Driving-Video Dehazing with Non-Aligned Regularization for Safety Assistance

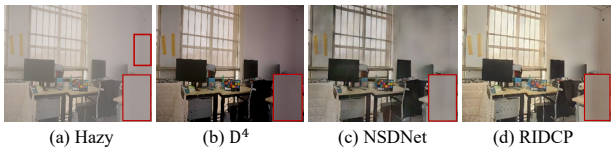
Supplementary Material

In this supplementary material, we provide an experiment on REVIDE dataset in Appendix A and more datasets details B. Next, we present additional ablation studies and discussions in Appendix C and Appendix D, respectively. In Appendix E, we showcase more visual results, including alignment results and video dehazing results.

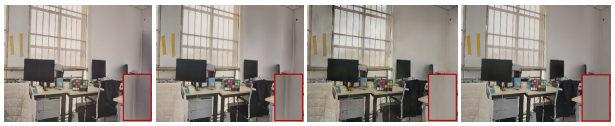
A. Experiment on REVIDE dataset.

Data Settings	Methods	REVIDE		Runtime (s)	Ref.
		PSNR ↑	SSIM ↑		
Unpaired	DCP [20]	11.03	0.7285	1.39	CVPR'09
	RefineNet [72]	23.24	0.8860	0.105	TIP'21
	CDD-GAN [6]	21.12	0.8592	0.082	ECCV'22
	D^4 [63]	19.04	0.8711	0.078	CVPR'22
Paired	PSD [7]	15.12	0.7795	0.084	CVPR'21
	RIDCP [59]	22.70	0.8640	0.720	CVPR'23
	PM-Net [38]	23.83	0.8950	0.277	ACMM'22
	MAP-Net [60]	24.16	**0.9043**	0.668	CVPR'23
Non-aligned	NSDNet [15]	23.52	0.8892	**0.075**	arXiv'23
	DVD (Ours)	**24.34**	0.8921	0.488	-

Table S1. Comparison of the proposed method and methods with aligned ground truth on REVIDE dataset.



(a) Hazy (b) D^4 (c) NSDNet (d) RIDCP



(e) PM-Net (f) MAP-Net (g) Ours (h) Reference

Figure S1. Visual comparison on REVIDE dataset.

To further verify the effectiveness of our proposed method, we evaluate the proposed method against SOTA methods that require aligned ground truths. Table S1 reports the evaluation results on the REVIDE dataset in terms of PSNR and SSIM. We can see that our proposed method obtains higher PSNR. In this work, we mainly focus on the real-world video dehazing in driving scenarios. However, we have also obtained good results on smoke data (REVIDE), indicating that our method is effective for both smoke/haze removal.

We further present visual observation comparisons in Fig. S1. The dehazing results of all the competitive methods contain artifacts, and the detail restoration is not ideal. In contrast, the proposed method generates much clearer results that are visually closer to the ground truth.

B. More datasets details

B.1. Spatio-temporal Misalignment Causes.

Here, due to real-world collection scenarios, as depicted in Fig. 1, avoidance maneuvers for pedestrians and vehicles on the road result in varying durations of collected hazy/clear video pairs with the same starting and ending points. Consequently, temporal misalignment occurs in hazy/clear video pairs. Similarly, avoidance maneuvers also lead to different shooting trajectories, causing spatial misalignment (*i.e.*, pixel misalignment). Additionally, the dynamic movement of pedestrians and vehicles contributes to spatial misalignment (*i.e.*, semantic misalignment).

B.2. Compare with Other Datasets

Compared to the 1981 pairs of indoor smoke data from the REVIDE [71] dataset, our non-aligned dataset GoProHazy consists of a total of 4256 pairs, and the no-reference DrivingHazy dataset comprises 1807 frames of hazy images. Moreover, our outdoor scenes are more numerous and realistic compared to indoor settings. Furthermore, in contrast to the large-scale synthetic dataset HazeWorld from MAP-Net [60], our proposed GoProHazy and DrivingHazy datasets represent real driving scenarios under real-world hazy weather conditions. This makes them more valuable for research aimed at addressing dehazing in videos captured under real-world conditions.

