International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02108-5

Towards Generalized UAV Object Detection: A Novel Perspective from Frequency Domain Disentanglement

Kunyu Wang[1] · Xueyang Fu[1] · Chengjie Ge[1] · Chengzhi Cao[1] · Zheng-Jun Zha[1]

Received: 12 October 2023 / Accepted: 26 April 2024

Abstract
When deploying unmanned aerial vehicle (UAV) object detection networks to complex, real-world scenes, generalization ability is often reduced due to domain shift. While most existing domain-generalized object detection methods disentangle domain-invariant features spatially, our exploratory experiments revealed a key insight for UAV object detection (UAV-OD): frequency domain contributions exhibit more pronounced disparities in generalization compared to generic object detection involving larger objects, since UAV-OD detects smaller objects. Therefore, frequency domain disentanglement stands out as a more direct, effective approach for UAV-OD. This paper proposes a novel frequency domain disentanglement method to improve UAV-OD generalization. Specifically, our framework leverages two learnable filters extracting domain-invariant and domain-specific spectra. Additionally, we design two contrastive losses: an image-level loss and an instance-level loss guiding training. These losses enable the filters to focus on extracting domain-invariant and domain-specific spectra, achieving better disentangling. Extensive experiments across multiple datasets, including UAVDT and Visdrone2019-DET, utilizing Faster R-CNN and YOLOv5, show our approach consistently and significantly outperforms baseline and state-of-the-art domain generalization methods. Our code is available at https://github.com/wangkunyu241/UAV-Frequency.

Keywords Unmanned aerial vehicles · Object detection · Domain generalization · Frequency domain disentanglement · Contrastive learning.

1 Introduction

Emerging as a crucial visual capability for unmanned aerial vehicles (UAV), UAV Object Detection (UAV-OD) (Cao et al., 2020; Wu et al., 2019; Chen et al., 2019; Zhang et al., 2019; Kiefer et al., 2022; Mittal et al., 2020; Du et al., 2018) continues to command attention from academic and industry researchers. Its significance is underscored by an array of intelligent applications, including civil infrastructure inspection, precision agriculture, and search and rescue operations (Duarte et al., 2022; Lygouras et al., 2019; Geraldes et al., 2019; San et al., 2018). However, due to the large mobility of UAVs, UAV-OD networks operate in a plethora of outdoor environments, often with diverse weather and illumination conditions. This creates a domain shift where the training dataset of UAV-OD networks (source domain) largely diverges from complex and unseen real-world scenes (target domain), leading to suboptimal performance. Consequently, enhancing the generalization ability of UAV-OD becomes a priority.

Domain Adaptation (DA) (Cao et al., 2023; Liu et al., 2023; Ganin & Lempitsky, 2015; Tzeng et al., 2017; Jiang et al., 2021a) proposes a compelling solution to address the problem by seeking aligned features between source and target domains. However, it has limitations when lacking target data availability, a common scenario within the complex and diverse deployment environments of UAV-OD. This hampers the practicality of DA approaches. While Domain Generalization (DG) (Liu et al., 2020a; Wu & Deng, 2022; Zhang et al., 2022; Xu et al., 2023; Lin et al., 2021; Vidit et al., 2023; Zhong et al., 2022; Zhao et al., 2023) aims to overcome this limitation by learning from a single or multiple related yet distinct source domains to ensure model generalization under distribution shifts. Most existing DG methodologies

Communicated by Hong Liu.

✉ Zheng-Jun Zha
zhazj@ustc.edu.cn

[1] School of Information Science and Technology and MoE Key Laboratory of Brain-Inspired Intelligent Perception and Cognition, University of Science and Technology of China, Hefei 230026, China

Published online: 14 June 2024

focus on object-related features in the spatial domain. However, we propose frequency domain disentanglement as an effective method for improving UAV-OD generalization. Our experimental validation, displayed in Sect. 3, reveals that the contributions of different frequency bands in generalization exhibit more pronounced disparities for UAV-OD, characterized by smaller-sized objects, compared to general object detection which involves larger objects.

Drawing on these findings, we suggest enhancing UAV-OD generalization via frequency domain disentanglement, a more direct and efficient method. Specifically, we firstly propose utilizing two learnable filters to extract domain-invariant and domain-specific spectrums. Then, we introduce two novel contrastive losses at image and instance level to facilitate the filter training for disentangling distinctive spectrums. This encourages the two captured spectrums to contain domain-invariant characteristics shared by target objects and domain-specific characteristics that vary across different domains. In this way, the UAV-OD network can generalize well on unseen target domains. For experiment settings, we focus on single-domain generalization, a notably challenging endeavour. We perform extensive experiments to validate our framework effectiveness, using multiple datasets UAVDT (Du et al., 2018) and Visdrone2019-DET (Zhu et al., 2019), with various object detection models like Faster-RCNN (Girshick, 2015) and YOLOv5 (Jocher et al., 2020). The results clearly reveal our proposed method enables superior generalization in the UAV-OD network compared to baseline and multiple state-of-the-art DG methods in unseen target domains.

To sum up, the contributions of this work are summarized as follows:

- Through exploratory experiments, we have gained a vital insight in the field of UAV-OD: the contributions of different frequency in generalization exhibit more pronounced disparities within UAV-OD, which deals with smaller-sized objects, compared to general object detection, which involves larger objects.
- Based on these findings, we make the earliest effort to improve the UAV-OD generalization via frequency domain disentanglement, which serves as a more direct and efficient method, providing a novel perspective to the field.
- We propose a novel frequency domain disentanglement framework that utilizes two learnable filters to extract the domain-invariant and domain-specific spectrums and design two novel contrastive losses at image and instance level to guide the disentangling process.

This paper expands on our previous conference version (Wang et al., 2023a) in three key aspects. Firstly, we clarify the motivation behind our approach and provide exploratory experiments, justifying frequency domain disentanglement's suitability for enhancing UAV-OD generalization. Secondly, we introduce a more efficient frequency domain disentanglement structure and new image-level contrastive loss on the methodological level. Through ablation experiments, we demonstrate the effectiveness of these enhancements, further improving UAV-OD networks generalization. Lastly, we conduct a comprehensive suite of experiments, providing extensive ablation studies, visualizations, and a thorough analysis based on the empirical findings.

2 Related Work

Since this paper discusses domain adaptive object detection, domain generalized object detection, frequency-based domain generalization, and UAV object detection, we provide a brief overview of these areas.

2.1 Domain Adaptive Object Detection

Domain adaptation strategies (Lu et al., 2023; Hsu et al., 2020a, b; Li et al., 2016) prioritize the transference of knowledge from the source domain to the target domain. Initially designed for image categorization, these methods were subsequently extended to object detection applications. For instance, Chen et al. (2018) and Saito et al. (2019) constructed methods to address domain shift scenarios by employing domain classifiers, consistency regularization, adversarial loss, and alignment of source and target image distributions. Zheng et al. (2020) introduced an ingenious coarse-to-fine feature adaptation approach for cross-domain object detection, which progressively and accurately aligns deep features. Zhuang et al. (2020) aligned feature distributions on the picture and instance levels to enhance generalization performance. A unique method based on vector decomposition was proposed by Wu et al. (2021a), to extract domain-neutral object representations for domain adaptive object detection. Li et al. (2022c) incorporated feature-level adversarial training, weak-strong augmentation, as well as mutual learning between teacher and student models to ensure domain-invariant features and minimize the production of low-quality pseudo labels. Chen et al. (2022b) employed uncertainty-guided consistency training and a novel Entropy Focal Loss to improve classification adaptation and localization adaptation. Chen et al. (2022a) introduced graph structures into the detection pipeline to deliberately model the intra- and inter-domain foreground object relations in both pixel and semantic spaces, thus extending the domain adaptive object detection model capability for relational reasoning. Li et al. (2022a) modeled impartial semantics with category knowledge, which brings semantic knowledge into global alignment and achieves semantic-conditioned adapta-

tion. Li et al. (2022b) utilized graph nodes and edges to set up semantic-aware node affinity and accomplishes refined adaptation through node-to-node graph matching. Liu et al. (2023) carried out cross-modality graph reasoning between linguistic and visual graphs to learn a generalized detector. Cao et al. (2023) integrated mean-teacher self-training with contrastive learning to address the domain gap in object detection.

However, the applicability of these techniques can be compromised by the requirement for specific target domain data, restricting their overall usability. The objective of this paper is to address this constraint by concentrating on domain generalization.

2.2 Domain Generalized Object Detection

Domain-generalized object detection has recently surged in popularity due to its performance superiority over domain-adaptive object detection and its ability to operate independently of target domain data. Liu et al. (2020a) proposed a data augmentation method that enriches the domain diversity of the initial small dataset, thereby heightening its generalization performance. Lin et al. (2021) integrated a disentangled network into Faster R-CNN. They acquired domain-invariant representation at both image and instance levels for a more generalizable object detection. Zhang et al. (2022) suggested a comprehensive evaluation benchmark along with a novel method termed 'region aware proposal reweighting.' This approach aids in eliminating dependence within the features of the Region of Interest and thus enhances the generalization capability of detectors amid varying distribution shifts. Focusing on single-domain generalization, Wu and Deng (2022) in his paper on Single-DGOD, developed a method known as cyclic-disentangled self-distillation. This process extracts domain-invariant representations without the necessity for domain-related annotations. Vidit et al. (2023) utilized a self-supervised vision-language model, CLIP, to instruct the training of an object detector for generalization to unseen target domains. He also proposed a semantic augmentation strategy that involves textual prompts from the pre-trained vision-language model, introducing semantic domain concepts. Xu et al. (2023) introduced a Multi-view Adversarial Discriminator-based domain generalization model. This model eradicates non-causal factors from standard features through multi-view adversarial training on source domains.

The above-mentioned domain-generalization methods primarily apply to generic object detection scenarios (Sun et al., 2021a; Zhou et al., 2023; Kajiura et al., 2021). However, taking into account the inherent small-object characteristics in UAV-OD, this paper proposes a domain-generalization method explicitly designed for UAV-OD scenarios. Through frequency domain disentanglement, we effectively disentangle domain-invariant features for UAV-OD in a more direct manner.

2.3 Frequency-Based Domain Generalization

Frequency-based domain generalization, owing to its distinctive features, has seen a surge in interest and research efforts recently. It broadly falls into two categories. One of them delves into the integration of the frequency domain prior into domain generalization. Xu et al. (2021), for instance, brought forth a Fourier-based strategy for augmentation, named amplitude mix. This strategy broadens the augmented images using a combination of Taylor expansion and inverse Fourier transformation. The model subsequently utilizes these images, along with their original labels, for enhancing generalization. Yang and Soatto (2020), on the other hand, proposed a method for unmonitored domain adaptability, leveraging Fourier domain adaptation. The plan here is to exchange the low-frequency spectrum of the source and target distributions, reducing the discrepancy between them. To face the challenges of federated domain generalization in medical imagery, Liu et al. (2021) proposed a new technique: Episodic Learning in Continuous Frequency Space. This method employs a continuous frequency space interpolation function to relay the distribution data across clients while preserving privacy. Meanwhile, Jeon et al. (2021) suggested an innovative feature stylization method to create novel domains while safeguarding discriminative class information. This process entails breaking down features into low and high frequency components, stylizing the low-frequency ones with novel domain styles, and preserving shape cues in the high-frequency components. Across a similar vein, Lee et al. (2023) addressed the issue of content variation in normalization for domain generalization. This approach considers amplitude and phase as style and content, respectively. Their method, PCNorm, removes style while retaining content through spectral decomposition.

The second category emphasizes learning generalization within the frequency domain. Huang et al. (2021) proposed image randomization in frequency space. This proposal involves maintaining unchanged domain invariant frequency components while randomizing only domain variant frequency components. The ultimate goal is to facilitate more controllable randomization and minimal effects on semantic structures of images and domain invariant features. Lin et al. (2023) introduced a method named Deep Frequency Filtering to explicitly modulate the frequency components of differing transfer difficulties during training across domains within the latent space. The primary drive behind this initiative is to boost the generalization competence of Deep Neural Networks. Lastly, Yang et al. (2023) proffered a novel learning strategy for multiple frequency domains. This strategy fragments the frequency domain of each original image into

subdomains, which forces the model to glean features from additional samples within the specifically limited spectrum. This augmented learning increases the chances of acquiring domain invariant features.

Our approach aligns with the second category. Noting the significant disparities various frequency bands contribute to the generalization of UAV-OD, we suggest decoupling domain-invariant features within the frequency domain to improve the generalization.

2.4 UAV Object Detection

Unmanned aerial vehicles object detection (UAV-OD) (Sun et al., 2021b; Mittal et al., 2020; Wu et al., 2021b) aims to detect objects of interest within UAV-captured images. However, in comparison to generic object detection (using surveillance or ground-based cameras), UAV equipped with cameras face greater challenges due to their high mobility and flexibility. These challenges include, but are not limited to:

Variations in object size. UAV operate across a range of altitudes when capturing images. Shooting at lower altitudes allows their cameras to capture finer details of objects. As the UAV ascends to higher altitudes, the camera survey a larger area, resulting in the capture of more objects. Therefore, UAV-OD should exhibit the capability to detect objects at various scales, including larger, intricately detailed objects, as well as smaller, less distinct ones. RR-Net (Chen et al., 2019) addressed the challenge of detecting objects of various scales in UAV-captured images by reducing bounding box prediction to key point and size estimation. Liu et al. (2020b) addressed the challenge of detecting small objects by increasing convolution operation at an early layer to enrich spatial information.

Variations in view angle. The mobility of UAVs enables them to capture images from diverse view angles. For instance, a UAV can observe an object from the front, side, and bird's-eye views within a very short period. This diversity of view angles leads to arbitrary orientations and aspect ratios of objects. Some view angles, such as the bird's-eye view, are rarely encountered in traditional ground-based object detection. Consequently, UAV object detection models must address the varied visual appearances of the same object from multiple view angles. Wu et al. (2019) introduced the Nuisance Disentangled Feature Transform framework, which leverages additional metadata alongside UAV images to acquire domain-robust features. This approach tackles the challenges posed by UAV-specific nuisances, including fluctuating flying altitudes, and dynamically shifting viewing angles.

