Supplementary material:

Video Dehazing via a Multi-Range Temporal Alignment Network with Physical Prior

Jiaqi Xu [1, 2, ⋆], Xiaowei Hu [2, ✉], Lei Zhu [3, 4, ✉], Qi Dou [1], Jifeng Dai [5, 2], Yu Qiao [2], Pheng-Ann Heng [1]
[1] The Chinese University of Hong Kong [2] Shanghai Artificial Intelligence Laboratory
[3] The Hong Kong University of Science and Technology (Guangzhou)
[4] The Hong Kong University of Science and Technology [5] Tsinghua University

⋆ : This work was done during Jiaqi Xu's internship at Shanghai Artificial Intelligence Laboratory.
✉ : Corresponding authors (huxiaowei@pjlab.org.cn; leizhu@ust.hk).

In this supplementary material, we provide more details about our HazeWorld dataset in Appendix A. Then we show network architecture details in Appendix B. In Appendix C, we present more experiment settings, comparisons to state-of-the-art (SOTA) methods, and ablation studies.

A. HazeWorld Dataset

Our HazeWorld dataset is a large-scale outdoor video dehazing dataset, and more dataset statistics are shown in Table 1. We select outdoor daytime videos from the original datasets [3, 6, 17, 19, 27, 31]. The frame rate and the resolution are mostly adjusted to keep each video with no more than 100 frames and around 720p, respectively. To obtain the transmission maps for synthesizing hazy videos based on the atmospheric scattering model, we use the geometric optimization-based robust video depth estimation method [12] to generate temporally consistent scale-less depth maps and manually convert them to the range of $[0m, 150m]$. Following [1, 23], four $\beta \in \{0.005, 0.01, 0.02, 0.03\}$ are adopted, which is corresponding to meteorological optical ranges of around $600m$, $300m$, $150m$, and $100m$, respectively. The example video frames with different β are shown in Fig. 1.

B. Network Architecture

The multi-scale feature maps are extracted from the encoder (*i.e.*, scales of 1/4, 1/8, 1/16, 1/32), which are then performed channel size reduction to a lower dimension using linear layers, *e.g.*, 64, for computational efficiency. Afterward, the feature maps with reduced channel size are fed into the prior and scene decoder separately.

In the overall U-Net architecture, the space-time deformable attention blocks (STDA) are used in multiple scene decoder layers. The input space-time flow $\tilde{\mathcal{O}}_{r\to i}$ of the first layer is initialized as zero, and it is refined gradually and trained in an end-to-end manner, which is similar to the coarse-to-fine optical flow estimation [21]. Note that the parameters in one STDA block are shared across multi-range alignment for computing flow offsets and other operations such as attention. But we record separate space-time flows $\mathcal{O}_{r\to i}$ for different ranges.

C. More Experimental Results

C.1. Settings

More implementation details. We use the AdamW optimizer and the polynomial scheduler with a power of 1.0. The learning rate is set as 2×10^{-4} with a warm-up start of 1,500 iterations. We use the pre-trained ConvNeXt-T [16] on ImageNet as the encoder. MPG is used in feature map scales of 1/16 and 1/32 for computational efficiency. We empirically set the number of prior tokens N as the number of frames i for each timestep i. On HazeWorld, the random crop and horizontal flip are adopted for data augmentation. The batch size is eight, and the patch size of input video frames is 256×256. During training, five hazy frames are used as inputs. The same training settings, *e.g.*, data augmentation and training iterations, are used for all compared methods. We use public implementations and retrain these models on our dataset to report their results. On REVIDE [13], we use heavier data augmentation, *e.g.*, resizing with random scales, since REVIDE only contains 42 training videos, following CG-IDN [29]. The patch size of input video frames is 384×384 as standard [9, 29].

Testing settings. We take the full videos as input for our method to restore the haze-free video frames.

C.2. More comparisons with SOTA methods

More quantitative comparison. Our method outperforms SOTA image dehazing, video dehazing, and video restoration methods on HazeWorld (see Table. 1 in the paper). Moreover, we compare our method against two SOTA image dehazing methods (*i.e.*, AECR [28] and Dehamer [7]) with the video alignment method BasicVSR++ [2]. Their PSNR/SSIM results are 23.47/0.9111 and 23.89/0.9220, respectively, which indicate that our method outperforms SOTA image dehazing with video alignment methods.

Note that since the codes of CG-IDN [29] and NCFL [9] are not officially released at the time of this work, we implement these two methods to obtain their quantitative results on our HazeWorld dataset.

More qualitative comparison. We show more visual comparisons of HazeWorld and REVIDE, as shown in Fig. 2 - Fig. 7 and Fig. 8, respectively. The examples demonstrate that our method can better remove the haze and generate visually appealing results.

C.3. Applications

We choose four different methods for downstream application validation, *i.e.*, VPSNet [11] for video panoptic segmentation on Cityscapes [3], PSPNet [30] for image semantic segmentation on Cityscapes [3], PackNet-SfM [6] for monocular depth estimation on DDAD [6], and STM [18] for video object segmentation on DAVIS [19]. We use the models trained on the original clear videos and obtain results on the input hazy videos, the dehazed videos, and the underlying haze-free videos. The β for hazy videos is 0.02 for video panoptic segmentation, depth estimation, and image semantic segmentation, and β is 0.03 for video object segmentation. We select 6 videos for video panoptic segmentation, 180 for image semantic segmentation, 50 for monocular depth estimation, and 25 for video object segmentation. The visual comparisons of these applications are shown in Fig. 9 - Fig. 12.

Table 1. **Statistics of our HazeWorld dataset.** Note that we use four different β to synthesize the videos, resulting in four times the number of hazy videos and frames in HazeWorld than shown in this table. The resolution indicates the shorter side of the video.