C. More Ablation Studies

The number of input frames. Table S2 demonstrates that optimal performance is achieved when using a three-frame input. This is attributed to the advantage of utilizing multiple frames to mitigate alignment issues, but it also introduces cumulative errors in alignment. As shown in Fig. S2, we also present the influence of different input frames on $\mathcal{L}_{\text{mfr}}$. Here, balancing efficiency considerations, we choose two frames as the input.

Number of Input frames	2 (Ours)	3	4
FADE ↓	0.7598	**0.7204**	0.7634
NIQE ↓	3.7753	**3.7392**	3.7984

Table S2. Ablation study for the number of input frames on GoProHazy dataset.

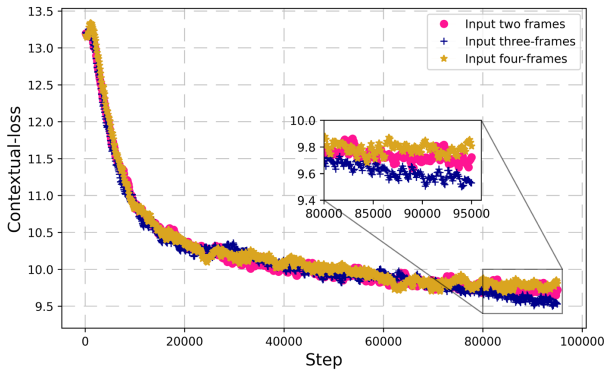


Figure S2. The influence of different input frames on $\mathcal{L}_{\text{mfr}}$

D. More Discussions

The impact of non-aligned scale. No doubt, the more aligned the hazy/clear frame pairs, the better the dehazing effect. However, our primary focus here is on the boundary issues related to non-aligned scales. In the ablation experiments of NSDNet [15], it was revealed that, compared to cases with ground truth (GT), a non-aligned pixel offset exceeding 90 pixels (for an image size of 256×256) results in a 0.7 dB decrease in PSNR, a 0.2 reduction in structural similarity (SSIM), and a decrease of 0.02 and 0.5 in FADE [8] and NIQE [41], respectively. We think that, in contrast to training with synthetic datasets, which may result in suboptimal dehazing in real-world scenes, the minor performance decline introduced by non-alignment is entirely acceptable. Moreover, during real-world data collection, we can easily control non-alignment within 90 pixels.

E. More Visual Results

The visualization of FCAS module. Here, we visualize the effectiveness of the flow-guided attention sampler (FCAS) in feature alignment, as shown in Fig. S3. We observe that the features aligned by the FCAS module are nearly consistent with the features of the current frame. The optical flow used to guide sampling is visualized in Fig. S3 (c). *Note that the ablation study on the FCAS is visualized in the main text.*

More visualizations of video dehazing. We present additional visual comparison results with state-of-the-art image/video dehazing methods on the GoProHazy dataset in Fig. S6. We observe that our proposed DVD method outperforms in dehazing performance, particularly in distant visibility and local detail restoration (*i.e.*, texture and brightness of scenes). The same dehazing issues are evident in the visual comparisons on the DrivingHazy and InternetHazy datasets. We present their visual comparisons separately in Fig. S7 and Fig. S8.

Applications. To highlight the benefits of dehazing results for downstream tasks, we employ the image segmentation

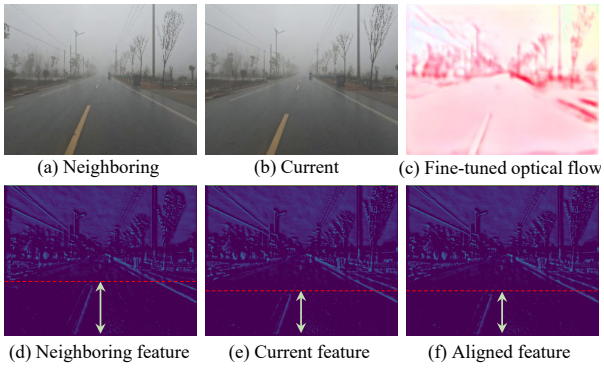


Figure S3. The visualization of FCAS module.