Real-time. UAV require sustained high real-time performance to prevent economic losses and potential threats to human life, such as in military, search and rescue, and surveillance areas. However, ensuring optimal performance and real-time capabilities for object detection on a UAV platform, constrained by limited computational and storage resources, presents a highly challenging endeavor. Zhang et al. (2019) addressed the challenge by enforcing channel-level sparsity of convolutional layers and pruning the less informative feature channels, thereby obtaining "slim" object detectors.

Class imbalance. Most existing UAV-OD datasets encounter class imbalance issue, where a small subset of frequently occurring classes typically dominates the majority of object instances. This leads to a long-tailed distribution, significantly impacting the detection performance of long tail classes. Yu et al. (2021) proposed a Dual Sampler and Head detection Network to solve the long-tail distribution problem in UAV datasets, which consists of two integral components: Class-Biased Samplers and Bilateral Box Heads.

Degradation. UAV operate in uncontrolled outdoor environments, facing unpredictable weather and lighting conditions. These conditions result in varying image degradation, significantly affecting object detection performance. Consequently, enhancing the generalization ability of UAV-OD is imperative.

This paper tackles the fifth challenge. Through exploratory experiments, we observe that compared to generic object detection, different frequency bands exhibit more pronounced disparities in contributions to the generalization of UAV-OD. Based on this insight, we introduce frequency domain disentanglement as a more direct and efficient decoupling method, greatly enhancing the generalization ability of UAV-OD.

3 Motivation

We first conduct experiments to explore how object size influences the contributions of various frequency bands to the network's generalization ability. Specifically, we use the UAVDT dataset (Du et al., 2018) as the foundational dataset for UAV object detection (UAV-OD). Within UAVDT, we employ the daylight scenes as the source training domain and the nighttime and foggy portions as the unseen target test domains to evaluate the model's generalization performance. In parallel, we examine generic object detection using the BDD100k (Yu et al., 2020), Cityscape (Cordts et al., 2016), and Foggy Cityscape (Sakaridis et al., 2018) datasets. To assess the model's generalization ability, we adopt the daylight scenes within BDD100k and Cityscapes clean images as the source training domain and the nighttime portion within BDD100k along with the Foggy Cityscapes foggy images as the unseen target test domains. To provide a comprehensive understanding, we meticulously collect statistics regarding the absolute sizes of objects in both the UAV-OD datasets and the generic object detection datasets, as shown in Fig. 1.

Table 1 The contributions of different frequency bands to the UAV-OD network's generalization

Frequency band	$[\alpha, \beta]$	Nighttime			Foggy			Average		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Full	[0, 1]	49.4	22.3	26.3	29.6	12.5	15.7	39.5	17.4	21.0
Low	[0, 0.02]	1.4	0.5	0.7	9.1	0.3	2.8	5.3	0.4	1.7
Middle	[0.02, 0.1]	25.1	10.4	12.3	22.7	9.1	11.2	23.9	9.8	11.8
High	[0.1, 1]	63.3	31.6	33.8	38.2	13.7	18.5	50.7	22.6	26.2

We utilize daytime scenes from UAVDT as the source domain and the specified bands of source images are reserved for training according to the $[\alpha, \beta]$. For testing, we adopt the nighttime and foggy portions within UAVDT as unseen target domains to evaluate the generalization performance. "Average" refers to the average generalization performance across unseen target domains

Table 2 The contributions of different frequency bands to the generic object detection network's generalization

Frequency band	$[\alpha, \beta]$	Nighttime			Foggy			Average		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Full	[0, 1]	63.2	29.9	31.9	48.9	31.7	29.6	56.1	30.8	30.8
Low	[0, 0.02]	12.4	9.5	9.0	8.6	7.1	7.3	10.5	8.3	8.2
Middle	[0.02, 0.1]	34.7	13.0	17.4	35.4	20.0	22.5	35.1	16.5	19.9
High	[0.1, 1]	55.3	23.1	28.5	45.7	29.4	28.2	50.5	26.2	28.4

We utilize daytime scenes from BDD100k and Cityscapes clean images as the source domain and the specified bands of source images are reserved for training according to the $[\alpha, \beta]$. For testing, we adopt the nighttime portion within BDD100k, along with the foggy images collected from Foggy Cityscapes as unseen target domains to evaluate the generalization performance

According to the definition in Cheng et al. (2023), the average absolute size of objects in the UAV-OD datasets falls within the small object range, whereas in the general object detection datasets, the average absolute size of objects exceeds the normal object range.

During training, we first convert each source domain image $x \in \mathbb{R}^{H\times W\times C}$ into frequency space through Fast Fourier Transform (FFT) (Nussbaumer & Nussbaumer, 1982):

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

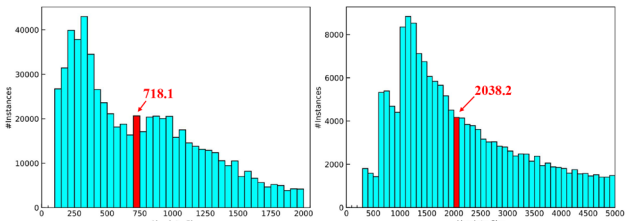
The frequency space signal $\mathcal{F}(x)$ can be further decomposed to an amplitude spectrum $\mathcal{A}(x)$ and a phase spectrum $\mathcal{P}(x)$, which is expressed as:

$$\mathcal{A}(x)(u,v) = \left[\mathcal{R}^2(x)(u,v) + \mathcal{I}^2(x)(u,v)\right]^{1/2}, \tag{2}$$

$$\mathcal{P}(x)(u,v) = \arctan\left[\frac{\mathcal{I}(x)(u,v)}{\mathcal{R}(x)(u,v)}\right], \tag{3}$$

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 $\mathcal{A}(x)$ outside the range of a certain upper threshold α and lower threshold β with a band reject filter



(a) UAV-OD (b) Generic object detection

Fig. 1 Statistical analysis of the absolute size of objects in UAV-OD and Generic Object Detection Datasets. The highlighted numbers and regions represent the average absolute size and their respective intervals. For visual clarity, intervals that account for less than 0.1% of the total object count have not been included

$f_s \in \mathbb{R}^{H\times W\times C}$ and obtain the remaining amplitude spectrum $\hat{\mathcal{A}}(x)$:

$$f_s(i,j) = \begin{cases} 1, & i \in [\frac{(1-\beta H)}{2}, \frac{(1-\alpha)H}{2}] \cup [\frac{(1+\alpha)H}{2}, \frac{(1+\beta)H}{2}] \\ & j \in [\frac{(1-\beta)W}{2}, \frac{(1-\alpha)W}{2}] \cup [\frac{(1+\alpha)W}{2}, \frac{(1+\beta)W}{2}] \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$\mathcal{A}(x) = \hat{\mathcal{A}}(x) \otimes f_s, \tag{5}$$

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 image $\hat{x}$ correspond to different frequency

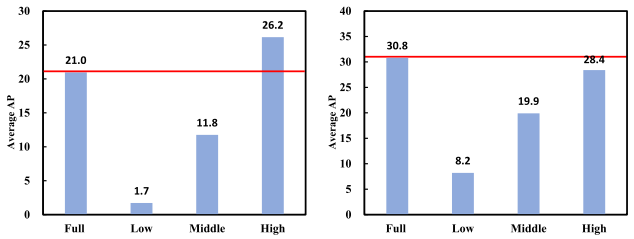


Fig. 2 Histograms of the average AP for Tables 1 and 2. We can clearly observe that, in contrast to generic object detection, different frequency bands make significantly distinct contributions to the generalization of UAV-OD

bands, which is utilized to train the network. The quantitative generalization results of different frequency bands for UAV-OD and generic object detection are shown in Tables 1 and 2. Two visualized bar charts are shown in Fig. 2.

We can observe that for generic object detection with larger object sizes, the most significant contribution to the network's generalization ability comes from high-frequency components, followed by mid-frequency components, and finally, low-frequency components. However, it is noteworthy that the generalization performance of any individual frequency band does not surpass that achieved when using all frequency bands together. Conversely, for UAV-OD with smaller object sizes, the ranking of contributions across frequency bands remains consistent, with high frequencies contributing the most to generalization, followed by mid and low frequencies. However, the distinctions in their contributions become even more pronounced for UAV-OD, with high-frequency components demonstrating a generalization performance that can surpass that achieved when using all frequency bands together. Based on these insightful observations, we can conclude that unlike generic object detection, the different frequency bands make notably distinct contributions to the generalization of UAV-OD. Specifically, high-frequency components play a significantly more prominent role in the generalization of UAV-OD compared to mid- and low-frequency components. This motivates us to use frequency domain information to improve the generalization performance.

4 Methodology

In this section, we begin by presenting the necessary preliminary knowledge in Sect. 4.1. In Sect. 4.2, we introduce the problem formulation, covering both problem definition and an overview of the methodology. A comprehensive description of the framework's architecture, including its components and the formulation of the loss functions, will be provided in Sect. 4.3. Moving on to Sect. 4.4, we delve into the training strategy and the inference process. Furthermore, we discuss the efficiency of our proposed method.

4.1 Preliminaries

Our proposed framework adopts Faster R-CNN (Girshick, 2015) and YOLOv5 (Jocher et al., 2020) as backbone detection models to substantiate the efficacy of our approach. Hence, we briefly review the these two backbone detection models.

Faster R-CNN, a two-stage detector, comprises of four components: backbone network, region proposal network, RoI pooling, and detection head. Its training loss is formulated as:

$$\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{reg}. \tag{6}$$

The $\mathcal{L}_{cls}$ term calculates the classification loss for each RoI, employing softmax to predict the likelihood of an RoI belonging to any object class. It quantifies the disparity between predicted and ground-truth class labels through cross-entropy. The $\mathcal{L}_{reg}$ term evaluates the difference between the predicted and actual box positions using a smooth $\mathcal{L}_1$ loss.

YOLOv5, a one-stage detector, comprises of four components: backbone network, neck network, anchor boxes, and detection head. Its training loss is:

$$\mathcal{L}_{det} = \mathcal{L}_{obj} + \mathcal{L}_{cls} + \mathcal{L}_{reg}. \tag{7}$$

The $\mathcal{L}_{obj}$ term measures the model's confidence in detecting objects within a bounding box. It uses binary cross-entropy loss to penalize confidence score errors. The $\mathcal{L}_{cls}$ term computes the error between predicted class probabilities and ground-truth class labels for each bounding box. It utilizes categorical cross-entropy loss. The $\mathcal{L}_{reg}$ term is applied to fine-tune the predicted bounding box coordinates to match the ground-truth.

4.2 Problem Formulation

Our aim is to learn a UAV-OD network trained on a source domain D_s that can generalize well on multiple unseen target domains D_t. Let X_s and $X_t \subset \mathbb{R}^{H\times W\times C}$ denote source domain D_s and target domain D_t images with height H, width W, and number of channel C, Y_s and $Y_t \subset \mathbb{R}$ denote the category labels of X_s and X_t, B_s and $B_t \subset \mathbb{R}^4$ denote the bounding boxes of X_s and X_t. The source domain can be formulated as $D_s = \left\{x_s^i, \left\{y_s^{ij}, b_s^{ij}\right\}_{j=1}^{N_i}\right\}_{i=1}^{N}$, 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.

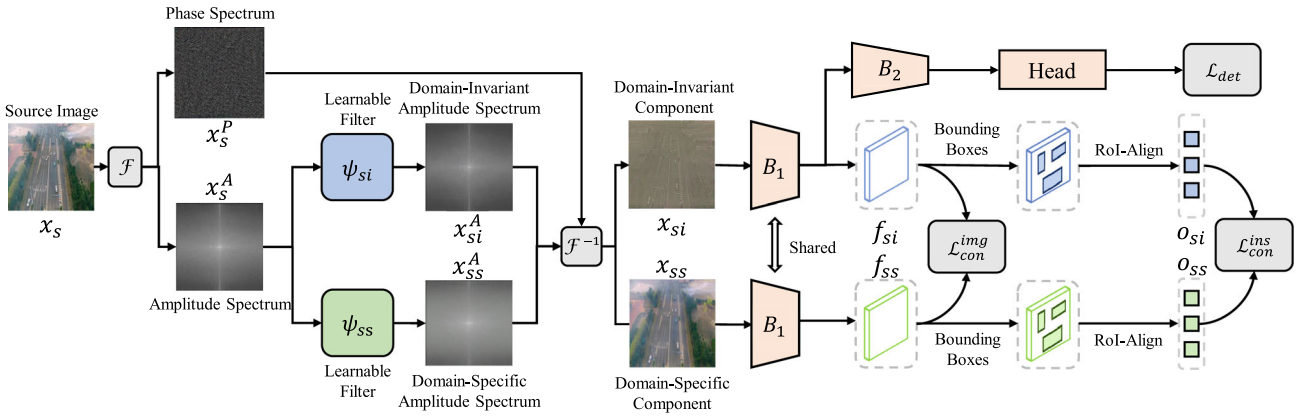


Fig. 3 Overview of the proposed framework. $\mathcal{F}$ and $\mathcal{F}^{-1}$ indicate FFT and IFFT. The backbone network is divided into B_1 and B_2. Head represents the detection head of UAV-OD network. ROI-Align indicates ROI-Alignment operation (He et al., 2017). We employ two learnable filters 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 two image-level and 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

Each target domain D_t, for simplicity, can be formulated as $D_t = \left\{x_t^i, \left\{y_t^{ij}, b_t^{ij}\right\}_{j=1}^{O_i}\right\}_{i=1}^{O}$, which includes O images and each image has O_i pairs of category labels $y_t^{ij} \in Y_t$ and bounding boxes $b_t^{ij} \in B_t$. During training, we can only access the source domain data. The trained network is tested on multiple unseen target domains.

To solve the problem, we propose a unique method to improve the UAV-OD's generalization ability, using frequency domain disentanglement. This scheme aims to segregate the domain-invariant feature within a frequency domain. To realize this goal, our approach uses two learnable filters tasked with extracting both the domain-invariant and domain-specific spectral components from the input image. The domain-invariant section of the input image then gets used for the object detection process. Moreover, we introduce a pair of new contrastive loss functions at the image-level and instance-level. These functions guide the training of the two learnable filters, focusing them on separating the two differing spectrums. With regard to the training strategy, we partition the entire framework's learnable parameters into two sets and utilize alternative optimization to efficiently handle the frequency domain disentanglement while simultaneously minimizing potential conflicts in UAV object detection. An overview of our framework is provided in Fig. 3.