Source Dataset	Scenario	#Videos	#Train/Test	#Frames	Diversity	Downstream	Resolution
Cityscapes [3]	Driving	540	360/180	16,200	Medium	Segmentation	720
DDAD [6]	Driving	200	150/50	16,600	Medium	Depth	720
UA-DETRAC [27]	Surveillance	83	53/30	8,235	Low	Detection	540
VisDrone [31]	Drone	70	52/18	6,121	Medium	Tracking	540, 720
DAVIS [19]	Generic	117	51/66	8,198	High	Segmentation	720
REDS [17]	Life	261	231/30	26,100	Medium	-	720

Table 2. **Comparison of model size, FLOPs, and runtime.**

Method	Params (M)	FLOPs (G)	Runtime (ms)	PSNR (dB)
MSBDN [5]	31.35	24.53	145	23.70
Dehamer [7]	132.4	48.26	202	22.92
DehazeFormer [24]	4.6	47.32	427	25.44
EDVR [26]	20.9	31.98	408	22.91
BasicVSR++ [2]	7.4	28.10	96	26.06
Our method	28.8	8.21	101	27.12

Table 3. **Comparison of temporal stability.** The relative standard deviation (RSD) of PSNR is reported in percentage and warping error (E_{warp}) is reported in the scale of $\times 10^{-3}$.

Method	MSBDN [5]	Dehamer [7]	EDVR [26]	BasicVSR++ [2]	Ours
E_{warp} ↓	1.38	1.44	1.53	1.26	**1.17**
RSD ↓	3.62	4.37	3.81	4.20	**3.33**
PSNR ↑	23.70	22.92	22.91	26.06	**27.12**

Table 4. **Ablation studies of the memory in MPG.**

(a) Discussion on the memory usage for the prior feature enhancement.

Memory		✓
PSNR	26.79	**27.12**

(b) Discussion on the number of transmission categories in MPG.

#Category	w/o	16	32 (Ours)	64
PSNR	26.87	27.01	**27.12**	26.96

C.4. More Analysis of Our Method

Model Efficiency. We compare the number of parameters (denoted as Params), FLOPs, and running time of our network and state-of-the-art methods on a TITAN RTX GPU. FLOPs are calculated at the input size of 256×256, and the running time per frame is tested at the spatial dimension of 1,280×720, respectively, and the sequence length is 5. As shown in Table 2, our method obtains the best PSNR value with reasonable FLOPs and running time, which indicates that MAP-Net performs better with a fast inference.

Temporal stability. Following [13], we utilize the flow warping error (E_{warp}) to measure the temporal consistency of two consecutive frames quantitatively. We also use the relative standard deviation (RSD) of PSNR to measure the temporal coherence of dehazed videos. The optical flows used for computing E_{warp} are obtained by FlowNet2 [10] on the haze-free videos. Table 3 shows that our network has the smallest E_{warp} value and the smallest RSD value among all methods. It indicates that the dehazed video frames of our network are clearer and more temporally stable than compared methods.

Comparisons using different training data. We train our network using REVIDE [29] and HazeWorld separately and test them on real outdoor hazy videos. As shown in Fig. 13, we can observe that the model trained on REVIDE may retain the haze, introduce color distortion, and produce artifacts. On the contrary, the model trained on HazeWorld can produce visually appealing and naturally dehazed results, which shows the advantage of our HazeWorld.

C.5. More Ablation Studies

Discussion on two major modules.

We provide qualitative results to understand the effectiveness of our proposed memory-based physical prior guidance (MPG) module and multi-range scene radiance recovery (MSR) module. As shown in Fig. 14, since MPG provides haze prior information as guidance, *e.g.*, transmission, "Basic+MPG" is able to produce a more uniform color for regions with similar transmission values, such as the road. On the other hand, "Basic+MSR" can capture the scene and haze clues from multiple adjacent frames, which restores more fine details (see cropped regions). By combining these two complementary contributions, our method clearly removes the haze and better recovers the scene structures.

Discussion on our MPG module. We first discuss the accuracy of the estimated physical prior-related components (*i.e.*, transmission and atmospheric light) to understand the learned prior feature. Then, we discuss the effectiveness of memory in our MPG. Note that we do not use ground truth transmission t or atmospheric light A for supervision during training. Since the ground truths of t and A are not always available, we employ a reconstruction loss on the reconstructed input image from estimated t and A to guarantee the estimation of t and A, *i.e.*, physical haze prior. This setting can apply to datasets without transmission or atmospheric light labels, *e.g.*, REVIDE. Since we have the ground truths of t and A for the testing set of our dataset, we can compute the quantitative mean absolute errors (MAE)

Table 5. **Discussion on the alignment layer and flow loss when the number of ranges is one.**

Layer	WA	DCN	DWA	DWA+$\mathcal{L}_{flow}$ (Ours)
PSNR	25.83	24.11	25.75	**26.24**

for the transmission and atmospheric light estimation, and the MAE values are 0.0419 and 0.0284, respectively. With such t and A estimation errors, our experimental results on real videos and benchmarks demonstrate that our method outperforms state-of-the-art methods. Further, as shown in Fig. 15, the reconstructed hazy image and the estimated transmission are very close to the ground truth, indicating that our model can recover the haze priors. Hence, the features encoded with prior information from MPG provide haze clue guidance on scene recovery.

