Generalized UAV Object Detection via Frequency Domain Disentanglement : Supplementary Material

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

1. Inference Stage

Fig. 1 shows the network details for the inference stage. During inference, the domain-invariant learnable filter 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.

2. Ablation Analysis of Hyper-parameter λ

Tab. 1 gives the quantitative results of the ablation analysis of the proposed framework's hyper-parameter λ. We investigate how varying the setting of hyper-parameter λ affects the network's generalization performance. We can see that different settings of λ affect the generalization performance. Setting λ to values too large or too small can degrade the network's generalization performance. When λ is set to 0.15, the performance is the best. The ablation analysis of the contrastive loss can be also found in Tab. 1, verifying the efficacy of the contrastive loss.

3. Ablation Analysis of the Backbone Division

Tab. 2 shows the quantitative results of the ablation analysis of the backbone division. To demonstrate the effect of different backbone divisions on the generalization capability of the UAV-OD network, we select different backbone divisions based on the structure of the backbone and conduct experiments. We can observe that selecting block four as a partition achieves the optimal results.

4. More Visualization Analysis

Image-level visualization. In Fig. 2, we provide more full-resolution image-level visualization examples. We can observe that despite the appearance of the image varies across domains, the domain-invariant component from each domain appears similar. For the domain-specific component, there is a clear separation between the foreground and background, with the foreground consisting of a darker color to indicate less attention and the background consisting of a brighter color to indicate more attention.

Feature-level visualization. As shown in Fig. 3, we provide more full-resolution feature-level visualizations of our method. For the domain-invariant feature, the foreground region is typically brighter than the background region, indicating that the domain-invariant feature focuses more on the image's foreground. For the domain-specific feature, the background region is typically brighter than the foreground region, indicating that the domain-specific feature emphasizes the image's background. Therefore, our method effectively separates the invariant and specific features.

5. Comparisons of the Training Time

Tab. 3 (a) demonstrates that the training time of our approach is comparable to other methods. Contrastive learning does not require a lot of time-consuming.

6. Discussion of Diverse Illumination Results.

Diverse illumination is a domain shift that varies more from global properties than others. As the frequency domain obeys global modeling, our method can handle it better. In addition, we have conducted experiments on the relighting and the results are shown in Tab. 3 (b). Due to the gap between low- and high-level vision tasks, relighting will hinder the generalizability.

7. Limitation

For limitation, our approach is an initial exploration of learning domain generalized UAV-OD network via frequency domain disentanglement. More subtle designs can be considered, leaving enough space for further development. Furthermore, our method is evaluated on three unseen target domains: various scene structures, diverse illumination conditions, and adverse weather conditions. We will consider more unseen target domains in future work to validate our method's effectiveness.

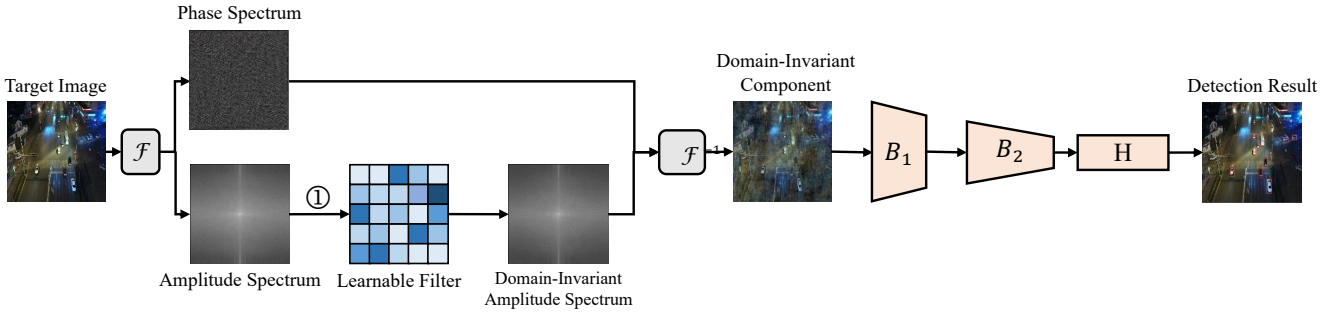


Figure 1. Illustration of the proposed method for the testing stage. $\mathcal{F}$ and $\mathcal{F}^{-1}$ indicate FFT and IFFT. The backbone of UAV-OD network is divided into B_1 and B_2. H represents the detection head of UAV-OD network, the lines marked with ① represent element-wise multiplication. During inference, we directly use the domain-invariant component to make predictions.