4.3 Framework Architecture

Given a source 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):

$$x_s^{\mathcal{F}} = \mathcal{F}(x_s). \tag{8}$$

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:

$$x_s^{\mathcal{A}} = \mathcal{A}(x_s^{\mathcal{F}}), \quad x_s^{\mathcal{P}} = \mathcal{P}(x_s^{\mathcal{F}}). \tag{9}$$

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 $x_s^{\mathcal{A}}$. Unlike the conference version, here we explore several potential filter structures, including the learnable tensors of shape $H \times W$ that can be element-wise multiplied with the amplitude spectrum, the learnable tensors of size $H \times W \times C$ that can be multiplied with the amplitude spectrum, and the conv blocks performing convolution on the amplitude spectrum, as illustrated in Fig. 4.

From the experimental observations, the superior generalization performance was associated with the second structure. Comprehensive experimental results are elaborated in Sect. 5.3, specifically within the "Frequency disentanglement structure" subsection. Consequently, we have chosen the second structure to design the learnable filters ψ_{si} and ψ_{ss}. The procedure for decoupling the domain-invariant amplitude spectrum $x_{si}^{\mathcal{A}}$ and the domain-specific amplitude

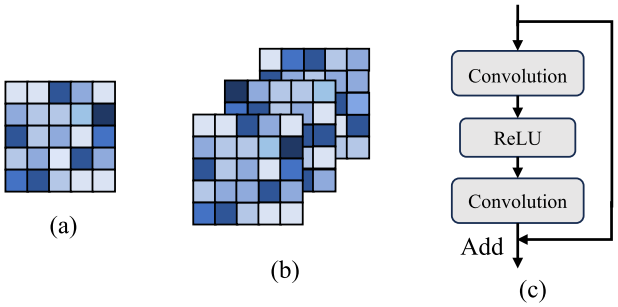


Fig. 4 Three potential frequency filter structure. **a** The learnable tensors of shape H × W that can be element-wise multiplied with the amplitude spectrum, **b** the learnable tensors of size H × W × C that can be multiplied with the amplitude spectrum, **c** the conv blocks performing 1x1 conv on the amplitude spectrum

spectrum $x_{ss}^{\mathcal{A}}$ is:

$$x_{si}^{\mathcal{A}} = x_s^{\mathcal{A}} \otimes \psi_{si}, \quad x_{ss}^{\mathcal{A}} = x_s^{\mathcal{A}} \otimes \psi_{ss}, \tag{10}$$

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}.

Furthermore, we design two novel image-level and instance-level contrastive loss to facilitate the learnable filters to focus on extracting domain-invariant and domain-specific spectrums. Specifically, we first divide the backbone network of detection models 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:

$$f_{si} = B_1(x_{si}), \quad f_{ss} = B_1(x_{ss}) \tag{11}$$

Let $\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}/\|\boldsymbol{u}\|\|\boldsymbol{v}\|$ denote the dot product between $\mathcal{L}_2$ normalized u and v, τ denotes the temperature hyper-parameter (He et al., 2020). Based on the f_{si} and f_{ss}, the image-level contrastive loss $\mathcal{L}_{\text{con}}^{\text{img}}$ is calculated as follows:

$$\begin{aligned}\mathcal{L}_{\text{con}}^{\text{img}} = &\sum_{f_{ss}\in F_{ss}} \frac{-1}{\text{Bs}} \sum_{\substack{\hat{f}_{ss}\in F_{ss}\\ f_{ss}\neq\hat{f}_{ss}}} \log \frac{\exp\left(\text{sim}\left(f_{ss}, \hat{f}_{ss}\right)/\tau\right)}{\sum_{\substack{f_s\in F_s\\ f_{ss}\neq f_s}} \exp\left(\text{sim}\left(f_{ss}, f_s\right)\right)} \\ &+ \sum_{f_{si}\in F_{si}} \frac{-1}{\text{Bs}} \sum_{\substack{\hat{f}_{si}\in F_{si}\\ f_{si}\neq\hat{f}_{si}}} \log \frac{\exp\left(\text{sim}\left(f_{si}, \hat{f}_{si}\right)/\tau\right)}{\sum_{\substack{f_s\in F_s\\ f_{si}\neq f_s}} \exp\left(\text{sim}\left(f_{si}, f_s\right)\right)},\end{aligned} \tag{12}$$

where Bs is the training batch size, $F_{si} = \{f_{si}^j\}_{j=1}^{\text{Bs}}$ denotes the set of f_{si} for each sample in a batch of source training data, $F_{ss} = \{f_{ss}^j\}_{j=1}^{\text{Bs}}$ denotes the set of f_{ss} for each sample in a batch of source training data, $F_s = F_{si} \cup F_{ss}$ denotes the intersection of F_{si} and F_{ss}.

After that, 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 $\{o_{si}^1, \cdots, o_{si}^n\}$ from x_{si} and the domain-specific instance-level features $\{o_{ss}^1, \cdots, o_{ss}^n\}$ 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 (He et al., 2017) to align the spatial size of all instance-level features:

$$o_{si} = \{\hat{o}_{si}^1, \cdots, \hat{o}_{si}^n\} = \text{RoIAlign}(\{o_{si}^1, \cdots, o_{si}^n\}), \tag{13}$$

$$o_{ss} = \{\hat{o}_{ss}^1, \cdots, \hat{o}_{ss}^n\} = \text{RoIAlign}(\{o_{ss}^1, \cdots, o_{ss}^n\}), \tag{14}$$

where $\hat{o}_{si}^j, \hat{o}_{ss}^j \in \mathbb{R}^{s\times s\times c}$, $j \in \{1, \cdots, n\}$, s indicates the output size of RoI-Alignment. Based on the o_{si} and o_{ss}, the instance-level contrastive loss $\mathcal{L}_{\text{con}}^{\text{ins}}$ is fomulated as:

$$\begin{aligned}\mathcal{L}_{\text{con}}^{\text{ins}} = &\sum_{\hat{o}_{ss}\in O_{ss}} \frac{-1}{\text{Bs}\times\text{n}} \sum_{\substack{\overline{o}_{ss}\in O_{ss}\\ \hat{o}_{ss}\neq\overline{o}_{ss}\\ \hat{y}_{ss}\neq\overline{y}_{ss}}} \log \frac{\exp\left(\text{sim}\left(\hat{o}_{ss}, \overline{o}_{ss}\right)/\tau\right)}{\sum_{\substack{o_s\in O_s\\ \hat{o}_{ss}\neq o_s}} \exp\left(\text{sim}\left(\hat{o}_{ss}, o_s\right)\right)} \\ &+ \sum_{\hat{o}_{si}\in O_{si}} \frac{-1}{\text{Bs}\times\text{n}} \sum_{\substack{\overline{o}_{si}\in O_{si}\\ \hat{o}_{si}\neq\overline{o}_{si}\\ \hat{y}_{si}\neq\overline{y}_{si}}} \log \frac{\exp\left(\text{sim}\left(\hat{o}_{si}, \overline{o}_{si}\right)/\tau\right)}{\sum_{\substack{o_s\in O_s\\ \hat{o}_{si}\neq o_s}} \exp\left(\text{sim}\left(\hat{o}_{si}, o_s\right)\right)},\end{aligned} \tag{15}$$

where $O_{si} = o_{si}^1 \cup, \cdots, \cup o_{si}^{\text{Bs}}$ denotes the intersection of o_{si} for each sample in a batch a source training data, $O_{ss} = o_{ss}^1 \cup, \cdots, \cup o_{ss}^{\text{Bs}}$ denotes the intersection of o_{ss} for each sample in a batch a source training data, $O_s = O_{si} \cup O_{ss}$ denotes the intersection of O_{si} and O_{ss}. Note that "Bs × n" assumes that each sample within a batch of training data contains n objects. If this assumption is not met, it should be the sum of objects in each sample.

In the context of image-level contrastive loss, domain-invariant image-level features within a training data batch are considered positive samples for each other, as are domain-specific image-level features. Conversely, domain-invariant image-level features are treated as negative samples compared to domain-specific image-level features. For the instance-level contrastive loss, domain-invariant instance-level features of the same class within a training data batch form positive pairs, as do domain-specific instance-level features of the same class. Conversely, instance-level features from different classes or from domain-invariant and

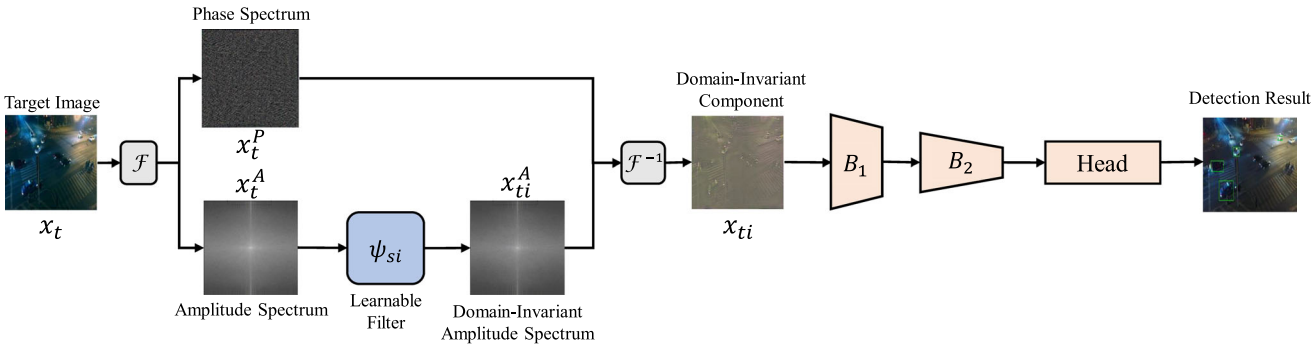


Fig. 5 Inference stage of the proposed framework. The domain-invariant learnable filter is utilized for prediction, while the domain-specific learnable filter is not used in the inference process

domain-specific categories respectively are regarded as negative pairs. By optimizing $\mathcal{L}_{\text{con}}^{\text{img}}$ and $\mathcal{L}_{\text{con}}^{\text{ins}}$, we bring positive pairs closer together while pushing negative pairs apart. This assists the two learnable filters in extracting domain-invariant and domain-specific features separately from both a global, image-level perspective and a local, instance-level perspective. As a result, frequency domain disentanglement is achieved.

4.4 Training and Inference

For the training stage, let us first denote the total loss function of our framework as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda \times (\mathcal{L}_{\text{con}}^{\text{img}} + \mathcal{L}_{\text{con}}^{\text{ins}}), \tag{16}$$

where $\mathcal{L}_{\text{det}}$ denotes the loss function of detection models, as discussed in Sect. 4.1, λ is the hyper-parameter for balancing detection loss and contrastive loss. 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 and detection head H. 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:

$$\theta = \{\psi_{si}, \psi_{ss}\}, \quad \eta = \{B_1, B_2, H\}. \tag{17}$$

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

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

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

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

where 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.

For the inference stage, as shown in Fig. 5, the domain-invariant learnable filter ψ_{si} is utilized to extract the domain-invariant amplitude spectrum, subsequently used to construct the domain-invariant component. We employ the domain-invariant component directly for prediction, while the domain-specific component is not used in the reference process.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate the proposed method using popular UAV-OD benchmarks: UAVDT (Du et al., 2018) and Visdrone2019-DET (Zhu et al., 2019). UAVDT consists of 41k frames with 840k bounding boxes, divided into three classifications; cars, trucks, and buses. Given that the distribution in UAVDT is disproportionately skewed with the last two classes constituting less than 5% of bounding boxes, we consolidate them into a unified class abiding by the authors' convention in Du et al. (2018). To set apart the source and target domains, we handpicked 20,891 daylight images, 11,489 images taken at night, and 5,179 foggy images from UAVDT based on the weather tags. The 5,179 UAVDT foggy images have encompassed both daytime and nighttime foggy conditions, consisting of 2,492 and 2,627 images, respectively. VisDrone2019-DET, on the other hand, contains 8,629 static images, inclusive of its training, validation, and testing sets. These snapshots are records from various drone platforms at different locations and heights. They are meticulously annotated with bounding boxes tagging objects falling into ten set classes like pedestrian, person, and vehicles like bicycles, cars, and tricycles.