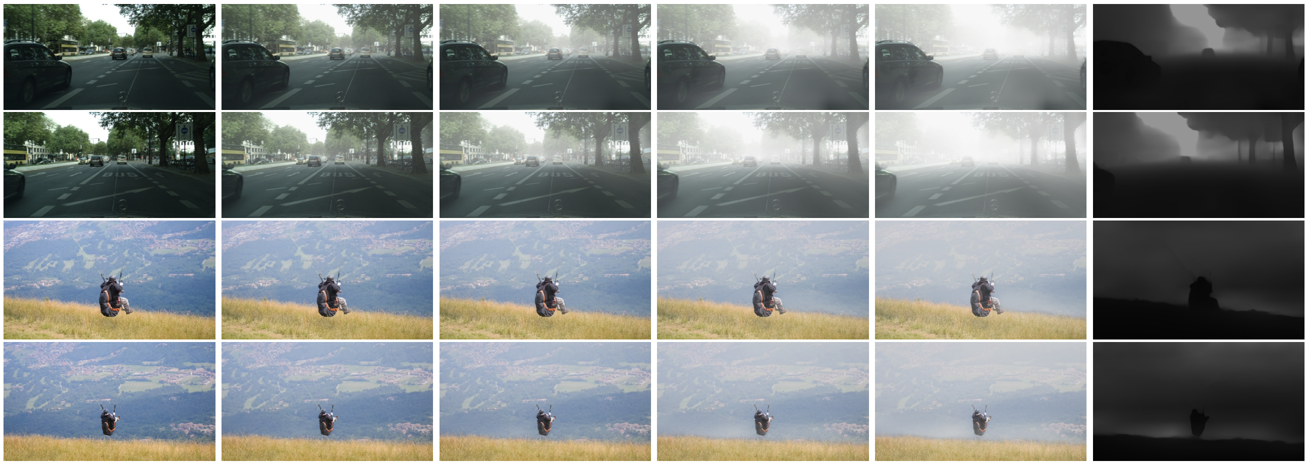
We then perform ablation study experiments on memory in MPG. Table 4 reports the quantitative results for the memory enhancement and the number of transmission categories for prior feature compression. Our method with memory brings a performance gain of 0.33 dB in PSNR compared to the baseline without memory for the prior feature enhancement due to the extracted long-range haze information in video sequences; see Table 4a. Moreover, we study the number of transmission categories used for prior feature compression. We also compare our method with the simple memory implementation without compression (denoted as w/o), which directly saves the flattened feature map into the memory. Only the features of scale 1/32 are used for memory in this baseline due to the expensive memory space requirement, especially for high-resolution videos. As shown in Table 4b, our compression strategy outperforms the baseline since more compact historical haze information is encoded in the memory. Besides, we observe that the network is not sensitive to the number of transmission categories. Hence, we empirically set the number of transmission categories as 32 in our implementation.

Discussion on our MSR module. To understand their effectiveness, we first visualize the alignment process in the space-time deformable attention block (STDA) and the aggregation weight maps in the guided multi-range aggregation block (GMRA). Then, we briefly discuss the alignment layers and flow loss.

Fig. 16 visualizes the warped images to showcase the alignment quality and activation maps to show the regions of interest. Some regions cannot be well aligned using the frame-by-frame alignment due to the occlusion (see the heads of women), which only considers one adjacent frame. In contrast, the learned space-time flow captures the correspondence from multiple frames for feature alignment, as illustrated by the warped images. Meanwhile, from the activation maps, the query in the target frame leverages temporal information from multiple adjacent frames but only focuses on corresponding regions. Hence, the multi-range alignment captures the temporal scene and haze clues from multiple space-time resolutions for scene recovery.

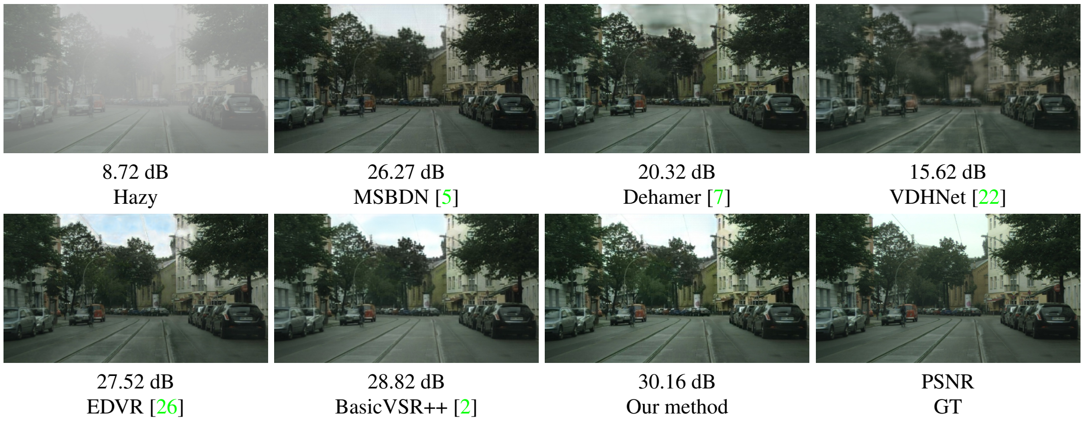
Fig. 17 shows the warping error maps and aggregation weight maps from different ranges and aggregation perspectives. The aggregation weight maps and warping error maps from multiple ranges together demonstrate that the model pays attention to different range features by considering the alignment quality. Moreover, the prior guidance provides haze clues on the aggregation process, as indicated by the aggregation weight maps concerning the transmission values from the prior and the scene perspectives.

We further conduct ablation experiments on the alignment layer and flow loss. As shown in Table 5, the local window attention (WA) [15] produces reasonable results (25.83 dB) because of its relatively large receptive field but is not able to capture correspondence for objects with noticeable misalignment in dynamic scenes (see examples in Fig. 16). The DCN layer [4] for temporal alignment suffers from unstable training with poor performance (24.11 dB), which is also observed in [2]. Lastly, our deformable attention with learned space-time flows trained using flow loss captures the scene and haze correspondence, which benefits from aligned pixel-wise temporal information and achieves better dehazing performance (26.24 dB).