Hyper-parameter	Various Scene			Diverse Illumination			Adverse Weather			Average		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
$\lambda = 1.0$	69.2	38.4	38.9	34.3	15.6	17.7	45.0	14.4	19.6	49.5	22.8	25.4
$\lambda = 0.5$	71.6	45.9	42.5	**39.5**	**18.5**	**20.8**	44.7	15.9	20.4	51.9	26.8	27.9
$\lambda = 0.3$	70.6	45.4	42.2	41.2	17.4	20.9	46.5	16.4	21.3	52.8	26.4	28.1
$\lambda = 0.15$	75.1	49.7	45.3	39.0	**18.5**	20.7	**48.0**	**17.2**	**22.3**	**54.0**	**28.4**	**29.4**
$\lambda = 0.1$	71.9	40.0	39.9	18.0	4.72	7.5	38.0	12.2	17.0	42.6	19.0	21.5
$\lambda = 0.05$	**83.1**	**63.7**	**54.2**	13.5	2.9	5.4	30.0	8.2	12.4	42.2	24.9	24.0
$\lambda = 0$	72.0	46.1	43.3	31.7	13.6	16.9	38.5	11.8	17.7	47.4	23.8	26.0

Table 1. Quantitative results of the ablation analysis of the proposed framework's hyper-parameter λ, which balances $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{det}}$.

Block	Various Scene			Diverse Illumination			Adverse Weather			Average		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
0	65.7	37.4	37.0	23.6	8.9	11.5	45.3	12.1	18.2	44.9	19.5	22.2
1	71.5	44.3	41.8	34.4	13.9	17.1	47.0	16.1	21.5	51.0	24.8	26.8
2	72.3	47.6	43.5	33.0	15.8	17.4	47.0	15.7	21.3	50.8	26.4	27.4
3	73.3	48.2	44.3	38.1	18.4	20.0	46.8	16.6	21.7	52.7	27.7	28.7
4	**75.1**	**49.7**	**45.3**	**39.0**	18.5	**20.7**	**48.0**	**17.2**	**22.3**	**54.0**	**28.4**	**29.4**
5	74.7	49.5	45.1	29.3	14.0	15.3	45.6	15.5	20.6	49.9	26.3	27.0
6	74.5	48.9	44.8	30.3	14.3	15.9	46.4	15.7	21.1	50.4	26.3	27.3
7	72.8	46.1	42.9	39.4	**18.6**	20.8	46.4	16.4	21.3	52.9	27.0	28.3
8	70.8	43.1	41.3	33.2	12.2	15.9	48.5	16.2	22.2	50.8	23.8	26.5
9	69.7	37.5	38.6	29.6	11.0	14.0	48.3	14.2	21.1	49.2	20.9	24.6

Table 2. Quantitative results of the ablation analysis of the backbone division. The UAV-OD network's backbone is divided according to the specified block.

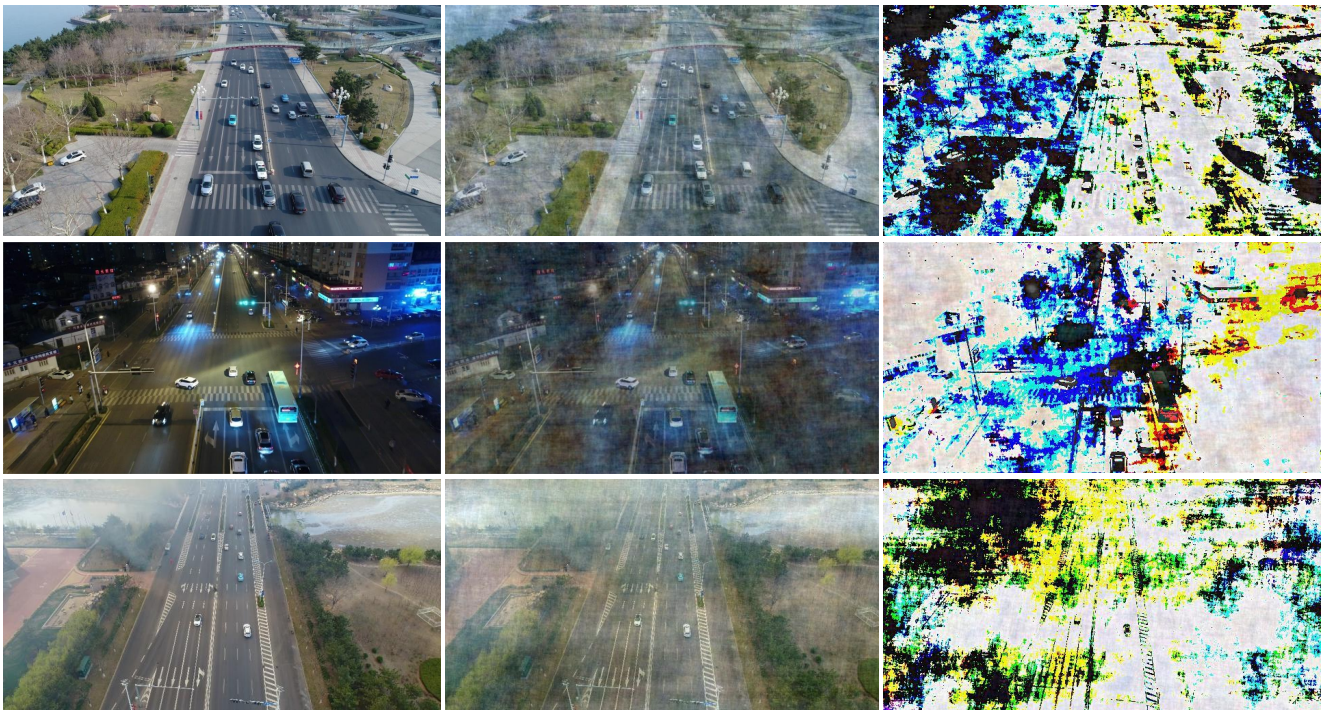
Method	Baseline	JiGen	RSC	StableNet	Single-DGOD	Ours
Duration	9h44m	16h33m	17h57m	21h55m	16h47m	16h42m