Figure S4. Visual results of object detection on the InternetHazy dataset.

method FastSAM [73] [1] to assess the gains brought by various image/video dehazing methods. The test results, as shown in Fig. S9, reveal that our proposed method achieves superior segmentation performance, particularly in the sky region. For the parameter settings of FastSAM, We employed the FastSAM-x model, setting the intersection over union (IoU) to 0.8 and the object confidence to 0.005.

In Fig. S4, we conducted an object detection (yolov8[2]) to validate the driving-safety assistance. We see that both vehicles and pedestrians are readily detected, enabling early detection by drivers and ensuring their safe operation.

Video demo. To validate the stability of our video dehazing results, we present a video result captured in a real driving environment and compare it with the latest video dehazing state-of-the-art method, MAP-Net [60]. We have included this video-demo.mp4 file in the supplementary materials.

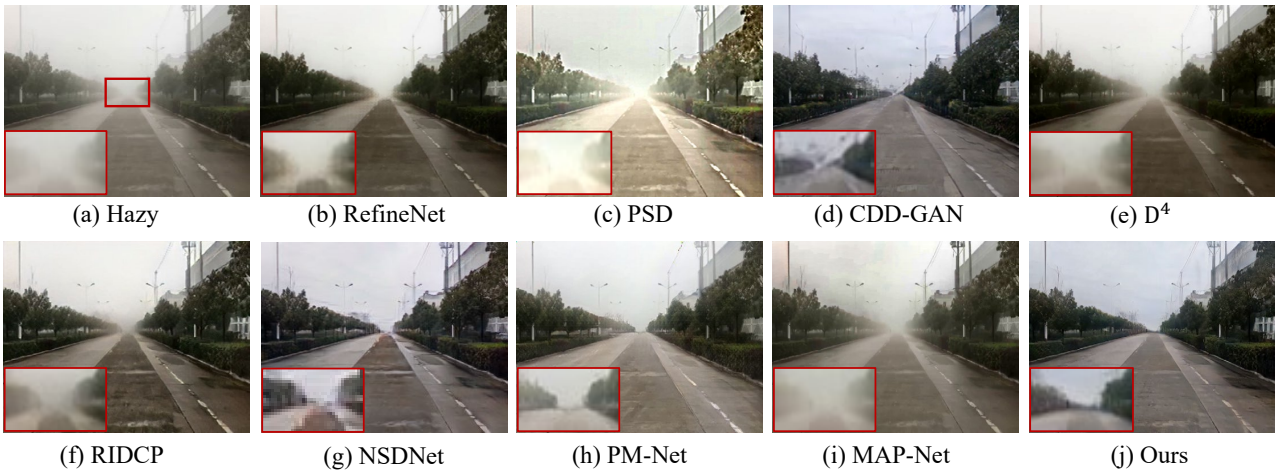
Limitations. In dense hazy scenarios, our method may exhibit slight artifacts in the sky region during dehazing. From the reported inference times in Table S1, it can be observed that our method still fails to meet real-time requirements.

Figure S5. An example of failure cases in sky region.

[1] https://replicate.com/casia-iva-lab/fastsam
[2] https://github.com/ultralytics/ultralytics

Figure S6. Comparsion of dehazing results on GoProHazy dataset.