(a) GT frames (b) $\beta = 0.005$ (c) $\beta = 0.01$ (d) $\beta = 0.02$ (e) $\beta = 0.03$ (f) Depth

Figure 1. Example video frames in our HazeWorld dataset with different β.



8.72 dB
Hazy

26.27 dB
MSBDN [5]

20.32 dB
Dehamer [7]

15.62 dB
VDHNet [22]

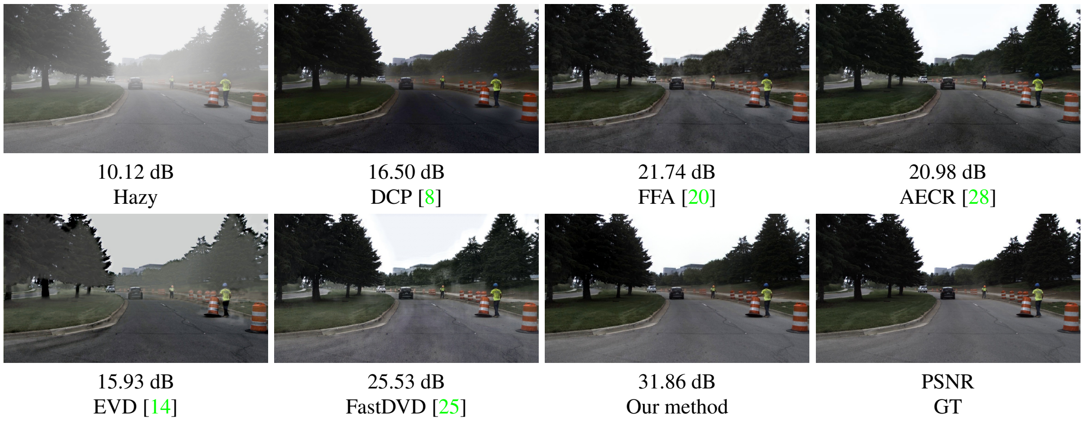
27.52 dB
EDVR [26]

28.82 dB
BasicVSR++ [2]

30.16 dB
Our method

PSNR
GT

Figure 2. Visual comparison of video dehazing results on HazeWorld #1.



10.12 dB
Hazy

16.50 dB
DCP [8]

21.74 dB
FFA [20]

20.98 dB
AECR [28]

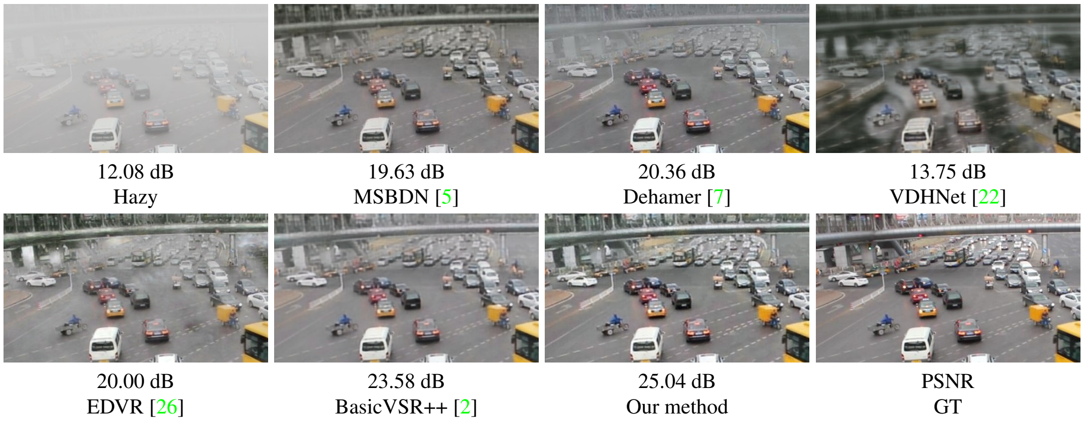
15.93 dB
EVD [14]

25.53 dB
FastDVD [25]

31.86 dB
Our method

PSNR
GT

Figure 3. Visual comparison of video dehazing results on HazeWorld #2.



12.08 dB
Hazy

19.63 dB
MSBDN [5]

20.36 dB
Dehamer [7]

13.75 dB
VDHNet [22]

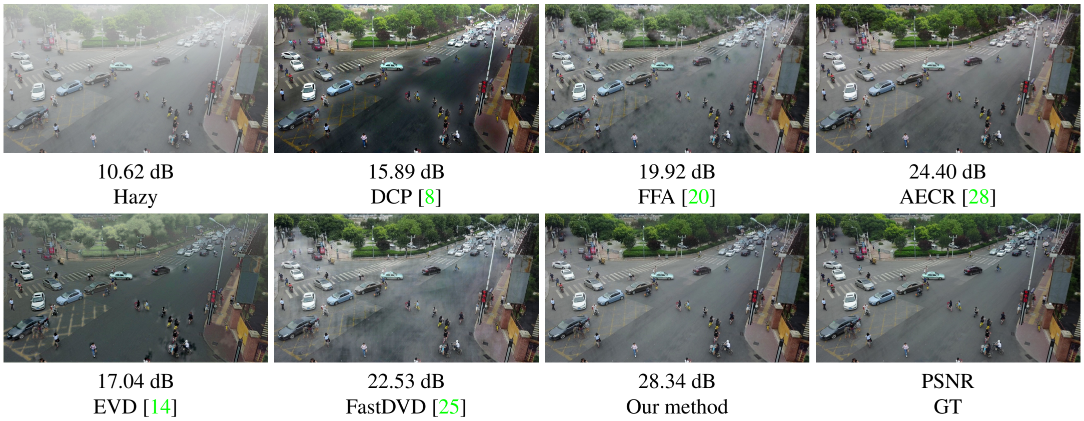
20.00 dB
EDVR [26]

23.58 dB
BasicVSR++ [2]

25.04 dB
Our method

PSNR
GT

Figure 4. Visual comparison of video dehazing results on HazeWorld #3.



