}$. The experimental results are presented in the Tab. 5. We discover that the alternating training strategy produces superior results compared with the joint training strategy, demonstrating its effectiveness.

Analysis of hyper-parameter λ. We investigate how varying the setting of hyper-parameter λ affects the network's generalization performance. The average generalization performances on three unseen domains are depicted in Fig. 3 (a). We can observe that When λ is set to a lower value, the network's average generalization performance declines sharply. The average generalization performance rises as λ increases, with 0.15 being the optimal value.

Analysis of the backbone divisions. As mentioned be-



(a) Input image (b) Domain-invariant feature of [28] (c) Domain-specific feature of [28] (d) Domain-invariant feature of Ours (e) Domain-specific feature of Ours

Figure 4. Comparisons of domain-invariant and domain-specific features of Single-DGOD [30] and our method. The first and second rows indicate the target domain with various scene structures and diverse illumination conditions. For each feature map, the average of all channels is selected for visualization. Our method achieves a more thorough disentanglement.

(a) Input image (b) Domain-invariant component

Figure 5. Visualization analysis of the domain invariant components extracted from different domains. The first and second rows indicate the target domain with various scene structures and diverse illumination conditions.

fore, we divide the UAV-OD's backbone into B_1 and B_2. B_1 is utilized to extract the corresponding features from the domain-invariant component x_{si} and the domain-specific component x_{ss} for calculating the instance-level contrastive loss. To demonstrate the influence of different backbone divisions of the UAV-OD network on the generalization ability, we select different backbone divisions according to the structure of the backbone and conduct extensive experiments, as shown in Fig. 3 (b). It can be noticed that selecting block 4 as partition yields the best results for generalization.

4.5. Statistic Analysis on Learnable Filters.

We conduct statistic analysis on two learnable filters, ψ_{si} and ψ_{ss}, to explore which spectrum band contains more domain-invariant information. Specifically, the learnable filters ψ_{si} and ψ_{ss} are separated into three parts, respectively, according to the Eq. (3): low-frequency part, middle-frequency part, and high-frequency part. In addition, we calculate the weights of ψ_{si} and ψ_{ss} for different spectrum bands by averaging different parts, as shown in Fig. 3 (c). We can observe that the weights of ψ_{si} for middle-frequency and high-frequency are higher than those of ψ_{ss}. Therefore, we can conclude that the image's middle-frequency and high-frequency parts contain more domain-invariant information than the low-frequency part.

4.6. Visualization Analysis

Image-level visualization. As illustrated in Fig. 5, we first visualize the domain-invariant components. Although the image's appearance from different domains varies, the domain-invariant component from different domains looks similar, indicating that the learnable filter does extract the domain-invariant spectrum among different domains, which demonstrates the effectiveness of our method.

Feature-level visualization. In Fig. 4, we compare the domain-invariant and domain-specific features of Single-DGOD [30] and our method extracted from three unseen target domains. Our method achieves a more thorough disentanglement of irrelevant background features, e.g., the advertising board in 1st row and the area illuminated by the car lamp in 2nd row are excluded from the invariant features. Our method achieves better disentanglement results.

5. Conclusion

In this paper, we propose enhancing the UAV-OD network's generalization via frequency domain disentanglement. Firstly, we employ two learnable filters to extract the domain-invariant spectrum that contributes positively to generalization and domain-specific spectrum that contributes negatively to generalization. Then, we designed an instance-level contrastive loss to facilitate learning the two learnable filters. Experimental results and visualization analysis demonstrated the superiority of our method. For limitation, our approach is an initial exploration of learning generalized UAV-OD via frequency domain disentanglement. More subtle and effective designs of frequency domain disentanglement can be considered, leaving enough space for further development.

References

[1] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019. 3

[2] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 6

[3] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2020. 1, 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 5

[5] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 1, 3

[6] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018. 5, 6

[7] Milan Erdelj and Enrico Natalizio. Uav-assisted disaster management: Applications and open issues. In *2016 international conference on computing, networking and communications (ICNC)*, pages 1–5. IEEE, 2016. 1

[8] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao. Uncertainty-aware unsupervised domain adaptation in object detection. *IEEE Transactions on Multimedia*, 24:2502–2514, 2021. 1, 3

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 5

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[11] Eija Honkavaara, Heikki Saari, Jere Kaivosoja, Ilkka Pölönen, Teemu Hakala, Paula Litkey, Jussi Mäkynen, and Liisa Pesonen. Processing and assessment of spectrometric, stereoscopic imagery collected using a lightweight uav spectral camera for precision agriculture. *Remote Sensing*, 5(10):5006–5039, 2013. 1