(a)

Method	LIME	Enlighten	ZeroDCE	Ours
AP	9.0	9.9	11.3	20.7

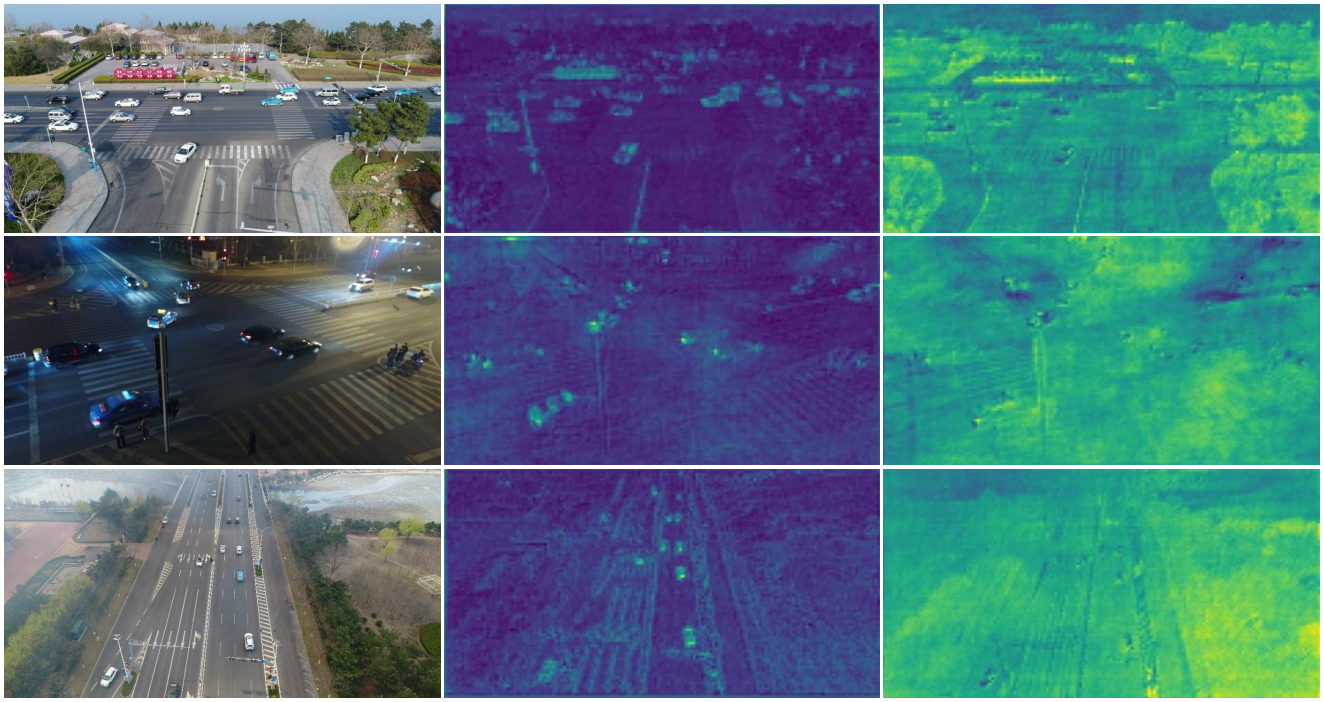
(b)

Table 3. (a) Comparisons of the training time. Duration are reported. (b) Comparisons of the relighting methods and ours. AP are reported.



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

Figure 2. Visualization analysis of the domain invariant and domain-specific components extracted from different domains. The first, second and third rows indicate the target domains with various scene structures, diverse illumination conditions, adverse weather condition.



(a) Input image (b) Domain-invariant feature (c) Domain-specific feature

Figure 3. Visualization analysis of the domain invariant and domain-specific features extracted from different domains. The first, second and third rows indicate the target domains with various scene structures, diverse illumination conditions, adverse weather conditions.