10.62 dB Hazy

15.89 dB DCP [8]

19.92 dB FFA [20]

24.40 dB AECR [28]

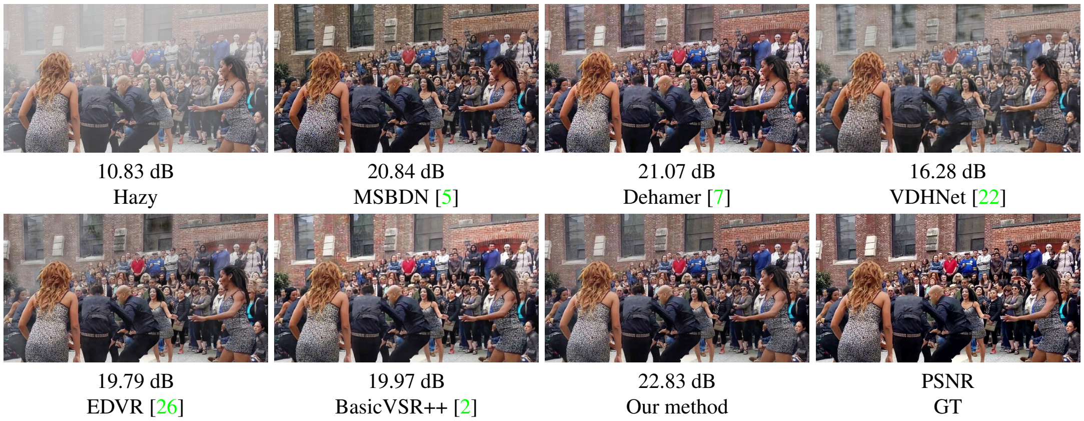
17.04 dB EVD [14]

22.53 dB FastDVD [25]

28.34 dB Our method

PSNR GT

Figure 5. Visual comparison of video dehazing results on HazeWorld #4.



10.83 dB Hazy

20.84 dB MSBDN [5]

21.07 dB Dehamer [7]

16.28 dB VDHNet [22]

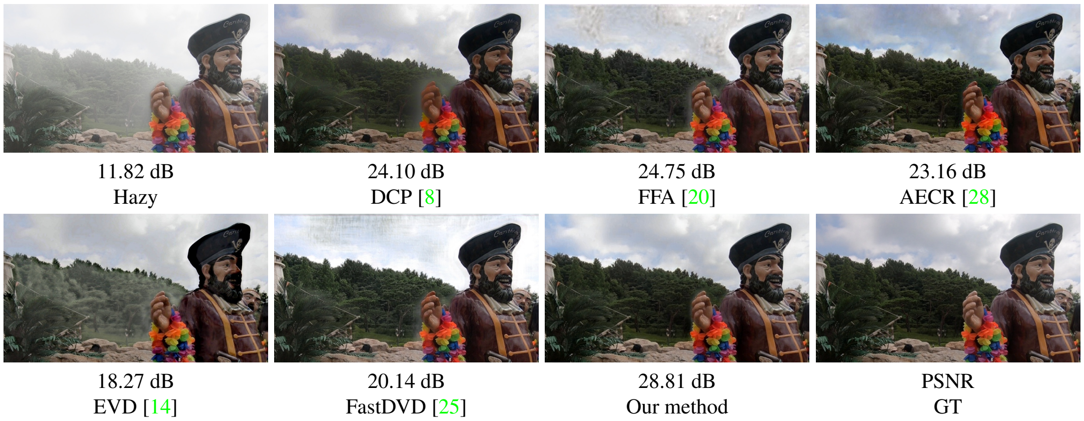
19.79 dB EDVR [26]

19.97 dB BasicVSR++ [2]

22.83 dB Our method

PSNR GT

Figure 6. Visual comparison of video dehazing results on HazeWorld #5.



11.82 dB Hazy

24.10 dB DCP [8]

24.75 dB FFA [20]

23.16 dB AECR [28]

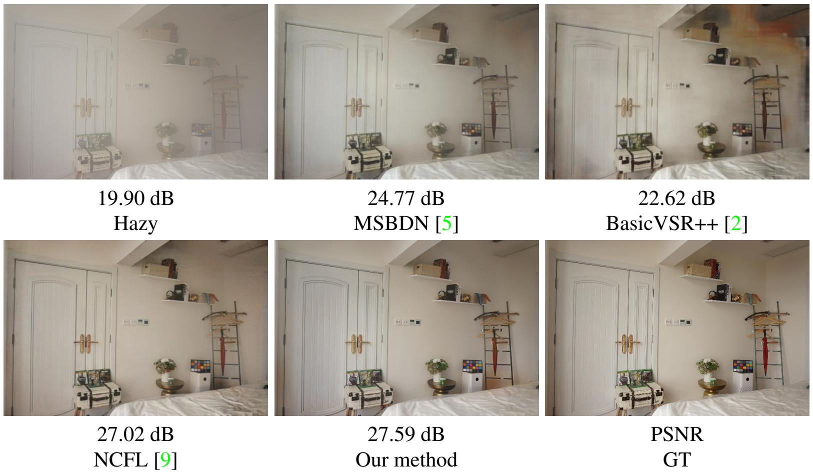
18.27 dB EVD [14]

20.14 dB FastDVD [25]

28.81 dB Our method

PSNR GT

Figure 7. Visual comparison of video dehazing results on HazeWorld #6.