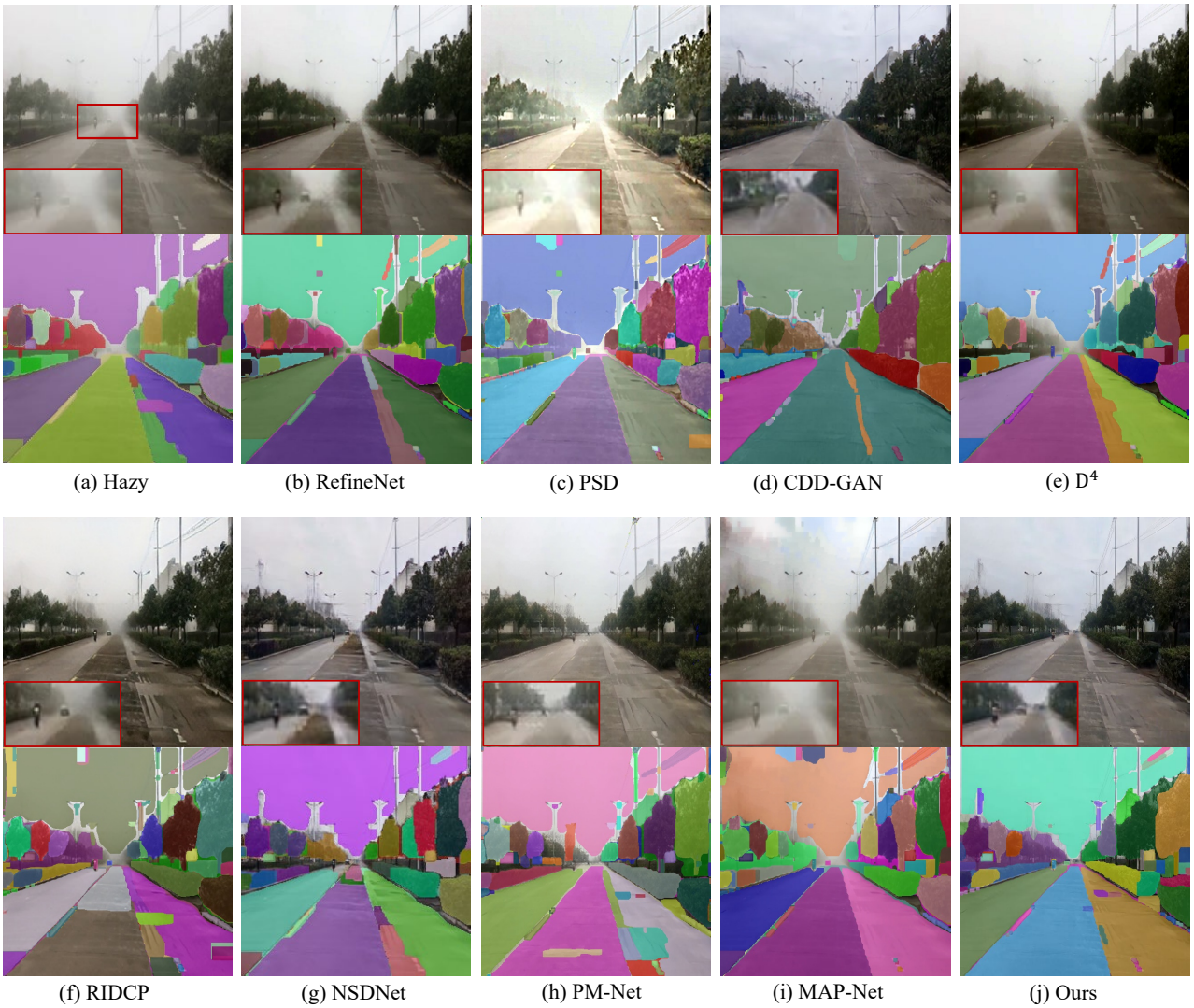
Figure S7. Comparsion of dehazing results on DrivingHazy dataset.



(a) Hazy (b) RefineNet (c) PSD (d) CDD-GAN (e) D^4

(f) RIDCP (g) NSDNet (h) PM-Net (i) MAP-Net (j) Ours

Figure S8. Comparsion of dehazing results on InternetHazy dataset.



(a) Hazy (b) RefineNet (c) PSD (d) CDD-GAN (e) D^4

(f) RIDCP (g) NSDNet (h) PM-Net (i) MAP-Net (j) Ours

Figure S9. Visual results of semantic segmentation on the InternetHazy dataset.