To facilitate cross-dataset generalization experiments, we adopt the category settings from UAVDT. We exclusively uti-

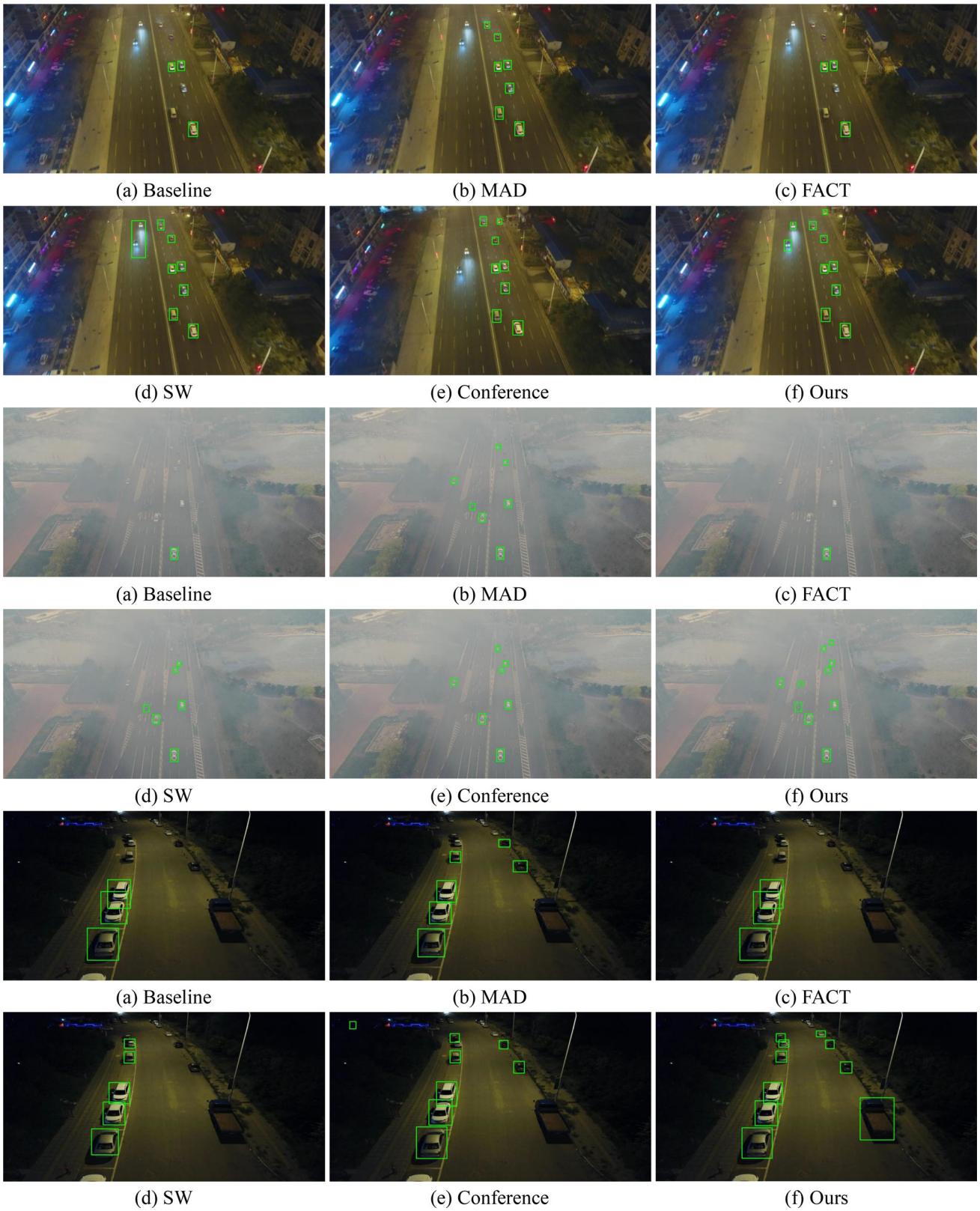


Fig. 6 Comparisons of the detection results of the baseline, the top-performing three comparative methods, the conference version, and our method. The images in the first group to the third group originate from UAVDT Nighttime, UAVDT Foggy, and Visdrone Nighttime, respectively. We employ Faster-RCNN as the detection model, with UAVDT Daylight as the source domain

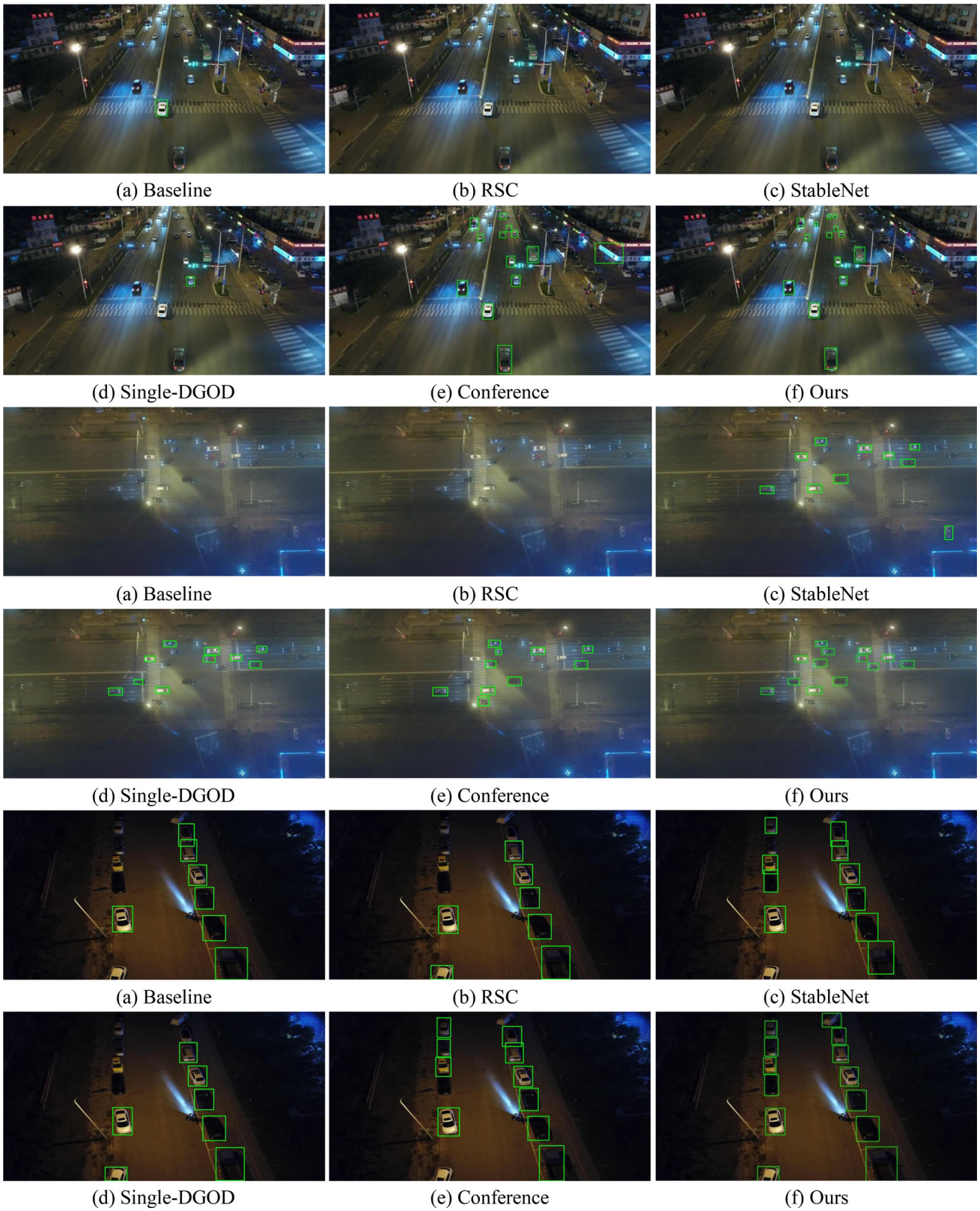


Fig. 7 Comparisons of the detection results of the baseline, the top-performing three comparative methods, the conference version, and our method. The images in the first group to the third group originate from UAVDT Nighttime, UAVDT Foggy, and Visdrone Nighttime, respectively. We employ YOLOv5 as the detection model, with UAVDT Daylight as the source domain

Table 3 Comparisons of the generalization performance of various DG methods

Methods	Avenue	UAVDT Nighttime			UAVDT Foggy			Visdrone Nighttime			Average		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Baseline (Girshick, 2015)	ICCV15	49.4	22.3	26.3	29.6	12.5	15.7	19.7	10.1	11.9	32.9	14.9	17.9
IBN-Net (Pan et al., 2018)	ECCV18	54.4	26.8	28.8	37.8	13.9	17.8	20.4	11.9	11.6	37.6	17.5	19.4
IterNorm (Huang et al., 2019)	CVPR19	41.6	14.5	19.2	28.6	11.9	14.9	12.5	9.1	8.4	27.6	11.8	14.2
JiGen (Carlucci et al., 2019)	CVPR19	49.7	21.6	25.2	29.1	12.4	15.1	18.4	7.9	9.6	32.4	14.0	16.6
SW (Pan et al., 2019)	ICCV19	64.0	30.1	33.0	42.5	12.3	18.5	21.9	11.1	12.5	42.8	17.8	21.3
RSC (Huang et al., 2020)	ECCV20	50.0	19.6	23.3	23.8	9.1	11.8	20.2	9.6	11.6	31.3	12.8	15.5
StableNet (Zhang et al., 2021)	CVPR21	44.8	20.4	23.5	30.2	13.1	15.1	17.5	6.9	9.0	30.8	13.4	15.9
FACT (Xu et al., 2021)	CVPR21	63.8	30.3	32.3	37.1	13.9	18.2	23.3	11.3	12.5	41.4	18.5	21.0
DIDN (Lin et al., 2021)	ICCV21	62.1	31.6	33.4	35.8	13.6	18.1	22.1	11.3	12.5	40.0	18.8	21.4
Single-DGOD (Wu & Deng, 2022)	CVPR22	60.3	30.3	32.6	38.4	13.3	17.8	20.1	10.3	11.1	39.6	18.0	20.5
MAD (Xu et al., 2023)	CVPR23	64.1	31.4	34.1	39.6	13.4	18.7	23.5	11.9	12.8	42.4	18.9	21.8
Conference (Wang et al., 2023a)	CVPR23	65.6	33.4	34.5	41.4	13.6	19.7	22.3	10.9	12.0	43.1	19.3	22.1
Ours	This work	**67.8**	**34.4**	**36.9**	**46.9**	**13.8**	**22.3**	**24.7**	**12.2**	**13.8**	**46.5**	**20.1**	**24.3**

We employ Faster-RCNN as the detection model, with UAVDT Daylight as the source domain, and UAVDT Nighttime, UAVDT Foggy, and Visdrone Nighttime as the unseen target domains for conducting experiments

We highlight the best results using **such** formatting

Table 4 Comparisons of the generalization performance of various DG methods

Methods	Avenue	UAVDT Nighttime			UAVDT Foggy			Visdrone Nighttime			Average		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Baseline (Girshick, 2015)	ICCV15	58.8	23.8	28.3	24.6	9.1	12.4	34.3	14.2	16.7	39.2	15.7	19.1
IBN-Net (Pan et al., 2018)	ECCV18	63.3	25.5	31.3	31.3	8.1	12.7	40.3	17.9	19.5	45.0	17.2	21.2
IterNorm (Huang et al., 2019)	CVPR19	56.5	22.1	27.7	29.2	9.1	12.7	28.0	13.2	14.6	37.9	14.8	18.3
JiGen (Carlucci et al., 2019)	CVPR19	58.6	24.1	29.2	27.9	9.1	12.3	34.5	14.5	17.6	40.3	15.9	19.7
SW (Pan et al., 2019)	ICCV19	55.7	21.1	26.0	23.9	8.6	11.4	32.8	13.6	15.0	37.5	14.4	17.5
RSC (Huang et al., 2020)	ECCV20	50.6	12.7	21.0	21.4	9.1	9.5	27.2	11.3	13.6	33.1	11.0	14.7
StableNet (Zhang et al., 2021)	CVPR21	61.4	27.2	31.0	31.4	10.2	14.4	32.7	14.7	16.4	41.8	17.4	20.6
FACT (Xu et al., 2021)	CVPR21	58.8	25.7	29.5	29.0	9.1	13.0	35.0	14.7	17.6	40.9	16.5	20.0
DIDN (Lin et al., 2021)	ICCV21	63.5	29.2	32.4	35.4	10.8	15.4	34.8	15.3	18.2	44.6	18.4	22.0
Single-DGOD (Wu & Deng, 2022)	CVPR22	61.5	26.8	30.9	29.7	9.1	14.5	34.4	15.0	17.9	41.9	16.9	21.1
MAD (Xu et al., 2023)	CVPR23	64.4	27.4	33.6	30.2	9.1	14.8	40.3	19.3	21.0	45.0	18.6	23.1
Conference (Wang et al., 2023a)	CVPR23	64.0	29.2	33.5	33.0	10.3	14.4	42.4	**20.0**	21.8	46.4	19.8	23.2
Ours	This work	**65.4**	**32.4**	**34.5**	**37.6**	**10.9**	**16.8**	**43.3**	**20.0**	**21.9**	**48.8**	**21.1**	**24.4**

We employ Faster-RCNN as the detection model, with Visdrone Daylight as the source domain, and UAVDT Nighttime, UAVDT Foggy, and Visdrone Nighttime as the unseen target domains for conducting experiments

We highlight the best results using **such** formatting

Table 5 Comparisons of the generalization performance of various DG methods

Methods	Avenue	UAVDT Nighttime			UAVDT Foggy			Visdrone Nighttime			Average		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Baseline (Girshick, 2015)	ICCV15	13.1	5.0	6.2	36.2	13.9	17.2	2.9	1.0	1.4	17.4	6.6	8.3
JiGen (Carlucci et al., 2019)	CVPR19	31.1	11.7	13.9	38.6	14.6	18.8	11.1	4.5	5.1	26.9	10.3	12.6
RSC (Huang et al., 2020)	ECCV20	17.0	7.2	8.5	44.3	19.9	22.7	6.5	2.2	2.9	22.6	9.8	11.4
StableNet (Zhang et al., 2021)	CVPR21	21.4	9.9	11.2	44.9	19.3	22.6	7.7	3.3	3.8	24.7	10.8	12.5
Single-DGOD (Wu & Deng, 2022)	CVPR22	34.1	15.3	17.4	43.2	17.8	21.1	11.7	4.8	5.8	29.7	12.6	14.8
Conference (Wang et al., 2023a)	CVPR23	42.4	21.0	22.5	46.2	17.5	18.0	19.7	5.3	8.1	36.1	14.6	16.2
Ours	This work	**52.1**	**29.8**	**30.4**	**49.3**	**19.8**	**21.1**	**24.7**	**8.2**	**10.9**	**42.0**	**19.3**	**20.8**

We employ YOLOv5 as the detection model, with UAVDT Daylight as the source domain, and UAVDT Nighttime, UAVDT Foggy, and Visdrone Nighttime as the unseen target domains for conducting experiments

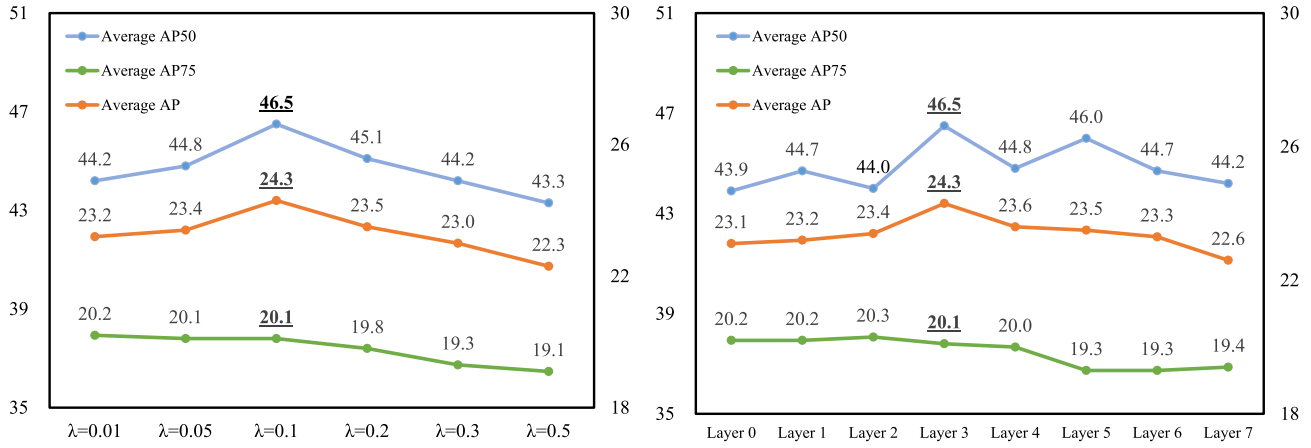
We highlight the best results using **such** formatting