19.90 dB
Hazy

24.77 dB
MSBDN [5]

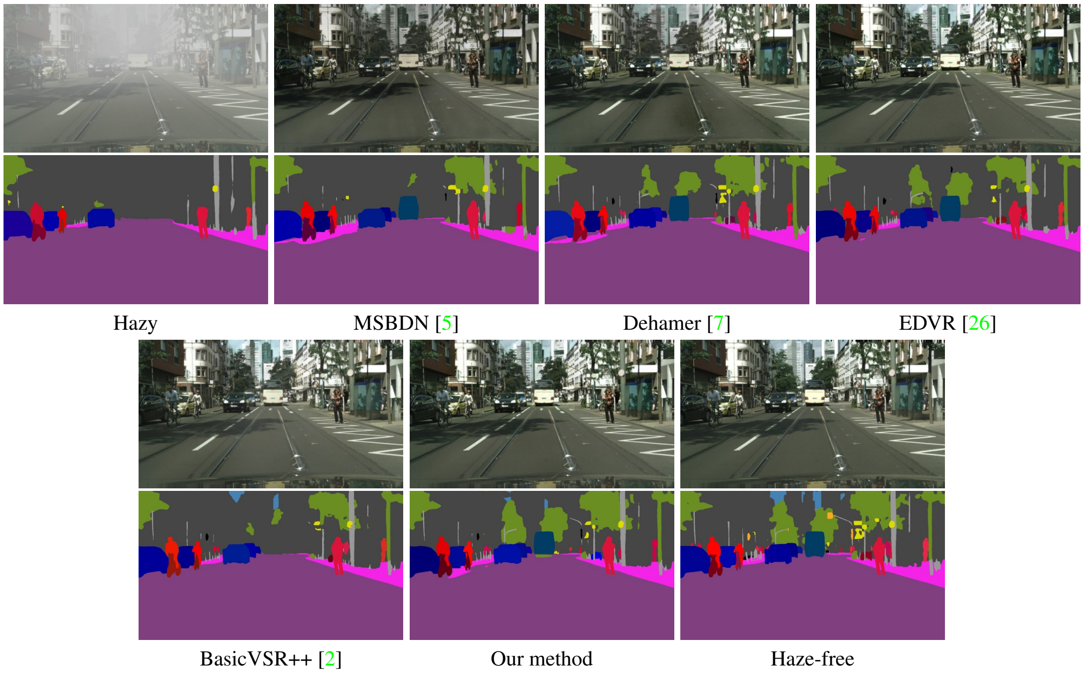
22.62 dB
BasicVSR++ [2]

27.02 dB
NCFL [9]

27.59 dB
Our method

PSNR
GT

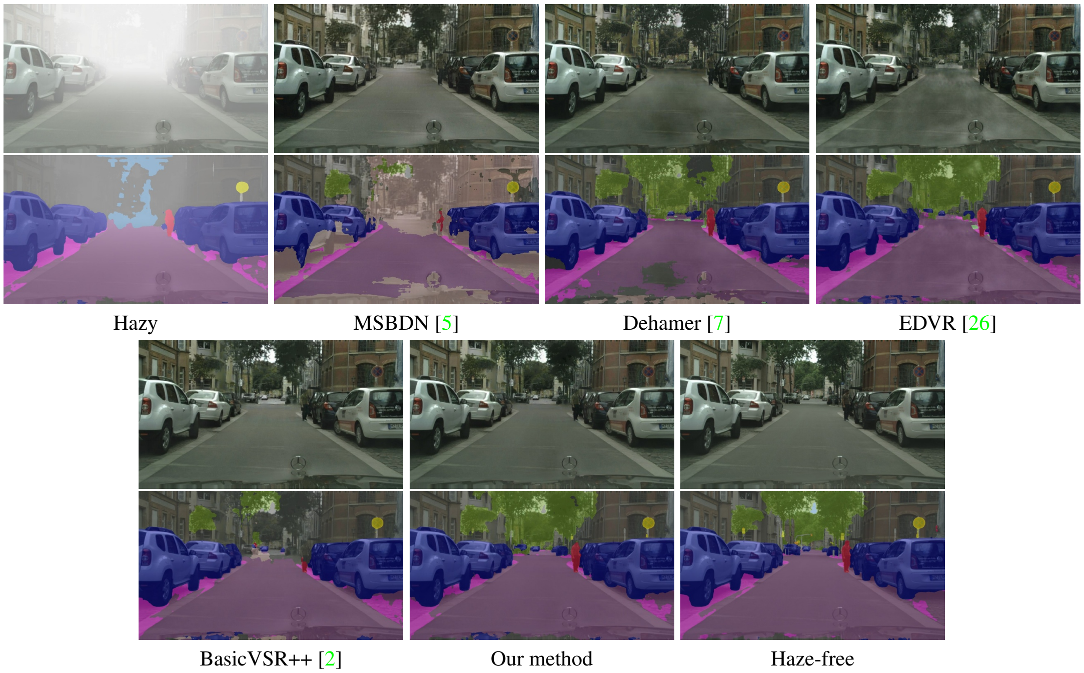
Figure 8. Visual results of *W002* video from REVIDE [29].



Hazy MSBDN [5] Dehamer [7] EDVR [26]

BasicVSR++ [2] Our method Haze-free

Figure 9. Visual results of video panoptic segmentation on HazeWorld. We show the hazy/dehazed/haze-free frames (the first row) and the corresponding results (the second row).