[12] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021. 3

[13] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pages 124–140. Springer, 2020. 6

[14] G. Jocher, K. Nishimura, T. Mineeva, and R. Vilariño. Yolov5. https://github.com/ultralytics/yolov5, 2020. Accessed: 2020-07-10. 4

[15] Benjamin Kiefer, Martin Messmer, and Andreas Zell. Diminishing domain bias by leveraging domain labels in object detection on uavs. In *2021 20th International Conference on Advanced Robotics (ICAR)*, pages 523–530. IEEE, 2021. 3

[16] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6092–6101, 2019. 1

[17] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019. 1

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8771–8780, 2021. 1, 2, 3

[20] Hong Liu, Pinhao Song, and Runwei Ding. Towards domain generalization in underwater object detection. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1971–1975. IEEE, 2020. 1, 3

[21] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021. 3

[22] Sasanka Madawalagama, Niluka Munasinghe, SDPJ Dampegama, and L Samarakoon. Low cost aerial mapping with consumer-grade drones. In *37th Asian Conference on Remote Sensing*, pages 1–8, 2016. 1

[23] Payal Mittal, Raman Singh, and Akashdeep Sharma. Deep learning-based object detection in low-altitude uav datasets: A survey. *Image and Vision Computing*, 104:104046, 2020. 1

[24] Henri J Nussbaumer. The fast fourier transform. In *Fast Fourier Transform and Convolution Algorithms*, pages 80–111. Springer, 1981. 2

[25] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 6

[26] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 6956–6965, 2019. 1, 3

[27] Karthik Seemakurthy, Charles Fox, Erchan Aptoula, and Petra Bosilj. Domain generalisation for object detection. *arXiv preprint arXiv:2203.05294*, 2022. 1, 3

[28] Eduard Semsch, Michal Jakob, Dušan Pavlicek, and Michal Pechoucek. Autonomous uav surveillance in complex urban environments. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 82–85. IEEE, 2009. 1

[29] Jingye Wang, Ruoyi Du, Dongliang Chang, Kongming Liang, and Zhanyu Ma. Domain generalization via frequency-domain-based feature disentanglement and interaction. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4821–4829, 2022. 3

[30] Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 847–856, 2022. 1, 2, 3, 6, 8

[31] Xin Wu, Wei Li, Danfeng Hong, Ran Tao, and Qian Du. Deep learning for unmanned aerial vehicle-based object detection and tracking: a survey. *IEEE Geoscience and Remote Sensing Magazine*, 10(1):91–124, 2021. 1

[32] Zhenyu Wu, Karthik Suresh, Priya Narayanan, Hongyu Xu, Heesung Kwon, and Zhangyang Wang. Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1201–1210, 2019. 3

[33] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020. 3

[34] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12355–12364, 2020. 3

[35] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. 3

[36] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 3

[37] Xingxu Yao, Sicheng Zhao, Pengfei Xu, and Jufeng Yang. Multi-source domain adaptation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3273–3282, 2021. 1, 3

[38] Hongyang Yu, Guorong Li, Weigang Zhang, Qingming Huang, Dawei Du, Qi Tian, and Nicu Sebe. The unmanned aerial vehicle benchmark: Object detection, tracking and baseline. *International Journal of Computer Vision*, 128(5):1141–1159, 2020. 1

[39] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5372–5382, 2021. 6

[40] Xingxuan Zhang, Zekai Xu, Renzhe Xu, Jiashuo Liu, Peng Cui, Weitao Wan, Chong Sun, and Chen Li. Towards domain generalization in object detection. *arXiv preprint arXiv:2203.14387*, 2022. 1, 2, 3

[41] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *European Conference on Computer Vision*, pages 86–102. Springer, 2020. 3

[42] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye. Adaptive object detection with dual multi-label prediction. In *European Conference on Computer Vision*, pages 54–69. Springer, 2020. 3

[43] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[44] Pengfei Zhu, Dawei Du, Longyin Wen, Xiao Bian, Haibin Ling, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. Visdrone-vid2019: The vision meets drone object detection in video challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 5, 6

[45] Chenfan Zhuang, Xintong Han, Weilin Huang, and Matthew Scott. ifan: Image-instance full alignment networks for adaptive object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13122–13129, 2020. 3