Table 6 Comparisons of the generalization performance of various DG methods

Methods	Avenue	UAVDT Nighttime			UAVDT Foggy			Visdrone Nighttime			Average		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Baseline (Girshick, 2015)	ICCV15	40.3	16.2	19.6	24.8	3.0	8.4	24.5	10.0	11.8	29.9	9.7	13.3
JiGen (Carlucci et al., 2019)	CVPR19	32.0	11.0	14.8	22.9	2.7	7.7	20.4	7.6	9.4	25.1	7.1	10.6
RSC (Huang et al., 2020)	ECCV20	44.3	17.0	21.1	23.9	2.7	7.8	30.3	12.7	14.6	32.8	10.8	14.5
StableNet (Zhang et al., 2021)	CVPR21	51.8	20.0	24.9	29.2	3.1	9.6	29.5	12.9	14.6	36.8	12.0	16.4
Single-DGOD (Wu & Deng, 2022)	CVPR22	51.5	20.2	24.7	29.5	3.4	9.8	30.1	12.9	14.7	37.0	12.2	16.4
Conference (Wang et al., 2023a)	CVPR23	59.3	30.8	31.6	32.2	5.0	11.8	34.6	17.1	18.0	42.0	17.6	20.5
Ours	This work	**61.0**	**30.9**	**32.4**	**38.5**	**5.9**	**13.9**	**41.5**	**21.7**	**22.3**	**47.0**	**19.5**	**22.9**

We employ YOLOv5 as the detection model, with Visdrone Daylight as the source domain, and UAVDT Nighttime, UAVDT Foggy, and Visdrone Nighttime as the unseen target domains for conducting experiments

We highlight the best results using **such** formatting



(a) Ablation study of the Hyper-parameter λ

(b) Ablation study of the Plug-in Layer

Fig. 8 The visual charts of the ablation study of hyper-parameter λ and plug-in layer to improve readability

lize labels corresponding to the car, van, bus, and truck classes in the Visdrone2019-DET dataset and treat them as a single class. Given the lack of weather labels in Visdrone2019-DET, we manually picked segments with dim lighting to compile a dataset comprising 5,709 daylight and 1,698 nighttime images.

Ultimately, we treat the daylight portion from both UAVDT and Visdrone2019-DET as the source domain, respectively. Meanwhile, we considered the nighttime and foggy portions from UAVDT, along with the nighttime portion from Visdrone2019-DET, as three distinct unseen target domains for conducting experiments on both intra-dataset and cross-dataset generalization validations. For the exploratory experiments in the introduction, we utilize the daylight portion of UAVDT as the source domain for UAV-OD scenarios and the nighttime and foggy portions of UAVDT are considered as unseen target domains. In the case of generic object detection scenarios, we selected 26,798 daylight images from the BDD100k dataset (Yu et al., 2020) and 2,950 Cityscapes dataset daylight images (Cordts et al., 2016) as the source domain. Additionally, we choose 14,702 nighttime images from the BDD100k dataset and 2,950 foggy images from the Foggy Cityscapes dataset (Sakaridis et al., 2018) as unseen target domains.

Implementation Details. We evaluate the proposed method on the most popular detection networks, including Faster-RCNN (Girshick, 2015) with ResNet50 (He et al., 2016) as the backbone network and YOLOv5 (Jocher et al., 2020). Our framework is implemented in Pytorch with eight NVIDIA 1080ti GPUs. When training with Faster-RCNN, we train the framework for 30 epochs with a batch size of 16. We employ the SGD optimizer with a learning rate of 10^{-4}, a momentum of 0.9, and step learning rate decay. In the case of YOLOv5, the training is conducted for 300 epochs with a learning rate of 10^{-3}, a momentum of 0.9, lambda learning rate decay, and linear warmup for the initial five epochs. Regarding the learnable filters within our framework, these filters are optimized using SGD with a learning rate of 10^{-3}, a momentum of 0.9, and step learning rate decay. The temperature parameter τ is set to 0.7, while the hyper-parameter λ takes on a value of 0.1. For the evaluation protocol, we use the widely accepted criteria, including AP, AP_{50}, and AP_{75}.

5.2 Comparison with State-of-the-Arts

This section involves evaluating our method by drawing comparisons with other advanced Domain Generalization (DG) approaches. Beyond established domain-generalized object detection techniques such as DIDN (Lin et al., 2021), Single-DGOD (Wu & Deng, 2022) and MAD (Xu et al., 2023), we adopt and enhance various model-agnostic domain-generalization strategies to construct a unified domain-generalized object detection network for comparative analysis. Techniques like IBN-Net (Pan et al., 2018), Iternorm (Huang et al., 2019), and SW (Pan et al., 2019) aspire to augment network generalization using innovative normalization methods. We incorporated their proposed structures into the core network of UAV-OD and utilized the pretrained weights they provided for each approach. Jigen Carlucci et al. (2019) is a strategy designed to enhance DG representative depiction in a self-directed way. Following the scheme put forth in their research, we supplemented a supportive jigsaw classifier to UAV-OD networks, seeking to minimize the overall jigsaw loss at the image level. RSC Huang et al. (2020) is a dropout-based DG method that progressively eliminates dominant features activated during the training data cycle. StableNet Zhang et al. (2021) proposes sample reweighting to enhance generalization under distribution shifts. We directly compute RFF for image representations and implement image-wise reweighting.

We commence a series of comparative experiments engaging Faster R-CNN as our detection model. Separately, we utilize UAVDT Daylight and Visdrone Daylight as the source domain to train the UAV-OD network and acknowledge UAVDT Nighttime, UAVDT Foggy, and Visdrone Nighttime as unseen target domains for both intra-dataset and cross-dataset generalization testing. The experimental results are presented in Tables 3 and 4.

Revealed in Table 3, our method notably excels all others across every one of the twelve measures. In particular, concerning the average AP_{50}, AP_{75} and AP metrics, our method achieves an impressive increase over the baseline method by 13.6%, 5.2% and 6.4% while consistently exceeding the conference version by 3.4%, 0.8%, and 2.2%. We also surpass the best-performing method among comparative procedures by 3.7%, 1.2%, and 2.5%.

Similarly, as exposed in Table 4, our method consistently displays superior results across all metrics. There was a significant enhancement over the baseline method, with increases of9.6%, 5.4%, and 5.3% for average AP_{50}, AP_{75} and AP, respectively. Relative to the conference model, ours did better by 2.4%, 1.3%, and 1.2%. Additionally, we exceeded the best-performing method among other compared approaches by 3.8%, 2.5%, and 1.3%.

Furthermore, to validate our method's effectiveness, we apply YOLOv5 as the detection model. Given many comparative procedures mentioned previously are not readily adaptable to YOLOv5, we perform a comparative analysis with the remaining applicable approaches. We also train individually on two source domains measuring the UAV-OD generalization performance on three unseen target domains. As shown in Tables 5 and 6 for the test results, our proposed method leads across all metrics, establishing a new high standard in all benchmarks, especially as it frequently outperforms comparative strategies by a significant margin.

To conclude, considering the quantitative results shown in the tables, we select three best-performing comparative methods, the baseline, conference version, and our method, and analyze qualitative detection results on three unseen target domains, as exhibited in Figs. 8 and 10. Evidently, our approach displays superior detection precision, leading to fewer false positives and false negatives.

5.3 Ablation Study

In this section, we carried out ablation tests with Faster R-CNN as the detection network and UAVDT Daylight as the training dataset, unless otherwise explicitly suggested. These experiments are designed to delve into the contribution of individual components of our proposed methodology to its overall success.

Framework ablation. We first run an ablation study to verify the efficacy of each component within our proposed framework, as detailed in Table 7. We initially include frequency learnable filters as extra learnable parameters at the front end of the UAV-OD network without added supervision. Just this single addition results in a performance enhancement of 4.7%, 3.2%, and 2.2% regarding the average AP_{50}, AP_{75}, and AP metrics, respectively. This indicates the appropriateness of frequency learning for the UAV-OD task, as it can seize features conducive to generalization that in turn may not be assimilated in the spatial domain. Following this, the inclusion of both image-level and instance-level contrastive losses into the framework further boosts the generalization of UAV-OD. This substantiates our premise that our proposed two contrastive losses can aid the frequency learnable filter to extract a domain-invariant spectrum. These experiments conclusively verify the efficacy of each component in our framework.

Spatial or Frequency. We conduct experiments to compare the performance of spatial and frequency domain disentanglement. For the spatial domain disentanglement, we employ the spatial convolution block similar to the Single-DGOD (Wu & Deng, 2022) method, disentangling features at the feature level. Through experiments, as shown in Table 8, we observed that, for enhancing the generalization of UAV-OD, frequency domain disentanglement exhibits superior performance compared to spatial disentanglement.

Frequency disentanglement structure. We perform experiments to explore the comparative efficacy of different frequency disentanglement structures. We experimented with five structures incorporating learnable HW-shaped tensor, learnable HWC-shaped tensor, vanilla convolution, dilated convolution, and deformable convolution. As can be clearly seen from Table 9, among the five disentanglement methods in the frequency domain, the HWC-shaped learnable tensors outperform other structures.

Amplitude or Phrase. We conducted experiments to validate whether disentanglement is performed on the amplitude spectrum or the phase spectrum in the frequency domain. The experiment results are shown in Table 10. Our observations indicate that the generalization performance of decoupling the phase spectrum and simultaneously decoupling both the amplitude and phase spectra is inferior to that of decoupling the amplitude spectrum. A plausible explanation is that the phase spectrum is related to the structural information of the image, whereas the amplitude spectrum is associated with style information. Given the critical role of structural information in object detection, encompassing the target's positional details, decoupling the phase spectrum is likely to disrupt this crucial structural information, leading to a reduction in detection performance. Therefore, we only adjust the frequency domain amplitude spectrum.

Different training strategy. To evaluate the effectiveness of the alternating training strategy, we further implemented it within our framework. The alternating training approach

Table 7 Ablation study of each component in the proposed framework

Base	Filter	Ins-level	Img-level	UAVDT Nighttime			UAVDT Foggy			Visdrone Nighttime			Average		
				AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
✓				49.4	22.3	26.3	29.6	12.5	15.7	19.7	10.1	11.9	32.9	14.9	17.9
✓	✓			59.0	29.4	31.8	31.1	13.3	16.4	22.5	11.5	12.2	37.6	18.1	20.1
✓	✓	✓		66.9	34.0	36.5	42.0	13.6	20.2	24.1	11.9	13.1	44.3	19.8	23.3
✓	✓		✓	66.7	33.8	36.2	39.7	14.0	19.7	24.1	11.8	12.5	43.5	19.9	22.8
✓	✓	✓	✓	67.8	34.4	36.9	46.9	13.8	22.3	24.7	12.2	13.8	46.5	20.1	24.3

'Base' denotes baseline, 'Filter' denotes the learnable filters, 'Ins-level' denotes the instance-level contrastive loss, 'Img-level' denotes the image-level contrastive loss

Table 8 Ablation study of whether spatial or frequency domain disentanglement

Disentanglement	UAVDT Nighttime			UAVDT Foggy			Visdrone Nighttime			Average		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Spatial Convolution Block	64.4	32.8	35.1	40 0	13.3	19.2	22.2	10.8	11.4	42.1	19.0	21.9
Freqency Learnable HWC Tensor	67.8	34.4	36.9	46.9	13.8	22.3	24.7	12.2	13.8	46.5	20.1	24.3

For the spatial domain disentanglement, we employ the spatial convolution block similar to the Single-DGOD (Wu & Deng, 2022) method

Table 9 Ablation study of the frequency disentanglement structure

Disentanglement	UAVDT Nighttime			UAVDT Foggy			Visdrone Nighttime			Average		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Frequency Vanilla Convolution	67.5	31.9	34.6	37.3	13.9	19.1	24.9	12.4	13.5	43.2	19.4	22.4
Frequency Dilated Convolution	62.1	32.5	34.4	39.5	13.3	18.6	20.5	10.2	12.8	40.7	18.7	21.9
Frequency Deformable Convolution	64.1	32.8	34.3	39.8	13.4	19.0	22.1	10.7	11.0	42.0	19.0	21.4
Freqency Learnable HW Tensor	65.2	32.6	34.4	40.6	13.3	19.2	22.2	10.8	11.2	42.7	18.9	21.6
Freqency Learnable HWC Tensor	67.8	34.4	36.9	46.9	13.8	22.3	24.7	12.2	13.8	46.5	20.1	24.3

We adopt different disentanglement strategies, including HW-shaped and HWC-shaped learnable tensors, vanilla convolution, dilated convolution and deformable convolution

Table 10 Ablation study of whether frequency disentanglement is performed on amplitude spectrum or phase spectrum

Amptitude	Phrase	UAVDT Nighttime			UAVDT Foggy			Visdrone Nighttime			Average		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
✓		67.8	34.4	36.9	46.9	13.8	22.3	24.7	12.2	13.8	46.5	20.1	24.3
	✓	57.5	26.5	30.6	39.3	12.9	17.5	17.2	9.1	9.7	38.0	16.1	19.3
✓	✓	62.6	31.3	32.9	36.7	13.7	18.4	24.6	10.0	11.9	41.3	18.3	21.1

Table 11 Ablation study of the training strategy for the proposed framework

Training Strategy	UAVDT Nighttime			UAVDT Foggy			Visdrone Nighttime			Average		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Joint Training	67.4	34.9	37.0	43.4	13.9	20.8	24.4	12.0	12.4	45.1	20.3	23.4
Alternating Training	67.8	34.4	36.9	46.9	13.8	22.3	24.7	12.2	13.8	46.5	20.1	24.3