Hazy MSBDN [5] Dehamer [7] EDVR [26]

BasicVSR++ [2] Our method Haze-free

Figure 10. Visual results of image semantic segmentation on HazeWorld. We show the hazy/dehazed/haze-free frames (the first row) and the corresponding results (the second row).

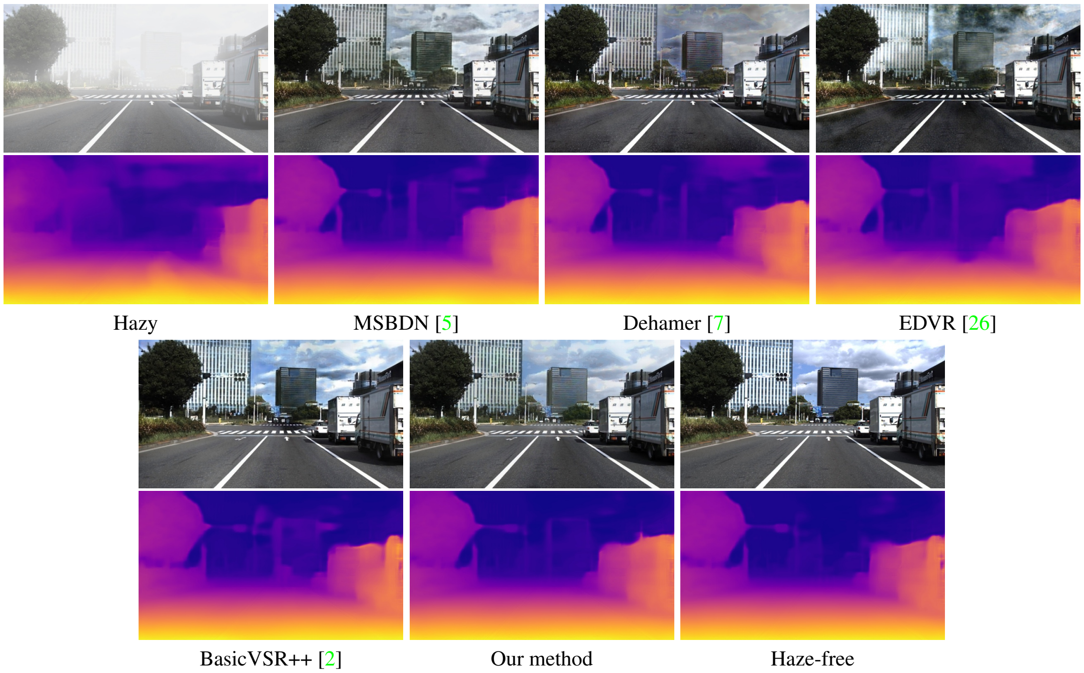

Hazy
MSBDN [5]
Dehamer [7]
EDVR [26]
BasicVSR++ [2]
Our method
Haze-free

Figure 11. Visual results of monocular depth estimation on HazeWorld. We show the hazy/dehazed/haze-free frames (the first row) and the corresponding results (the second row).

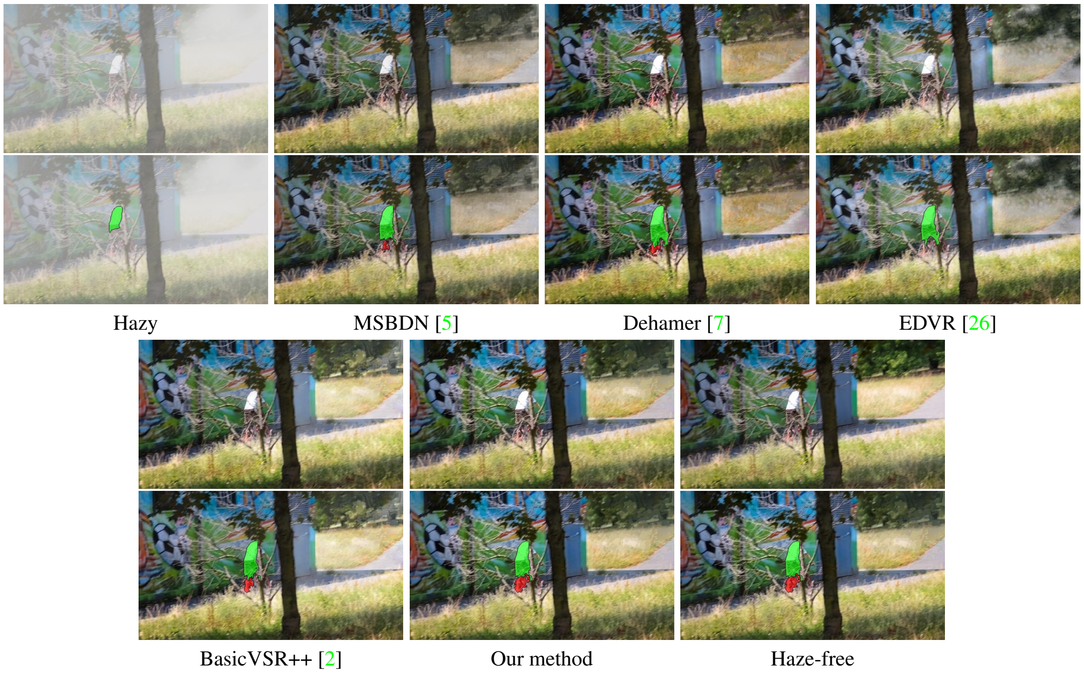

Hazy
MSBDN [5]
Dehamer [7]
EDVR [26]
BasicVSR++ [2]
Our method
Haze-free

Figure 12. Visual results of video object segmentation on HazeWorld. We show the hazy/dehazed/haze-free frames (the first row) and the corresponding results (the second row).