We adopt different strategies, including joint training and alternating training

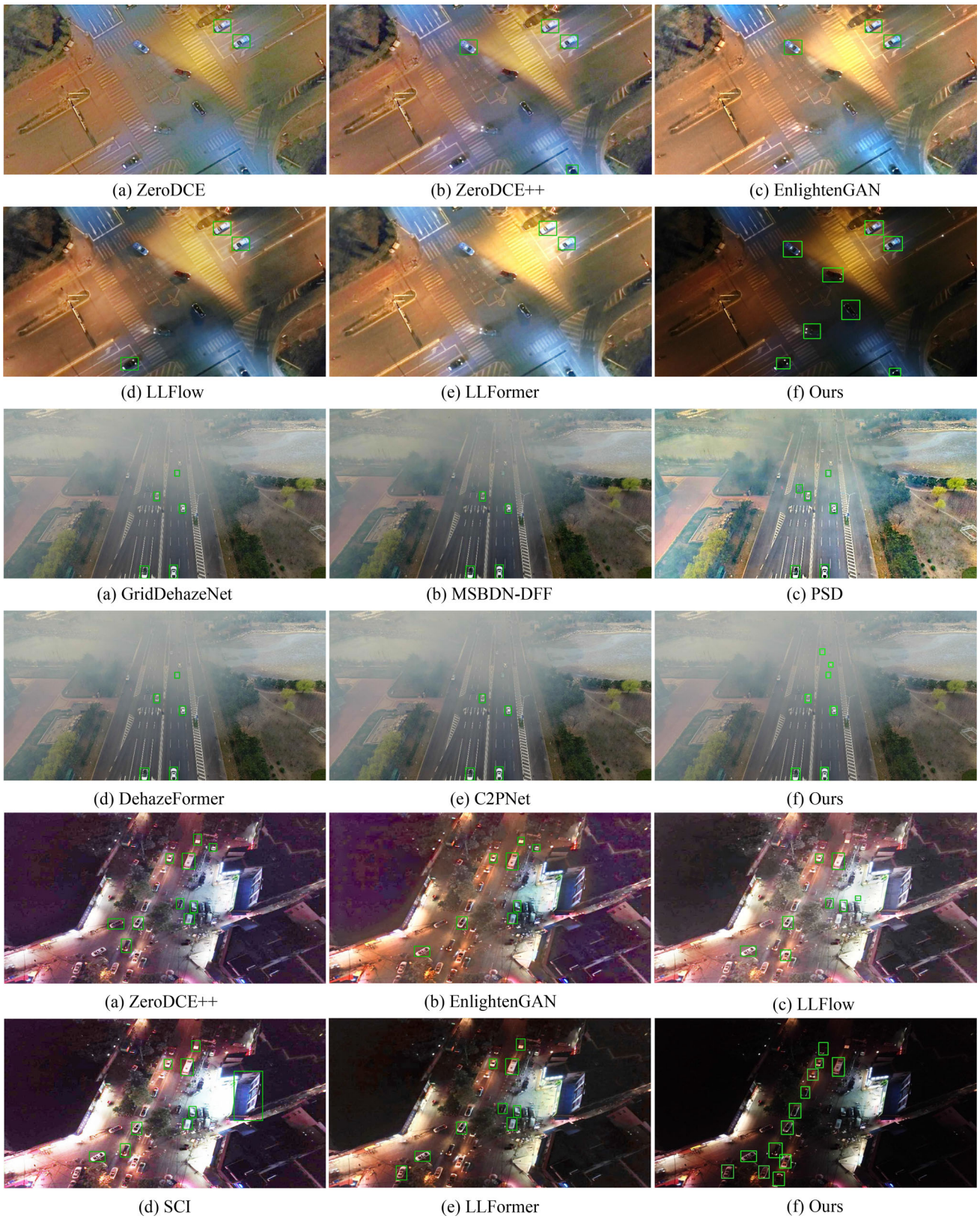


Fig. 9 Qualitative comparisons of the top-performing five two-stage methods, and our method. We employ Faster-RCNN as the detection model, with UAVDT Daylight as the source domain. The images in the first group to the third group originate from UAVDT Nighttime, UAVDT Foggy, and Visdrone Nighttime, respectively

Table 12 Ablation study of the hyper-parameter λ. We vary the setting of λ to investigate how λ affects the generalization

Hyper-parameter	UAVDT Nighttime			UAVDT Foggy			Visdrone Nighttime			Average		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
$\lambda = 0.01$	67.5	34.9	37.1	41.2	13.9	20.1	23.9	11.9	12.4	44.2	20.2	23.2
$\lambda = 0.05$	67.2	34.6	36.8	43.1	13.9	20.9	24.1	12.0	12.6	44.8	20.1	23.4
$\lambda = 0.1$	67.8	34.4	36.9	46.9	13.8	22.3	24.7	12.2	13.8	46.5	20.1	24.3
$\lambda = 0.2$	66.9	33.7	36.5	44.1	13.8	21.0	24.3	11.9	12.9	45.1	19.8	23.5
$\lambda = 0.3$	66.4	33.6	36.3	42.7	13.6	20.5	23.6	10.8	12.3	44.2	19.3	23.0
$\lambda = 0.5$	65.7	33.0	35.6	41.3	13.0	19.3	22.9	11.2	12.0	43.3	19.1	22.3

Table 13 Ablation study of the plug-in layer for calculating the contrastive loss

Plug-in Layer	UAVDT Nighttime			UAVDT Foggy			Visdrone Nighttime			Average		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Layer 0	67.2	34.6	37.0	40.6	14.0	19.9	24.1	11.9	12.3	43.9	20.2	23.1
Layer 1	67.6	34.7	37.0	42.4	14.0	20.6	24.1	11.9	12.0	44.7	20.2	23.2
Layer 2	67.4	34.9	37.1	40.7	14.0	20.0	24.1	11.9	13.1	44.0	20.3	23.4
Layer 3	67.8	34.4	36.9	46.9	13.8	22.3	24.7	12.2	13.8	46.5	20.1	24.3
Layer 4	66.8	34.2	36.3	43.5	13.9	21.0	24.0	12.0	13.6	44.8	20.0	23.6
Layer 5	67.2	34.1	36.6	46.9	13.8	22.3	23.8	10.0	11.5	46.0	19.3	23.5
Layer 6	65.8	32.2	35.5	44.3	13.9	21.5	24.0	11.9	12.9	44.7	19.3	23.3
Layer 7	67.3	33.0	36.4	41.9	13.7	20.1	23.3	11.4	11.4	44.2	19.4	22.6

We divide the backbone of Faster-RCNN into six segments and conduct experiments between each segment, including both the front-end and the back-end

Fig. 10 Visualization analysis of the domain-specific and domain-invariant components extracted from different domains. We also provide the high-frequency components of the input images as a reference. The images in the first row to the third row originate from UAVDT Nighttime, UAVDT Foggy, and Visdrone Nighttime, respectively

divides the learnable parameters of the entire framework into two portions: the first part comprises two learnable filters, and the other consists of the detection network. We individually optimize these two sets of parameters applying the contrastive loss and detection loss. We specifically fix the first part's parameters and optimize the second set of parameters using detection loss. Following this, we keep the second part of parameters constant and optimize the parameters of the first part using the contrastive loss. This cyclical process continues until the model converges. On the contrary, joint optimization doesn't partition the learnable parameters and instead optimizes them together based on the total loss function during the forward propagation of the network. As represented in Table 11, alternating optimization yields superior generalization performance. This is because alternating optimization can avoid clashes between two distinct optimization directions, thereby preventing suboptimal outcomes.

Hyper-parameter λ ablation. We then probe into how the changes in setting for hyper-parameter λ affect the generalization performance of the network. For this, we conduct several experiments with different λ values, and the outcomes are illustrated in Table 12 and Fig. 9a. Our evaluations clearly show that either too high or too low λ values result in a consistent drop in generalization performance. Interestingly, even a relatively small setting for lambda still brings about substantial enhancement in the generalization performance than when $\lambda = 0$, under which circumstances no extra contrastive losses are used for supervision. In the end, taking lambda as 0.1 emerges as the optimal hyper-parameter choice, leading to the highest generalization performance.

Contrastive loss after different layers. Finally, we proceed to present our experimental results where we computed the contrastive loss at different layers of the ResNet50. As shown in Table 13 and Fig. 9b, the analyses demonstrate that calculating the contrastive loss after the ResNet50's third layer leads to the highest generalization performance. This can be rationalized by acknowledging that the features drawn out from the more superficial layers possess high resolution, implying semantic sparsity, whereas features from the deeper layers embody low resolution, signaling semantic abstraction. Both of these conditions make them less appropriate as a feature space for optimizing the contrastive loss for feature disentanglement. As a result, the middle layers surface as a favourable compromise for this purpose.

5.4 Comparison with Two-Stage Methods

In this section, we evaluate our method by comparing it with several two-stage approaches. Two stage methods is to pre-process target domain images using off-the-shelf image restoration or enhancement methods and eliminate the negative effects of domain shift on person re-id. For nighttime images, we utilize ZeroDCE (Guo et al., 2020), ZeroDCE++ (Li et al., 2021), EnlightenGAN (Jiang et al., 2021b), LLFlow (Wang et al., 2022), SCI (Ma et al., 2022) and LLFormer (Wang et al., 2023b) as pre-processing low-light enhancement methods. For foggy images, we employ GridDehazeNet (Liu et al., 2019), FFA-Net (Qin et al., 2020), MSBDN-DFF (Dong et al., 2020), PSD (Chen et al., 2021), DehazeFormer (Song et al., 2023) and C2PNet (Zheng et al., 2023) as pre-processing restoration methods.

We integrate the aforementioned pre-processing strategies with Faster R-CNN that has been separately trained on UAVDT daylight and Visdrone daylight source domains. The quantitative and qualitative results are enumerated in Tables 14 and 15, and Fig. 11. It is evident that two-stage methods prioritize the enhancement of visual quality, demonstrating improvements in brightness and fog degradation for images with lower brightness or affected by fog after undergoing image enhancement. However, based on the detection results, it is apparent that these two-stage methods do not effectively consider the requirements of downstream detection networks. Consequently, they do not lead to a significant improvement in detection performance. In some instances, certain two-stage methods may even produce unnaturally restored images, creating a gap between the restored images and the natural images used to train Faster-RCNN. This gap can result in a decline in detection performance.

5.5 Comparison with Domain Adaption methods

To further evaluate the generalization ability, we compare our method with some domain adaptation method, including DAF (Chen et al., 2018), SWDA (Saito et al., 2019), HTCN (Chen et al., 2020), ICCR (Xu et al., 2020), VDD (Wu et al., 2021a) and TIA (Zhao & Wang, 2022).

To accommodate the configuration of DA methods, we further divide the 11,489 UAVDT nighttime images, 5,179 UAVDT foggy images, and 1,698 Visdrone nighttime images into subsets of 7,653 and 3,836 UAVDT nighttime images, 3,654 and 1,525 UAVDT foggy images, and 1,209 and 489 Visdrone nighttime images. The first subsets are used as training images in the target domain, while the second subsets are utilized for testing within the target domain. For the detection network, we employ Faster R-CNN and train it on the UAVDT daylight source domain. The experimental results are tabulated in Table 16.

When compared to the baseline, the DA methods have shown substantial improvements, with TIA's performance even exceeding that of our conference version. However, none of these methods can outdo our newly proposed approach. The main drawback of DA methods may be the limited quantity of data available from the target domain. Their performance might continue to improve and may even surpass our method if more target domain data were included.

Table 14 Comparisons of the generalization performance with several two-stage methods

Methods	Avenue	UAVDT Nighttime			Visdrone Nighttime		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Baseline (Girshick, 2015)	ICCV15	49.4	22.3	26.3	19.7	10.1	11.9
ZeroDCE (Guo et al., 2020)	CVPR20	47.1	19.7	24.1	15.2	6.3	7.8
ZeroDCE++ (Li et al., 2021)	TPAMI21	53.0	21.3	26.5	20.1	10.1	10.7
EnlightenGAN (Jiang et al., 2021b)	TIP21	43.3	19.1	22.4	16.2	9.5	9.9
LLFlow (Wang et al., 2022)	AAAI22	53.8	23.7	27.5	20.7	10.4	10.8
SCI (Ma et al., 2022)	CVPR22	44.4	17.5	22.2	17.1	9.1	10.0
LLFormer (Wang et al., 2023b)	AAAI23	41.6	16.9	21.0	16.7	9.1	9.1
Ours	This work	**67.8**	**34.4**	**36.9**	**24.7**	**12.2**	**13.8**

Methods	Avenue	UAVDT Foggy		
		AP_{50}	AP_{75}	AP
Baseline (Girshick, 2015)	ICCV15	29.6	12.5	15.7
GridDehazeNet (Liu et al., 2019)	ICCV19	33.0	13.1	16.8
FFA-Net (Qin et al., 2020)	AAAI20	30.0	12.5	15.7
MSBDN-DFF (Dong et al., 2020)	CVPR20	30.1	11.8	15.5
PSD (Chen et al., 2021)	CVPR21	33.6	12.0	14.2
DehazeFormer (Song et al., 2023)	TIP23	33.2	13.1	16.7
C2PNet (Zheng et al., 2023)	CVPR23	30.0	12.4	15.8
Ours	This work	**46.9**	**13.8**	**22.3**

We employ Faster-RCNN as the detection model, with UAVDT Daylight as the source domain, and UAVDT Nighttime, UAVDT Foggy, and Visdrone Nighttime as the unseen target domains for conducting experiments We highlight the best results using **such** formatting

Table 15 Comparisons of the generalization performance with several two-stage methods

Methods	Avenue	UAVDT Nighttime			Visdrone Nighttime		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Baseline (Girshick, 2015)	ICCV15	58.8	19.6	28.3	34.3	14.2	16.7
ZeroDCE (Guo et al., 2020)	CVPR20	55.4	21.8	26.4	31.4	11.9	14.6
ZeroDCE++ (Li et al., 2021)	TPAMI21	57.7	17.6	27.2	33.5	13.8	15.8
EnlightenGAN (Jiang et al., 2021b)	TIP21	54.5	22.5	26.0	31.0	8.8	15.0
LLFlow (Wang et al., 2022)	AAAI22	56.3	22.7	27.3	33.2	13.7	16.2
SCI (Ma et al., 2022)	CVPR22	49.5	18.1	23.2	30.5	11.5	14.2
LLFormer (Wang et al., 2023b)	AAAI23	56.4	20.4	25.8	33.2	13.9	16.0
Ours	This work	**65.4**	**32.4**	**34.5**	**43.3**	**20.0**	**21.9**

Methods	Avenue	UAVDT Foggy		
		AP_{50}	AP_{75}	AP
Baseline (Girshick, 2015)	ICCV15	24.6	9.1	12.4
GridDehazeNet (Liu et al., 2019)	ICCV19	27.8	8.0	11.6
FFA-Net (Qin et al., 2020)	AAAI20	24.9	8.5	11.3
MSBDN-DFF (Dong et al., 2020)	CVPR20	28.1	8.2	11.7
PSD (Chen et al., 2021)	CVPR21	28.6	8.9	12.5
DehazeFormer (Song et al., 2023)	TIP23	27.9	8.3	12.0
C2PNet (Zheng et al., 2023)	CVPR23	24.8	8.4	11.1
Ours	This work	**37.6**	**10.9**	**16.8**

We employ Faster-RCNN as the detection model, with Visdonre Daylight as the source domain, and UAVDT Nighttime, UAVDT Foggy, and Visdrone Nighttime as the unseen target domains for conducting experiments We highlight the best results using **such** formatting

Table 16 Comparisons of the generalization performance with several DA methods

Methods	Avenue	UAVDT Nighttime			UAVDT Foggy			Visdrone Nighttime			Average		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Baseline (Girshick, 2015)	ICCV15	49.4	22.3	26.3	29.6	12.5	15.7	19.7	10.1	11.9	32.9	14.9	17.9
DAF (Chen et al., 2018)	CVPR18	62.2	29.5	32.7	32.5	13.1	18.0	22.7	8.7	11.3	39.1	17.1	20.7
SWDA (Saito et al., 2019)	CVPR19	64.8	28.5	32.5	35.9	13.6	18.6	23.5	12.0	13.1	41.4	18.0	21.4
HTCN (Chen et al., 2020)	CVPR20	61.8	32.1	29.7	41.2	12.9	19.3	23.4	12.2	13.5	42.1	19.1	20.9
ICCR (Xu et al., 2020)	CVPR20	59.7	29.9	28.8	33.5	13.9	17.4	23.5	9.4	11.2	38.9	17.7	19.1
VDD (Wu et al., 2021a)	ICCV21	60.4	31.0	30.0	37.5	13.1	18.6	20.0	10.1	11.5	39.3	18.0	20.0
TIA (Zhao & Wang, 2022)	CVPR22	66.5	**34.4**	**37.0**	44.9	14.3	19.6	23.7	12.3	13.4	45.0	20.3	23.3
Conference (Wang et al., 2023a)	CVPR23	65.6	33.4	34.5	41.4	13.6	19.7	22.3	10.9	12.0	43.1	19.3	22.1
Ours	This work	**67.8**	**34.4**	36.9	**46.9**	**13.8**	**22.3**	**24.7**	**12.2**	**13.8**	**46.5**	**20.1**	**24.3**

We employ Faster-RCNN as the detection model, with UAVDT Daylight as the source domain, and UAVDT Nighttime, UAVDT Foggy, and Visdrone Nighttime as the unseen target domains for conducting experiments
We highlight the best results using **such** formatting

However, this is precisely the primary challenge that DA methods face. Amassing substantial target domain data for UAV-OD scenarios is a complicated task. Given the adaptability and mobility of UAVs, they can be operational in a range of environments, making it difficult to guarantee the availability of the target domain data. In contrast, our method achieves superior generalization performance with sole reliance on source domain data.

5.6 Efficiency

To gain insight into the efficiency of our approach, we implemented an analysis using Faster R-CNN with a ResNet50 backbone as the detection model. This was compared to the baseline, ten domain generalization (DG) methods, and the conference version of our method. To evaluate each method's efficiency, three widespread metrics were applied: latency (calculated based on model inference on a GTX 1080ti graphics card), floating-point operations per second (FLOPs), and the number of parameters. We also noted the average generalization performance across three unseen target domains for object detection tasks, measured using mean average precision (mAP) at various intersection-over-union (IoU) thresholds. The experimental results are shown in Table 17.

Our approach strives to disentangle domain-invariant features in the frequency domain, which subsequently introduces additional parameters and computational overhead during inference. This mainly stems from the integration of domain-invariant learnable filters and the implementation of Fourier and inverse Fourier transforms on images. However, our learnable filter structure utilizes element-wise multiplication with the amplitude spectrum, which enables disentanglement at the image level with fewer channels. This counteracts the significant computation and parameter burden found in methods like Single-DGOD (Wu & Deng, 2022), which decouples features using convolutions across more channels. In addition, the Fast Fourier Transform (FFT) exhibits lower computational cost. For a given $M \times N$ two-dimensional image, the computational complexity of the 2D Fourier transform is $O(M^2 \times N^2)$, as it involves $M \times N$ additions and $M \times N$ multiplications. However, in the case of the 2D FFT, it can be considered as two consecutive one-dimensional FFTs. Due to the use of the Cooley-Tukey algorithm, its computational complexity is reduced to $O(M \times N \log(M \times N))$. Furthermore, in real-world applications, various hardware and software techniques can be employed, including specialized hardware acceleration, parallel and distributed processing, pipeline processing, and quantization. These approaches enhance the operational speed of Fourier transforms, alleviating concerns about additional burdens resulting from Fourier transform computations. Compared to the baseline and DG methods that do not introduce extra inference parameters, our approach only slightly increases latency by 6.1ms, FLOPs by 0.96G, and parameters by 1.03M, while significantly enhancing generalization. The deployment of multi-dimensional filters in our conference version led to minor improvements across all metrics. To sum up, our method boosts generalization performance while maintaining efficiency comparable to the baseline.

5.7 Visualization

Image-level visualization. We undertake a visual examination of both domain-invariant and domain-specific elements, using the high-frequency parts from the original images as a benchmark. Our intention is to analyze which frequency bands play a crucial role in elevating the generalization capability of UAV-OD networks. The visual understanding gleaned from this analysis is depicted in Fig. 6.

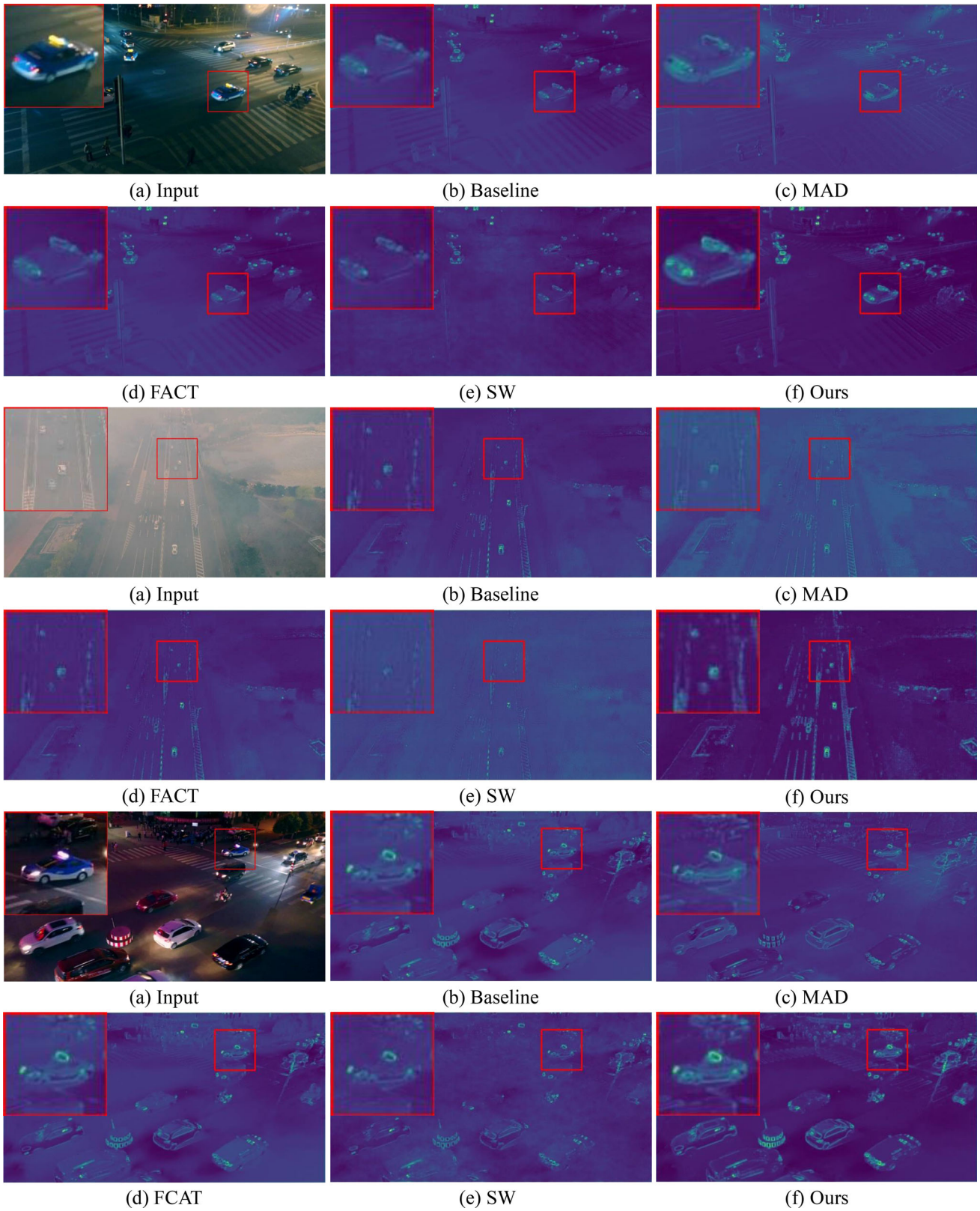


Fig. 11 Comparisons of the domain-invariant features extracted by the baseline, the top-performing three comparative methods, and our method. The images in the first group to the third group originate from UAVDT Nighttime, UAVDT Foggy, and Visdrone Nighttime, respectively

Table 17 Comparisons of the efficieny and performance of various DG methods

Methods	Avenue	Latency (ms)	FLOPs (G)	Params (M)	Average		
					AP_{50}	AP_{75}	AP
Baseline (Girshick, 2015)	ICCV15	84.6	128.03	28.29	32.9	14.9	17.9
IBN-Net (Pan et al., 2018)	ECCV18	90.8	135.94	28.29	37.6	17.5	19.4
IterNorm (Huang et al., 2019)	CVPR19	82.7	127.99	28.29	27.6	11.8	14.2
JiGen (Carlucci et al., 2019)	CVPR19	84.6	128.03	28.29	32.4	14.0	16.6
SW (Pan et al., 2019)	ICCV19	98.4	135.86	28.29	42.8	17.8	21.3
RSC (Huang et al., 2020)	ECCV20	84.6	128.03	28.29	31.3	12.8	15.5
StableNet (Zhang et al., 2021)	CVPR21	84.6	128.03	28.29	30.8	13.4	15.9
FACT (Xu et al., 2021)	CVPR21	84.6	128.03	28.29	41.4	18.5	21.0
DIDN (Lin et al., 2021)	ICCV21	84.6	128.03	28.29	40.0	18.8	21.4
Single-DGOD (Wu & Deng, 2022)	CVPR22	101.2	204.42	56.61	39.6	18.0	20.5
MAD (Xu et al., 2023)	CVPR23	84.6	128.03	28.29	42.4	18.9	21.8
Conference (Wang et al., 2023a)	CVPR23	90.3	128.88	28.63	43.1	19.3	22.1
Ours	This work	90.7	128.99	29.32	**46.5**	**20.1**	**24.3**

For efficiency, Latency, FLOPs, and Params are reported. For performance, the average generalization performance across three unseen target domain are reported. We employ Faster-RCNN as the detection model, with UAVDT Daylight as the source domain, and UAVDT Nighttime, UAVDT Foggy, and Visdrone Nighttime as the unseen target domains for conducting experiments

Firstly, the observation is made that despite variations in the visual appearances of images over different domains, the domain-invariant elements from these domains display a significant degree of resemblance. This suggests that the learnable filters are successful in capturing the domain-invariant spectrums, bridging gaps across diverse domains.

Secondly, in spite of our early experimental results suggesting that harnessing high-frequency information from images outperforms incorporating the entire frequency spectrum from a generalization standpoint, and that including mid-frequency or low-frequency data could result in detrimental effects, our visual results tell a somewhat different story. When comparing the domain-invariant parts to the high-frequency portions, it is revealed that the domain-invariant parts still retain some elements of low and mid-frequencies. This suggests that the relevance of mid-frequency and low-frequency data in images should not be entirely overlooked. It is clear that certain portions of low-frequency and mid-frequency information positively contribute to UAV-OD generalization. This highlights the benefits of using learning-based methods. In contrast to frequency-prior-based Domain Generalization (DG) techniques, our method allows for dynamic adaptation and helps in understanding which frequency bands are most beneficial for generalization, based on the peculiar characteristics of the task at hand.

Feature-level visualization. In Fig. 7, we present the domain-invariant features drawn by the baseline, the top three comparative methods, and our approach across three unseen target domains. It is evident that our method is capable of extracting more foreground object-related information while suppressing background-related information. This leads to a more robust disentanglement, as evidenced by such examples as the cars in the first and third night scenes, as well as the car in the second foggy scene.

5.8 Generality of the Framework

In this section, we delve into the generality of our framework. Within our framework, the frequency domain disentanglement module operates at the image level, and the computation of the contrastive loss takes place during the feature extraction stage. Consequently, our framework is not dependent on any particular design of the object detection network. We simply need to position the frequency domain disentanglement module ahead of the network and calculate the contrastive loss during the network's feature extraction stage to seamlessly integrate it into our proposed framework. In this paper, we have explored representative one-stage methods like Faster-RCNN and two-stage methods such as YOLOv5. Regarding transformer-based methods like DETR Carion et al. (2020), the integration process is straightforward: we simply incorporate the frequency domain disentanglement module before the network and compute the contrastive loss during the feature extraction stage of the transformer module. Consequently, our framework can be seamlessly combined and applied with transformer-based methods.

6 Limitation and Social Impact

Limitation. As illustrated in Table 17, although our method significantly enhances the generalization performance of UAV-OD, it introduces an additional frequency domain disentanglement module. Consequently, it brings extra parameters, inference time, and computational load during inference, creating an additional burden, particularly for UAV platforms with edge computing capabilities. In the future work, we would like to explore strategies for further reducing the additional overhead. One potential solution involves integrating the frequency domain disentanglement module with the convolutional module responsible for detection at the feature level. By replacing a convolutional module in the detection network with a frequency domain convolutional module (Chi et al., 2020), we establish a unified module for detection and disentanglement, thereby mitigating the additional computational costs.

Social impact. UAV applications are widespread in today's society, spanning various fields such as military operations, rescue missions, agriculture, and surveillance. This paper aims to enhance the generalization ability of UAV-OD in adverse environments, potentially yielding the following positive impacts on society: Firstly, it contributes to improving the security of UAV operations in military, rescue, and surveillance tasks. This enhancement helps to prevent misjudgments caused by interference in harsh environments, thereby averting significant economic losses and potential risks to personnel. Secondly, it facilitates the increased efficiency of UAV operations, diversifying the application scenarios of UAV and subsequently enhancing industrial and agricultural productivity. Lastly, it assists in more accurately monitoring and responding to natural disasters, thereby mitigating their impact on society.

7 Conclusion

This paper presents a novel approach to improving the generalizability of UAV-OD through frequency domain disentanglement, a more direct and efficient method of disentanglement. Initially, two adaptable filters are utilized to isolate the domain-neutral spectrums that positively affect generalization and the domain-specific spectrums that negatively influence it. Following this, we formulate two contrastive losses at both the image and instance levels, in order to guide the learning of these adaptable filters. Comprehensive tests with various detection models such as Faster-RCNN and YOLOv5, across numerous datasets including UAVDT and Visdrone2019-DET, showcase the superiority of our proposed technique.

Acknowledgements This work was supported by National Natural Science Foundation of China (NSFC) under Grants 62225207 and 62276243.

References

Cao, J., Cholakkal, H., Anwer, R.M., Khan, F. S., Pang, Y., & Shao, L. (2020). D2Det: Towards high quality object detection and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11485–11494).

Cao, S., Joshi, D., Gui, L. Y., & Wang, Y. X. (2023). Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23839–23848).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229). Springer.

Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B., & Tommasi, T. (2019). Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2229–2238).

Chen, C., Li, J., Zhou, H. Y., Han, X., Huang, Y., Ding, X., & Yu, Y. (2022a). Relation matters: Foreground-aware graph-based relational reasoning for domain adaptive object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(3), 3677–3694.

Chen, C., Zhang, Y., Lv, Q., Wei, S., Wang, X., Sun, X., & Dong, J. (2019). RRNet: A hybrid detector for object detection in drone-captured images. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*.

Chen, C., Zheng, Z., Ding, X., Huang, Y., & Dou, Q. (2020). Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8869–8878).

Chen, M., Chen, W., Yang, S., Song, J., Wang, X., Zhang, L., Yan, Y., Qi, D., Zhuang, Y., Xie, D., et al. (2022b). Learning domain adaptive object detection with probabilistic teacher. arXiv preprint arXiv:2206.06293

Chen, Y., Li, W., Sakaridis, C., Dai, D., & Van Gool, L. (2018). Domain adaptive faster R-CNN for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3339–3348).

Chen, Z., Wang, Y., Yang, Y., & Liu, D. (2021). PSD: Principled synthetic-to-real dehazing guided by physical priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7180–7189).

Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., Han, J. (2023). Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chi, L., Jiang, B., & Mu, Y. (2020). Fast Fourier convolution. *Advances in Neural Information Processing Systems, 33*, 4479–4488.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213–3223).

Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., & Yang, M. H. (2020). Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2157–2167).

Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., & Tian, Q. (2018). The unmanned aerial vehicle benchmark:

Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 370–386).
Duarte, A., Borralho, N., Cabral, P., & Caetano, M. (2022). Recent advances in forest insect pests and diseases monitoring using UAV-based data: A systematic review. *Forests, 13*(6), 911.
Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, PMLR (pp. 1180–1189).
Geraldes, R., Goncalves, A., Lai, T., Villerabel, M., Deng, W., Salta, A., Nakayama, K., Matsuo, Y., & Prendinger, H. (2019). UAV-based situational awareness system using deep learning. *IEEE Access, 7*, 122583–122594.
Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440–1448).
Guo, C., Li, C., Guo, J., Loy, C. C., Hou, J., Kwong, S., & Cong, R. (2020). Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1780–1789).
He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9729–9738).
He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
Hsu, C.C., Tsai, Y.H., Lin, Y.Y., & Yang, M.H. (2020a). Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *Computer Vision–ECCV 2020: 16th European conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16 (pp. 733–748). Springer.
Hsu, H. K., Yao, C. H., Tsai, Y. H., Hung, W. C., Tseng, H. Y., Singh, M., & Yang, M. H. (2020b). Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 749–757).
Huang, J., Guan, D., Xiao, A., Lu, S. (2021). FSDR: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6891–6902).
Huang, L., Zhou, Y., Zhu, F., Liu, L., & Shao, L. (2019). Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4874–4883).
Huang, Z., Wang, H., Xing, E. P., & Huang, D. (2020). Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16 (pp. 124–140). Springer.
Jeon, S., Hong, K., Lee, P., Lee, J., & Byun, H. (2021). Feature stylization and domain-aware contrastive learning for domain generalization. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 22–31).
Jiang, J., Chen, B., Wang, J., & Long, M. (2021a). Decoupled adaptation for cross-domain object detection. arXiv preprint arXiv:2110.02578
Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., & Wang, Z. (2021b). Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing, 30*, 2340–2349.
Jocher, G., Changyu, L., Hogan, A., Yu, L., Rai, P., Sullivan, T., et al. (2020). ultralytics/yolov5: Initial release. Zenodo
Kajiura, N., Liu, H., & Satoh, S. (2021). Improving camouflaged object detection with the uncertainty of pseudo-edge labels. In *ACM multimedia Asia* (pp. 1–7).
Kiefer, B., Ott, D., & Zell, A. (2022). Leveraging synthetic data in object detection on unmanned aerial vehicles. In *2022 26th international conference on pattern recognition (ICPR)* (pp. 3564–3571). IEEE.
Lee, S., Bae, J., & Kim, H.Y. (2023). Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11776–11785).
Li, C., Guo, C., & Loy, C. C. (2021). Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(8), 4225–4238.
Li, D., Huang, J.B., Li, Y., Wang, S., & Yang, M. H. (2016). Weakly supervised object localization with progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3512–3520).
Li, W., Liu, X., Yao, X., & Yuan, Y. (2022a). Scan: Cross domain object detection with semantic conditioned adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*, 1421–1428.
Li, W., Liu, X., Yuan, Y. (2022b). Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5291–5300).
Li, Y. J., Dai, X., Ma, C. Y., Liu, Y. C., Chen, K., Wu, B., He, Z., Kitani, K., & Vajda, P. (2022c). Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7581–7590).
Lin, C., Yuan, Z., Zhao, S., Sun, P., Wang, C., & Cai, J. (2021). Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8771–8780).
Lin, S., Zhang, Z., Huang, Z., Lu, Y., Lan, C., Chu, P., You, Q., Wang, J., Liu, Z., Parulkar, A., et al. (2023). Deep frequency filtering for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11797–11807).
Liu, H., Song, P., & Ding, R. (2020a). Towards domain generalization in underwater object detection. In *2020 IEEE international conference on image processing (ICIP)* (pp. 1971–1975). IEEE.
Liu, M., Wang, X., Zhou, A., Fu, X., Ma, Y., & Piao, C. (2020). UAV-YOLO: Small object detection on unmanned aerial vehicle perspective. *Sensors, 20*(8), 2238.
Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P. A. (2021). 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* (pp. 1013–1023).
Liu, X., Ma, Y., Shi, Z., & Chen, J. (2019). Griddehazenet: Attention-based multi-scale network for image dehazing. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7314–7323).
Liu, Y., Wang, J., Huang, C., Wang, Y., & Xu, Y. (2023). CIGAR: Cross-modality graph reasoning for domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23776–23786).
Lu, Y., Zhong, Z., & Shu, Y. (2023). Multi-view domain adaptive object detection on camera networks. In *AAAI*.
Lygouras, E., Santavas, N., Taitzoglou, A., Tarchanidis, K., Mitropoulos, A., & Gasteratos, A. (2019). Unsupervised human detection with an embedded vision system on a fully autonomous UAV for search and rescue operations. *Sensors, 19*(16), 3542.
Ma, L., Ma, T., Liu, R., Fan, X., & Luo, Z. (2022). Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5637–5646).

Mittal, P., Singh, R., & Sharma, A. (2020). Deep learning-based object detection in low-altitude UAV datasets: A survey. *Image and Vision Computing, 104*, 104046.

Nussbaumer, H. J., & Nussbaumer, H. J. (1982). *The fast Fourier transform*. Springer.

Pan, X., Luo, P., Shi, J., & Tang, X. (2018). Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 464–479).

Pan, X., Zhan, X., Shi, J., Tang, X., & Luo, P. (2019). Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1863–1871).

Qin, X., Wang, Z., Bai, Y., Xie, X., & Jia, H. (2020). FFA-Net: Feature fusion attention network for single image dehazing. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*, 11908–11915.

Saito, K., Ushiku, Y., Harada, T., & Saenko, K. (2019). Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6956–6965).

Sakaridis, C., Dai, D., & Van Gool, L. (2018). Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision, 126*, 973–992.

San, K. T., Mun, S. J., Choe, Y. H., & Chang, Y. S. (2018). UAV delivery monitoring system. In *MATEC web of conferences, EDP Sciences* (Vol. 151, p. 04011).

Song, Y., He, Z., Qian, H., & Du, X. (2023). Vision transformers for single image dehazing. *IEEE Transactions on Image Processing, 32*, 1927–1941.

Sun, K., Liu, H., Ye, Q., Gao, Y., Liu, J., Shao, L., & Ji, R. (2021a). Domain general face forgery detection by learning to weight. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*, 2638–2646.

Sun, W., Dai, L., Zhang, X., Chang, P., & He, X. (2021b). RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring. *Applied Intelligence* 1–16.

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7167–7176).

Vidit, V., Engilberge, M., & Salzmann, M. (2023). Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3219–3229).

Wang, K., Fu, X., Huang, Y., Cao, C., Shi, G., Zha, Z. J. (2023a). Generalized uav object detection via frequency domain disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1064–1073).

Wang, T., Zhang, K., Shen, T., Luo, W., Stenger, B., & Lu, T. (2023b). Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*, 2654–2662.

Wang, Y., Wan, R., Yang, W., Li, H., Chau, L. P., & Kot, A. (2022). Low-light image enhancement with normalizing flow. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*, 2604–2612.

Wu, A., & Deng, C. (2022). 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* (pp. 847–856).

Wu, A., Liu, R., Han, Y., Zhu, L., & Yang. Y. (2021a). Vector-decomposed disentanglement for domain-invariant object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9342–9351).

Wu, X., Li, W., Hong, D., Tao, R., & Du, Q. (2021). Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey. *IEEE Geoscience and Remote Sensing Magazine, 10*(1), 91–124.

Wu, Z., Suresh, K., Narayanan, P., Xu, H., Kwon, H., & Wang, Z. (2019). 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* (pp. 1201–1210).

Xu, C. D., Zhao, X. R., Jin, X., & Wei, X. S. (2020). Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11724–11733).

Xu, M., Qin, L., Chen, W., Pu, S., & Zhang, L. (2023). Multi-view adversarial discriminator: Mine the non-causal factors for object detection in unseen domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8103–8112).

Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q. (2021). A Fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14383–14392).

Yang, Q., Niu, H., Xia, P., Zhang, W., & Li, B. (2023). Frequency decomposition to tap the potential of single domain for generalization. arXiv preprint arXiv:2304.07261

Yang, Y., & Soatto, S. (2020). FDA: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4085–4095).

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., & Darrell, T. (2020). BDD100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2636–2645).

Yu, W., Yang, T., & Chen, C. (2021). Towards resolving the challenge of long-tail distribution in UAV images for object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3258–3267).

Zhang, P., Zhong, Y., & Li, X. (2019). Slimyolov3: Narrower, faster and better for real-time UAV applications. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*.

Zhang, X., Cui, P., Xu, R., Zhou, L., He, Y., & Shen, Z. (2021). Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5372–5382).

Zhang, X., Xu, Z., Xu, R., Liu, J., Cui, P., Wan, W., Sun, C., & Li, C. (2022). Towards domain generalization in object detection. arXiv preprint arXiv:2203.14387

Zhao, L., & Wang, L. (2022). Task-specific inconsistency alignment for domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14217–14226).

Zhao, Y., Zhong, Z., Zhao, N., Sebe, N., & Lee, G. H. (2023). Style-hallucinated dual consistency learning: A unified framework for visual domain generalization. *International Journal of Computer Vision*.

Zheng, Y., Huang, D., Liu, S., & Wang, Y. (2020). Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13766–13775).

Zheng, Y., Zhan, J., He, S., Dong, J., & Du, Y. (2023). Curricular contrastive regularization for physics-aware single image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5785–5794).

Zhong, Z., Zhao, Y., Lee, G. H., & Sebe, N. (2022). Adversarial style augmentation for domain generalized urban-scene segmentation. *Advances in Neural Information Processing Systems, 35*, 338–350.

Zhou, Z., Li, H., Liu, H., Wang, N., Yu, G., & Ji, R. (2023). Star loss: Reducing semantic ambiguity in facial landmark detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15475–15484).

Zhu, P., Du, D., Wen, L., Bian, X., Ling, H., Hu, Q., Peng, T., Zheng, J., Wang, X., Zhang, Y., et al. (2019). Visdrone-vid2019: The vision meets drone object detection in video challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*.

Zhuang, C., Han, X., Huang, W., & Scott, M. (2020). iFAN: Image-instance full alignment networks for adaptive object detection. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*, 13122–13129.