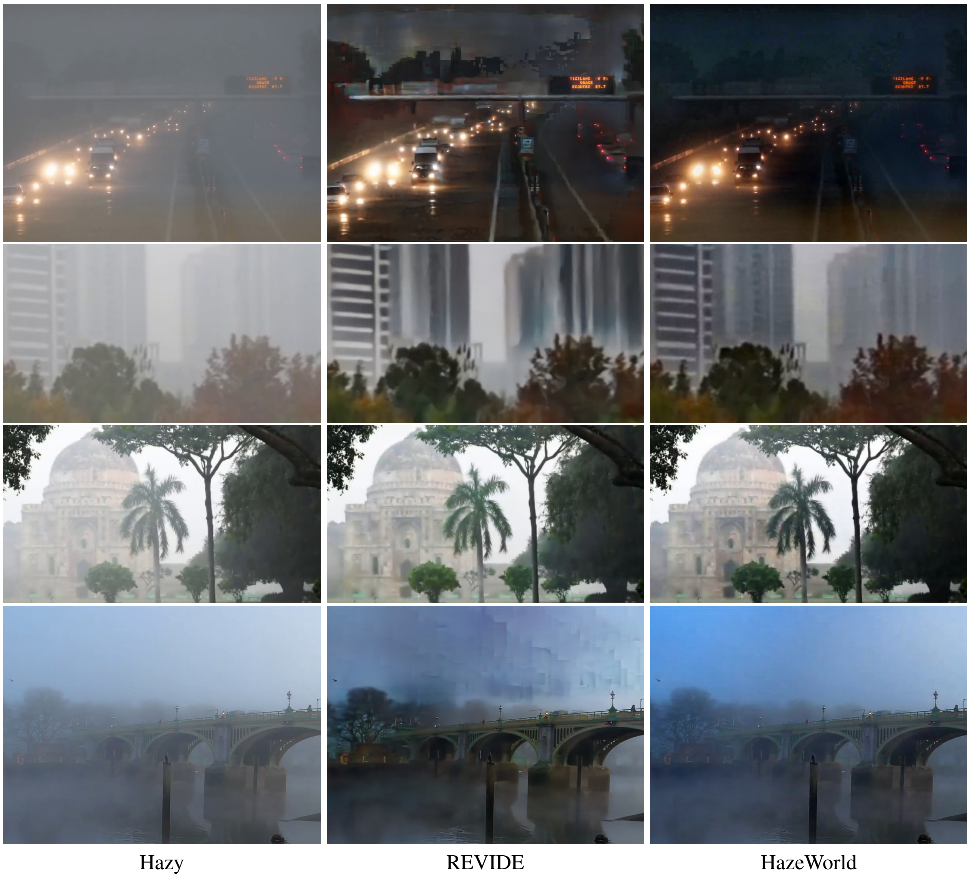
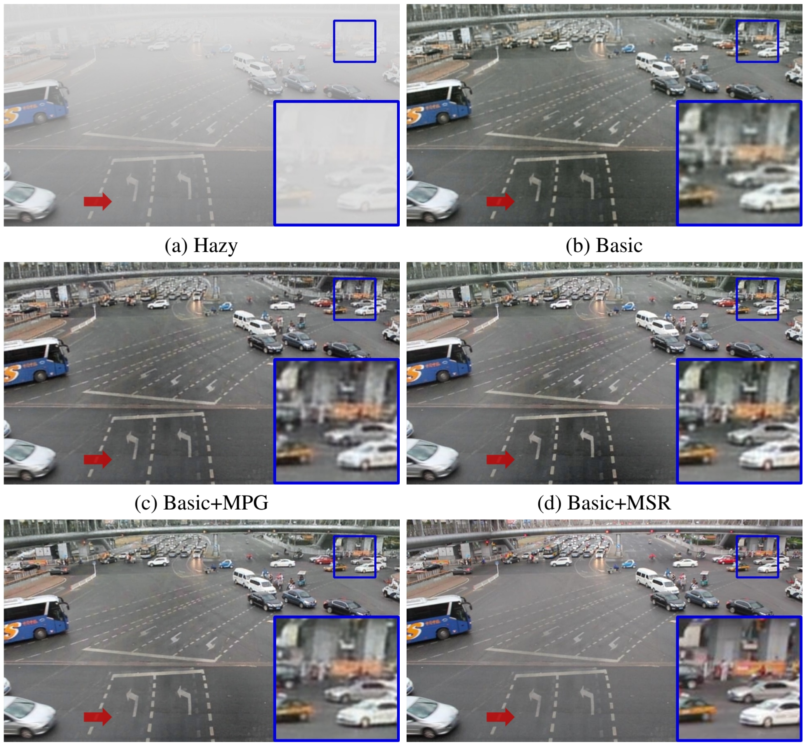


Figure 13. Visual comparison results on real hazy videos. The results are generated by our method trained using REVIDE and HazeWorld. The model trained on REVIDE tends to retain some haze, introduce color distortion, and produce artifacts.


(a) Hazy

(b) Basic

(c) Basic+MPG

(d) Basic+MSR

(e) Our method

(f) GT

Figure 14. Visual results on the effectiveness of the memory prior guidance (MPG) module and the multi-range scene recovery (MSR) module. (**c**) The result of “Basic+MPG” loses fine details because of the lack of temporal information from multiple adjacent frames (see the cropped regions). (**d**) The result of “Basic+MSR” suffers from non-uniform colors since no prior haze clues are provided, as indicated by the red arrows. (**e**) Our method with MPG and MSR generates a clearer and more visually appealing dehazed result, benefiting from two complementary contributions.

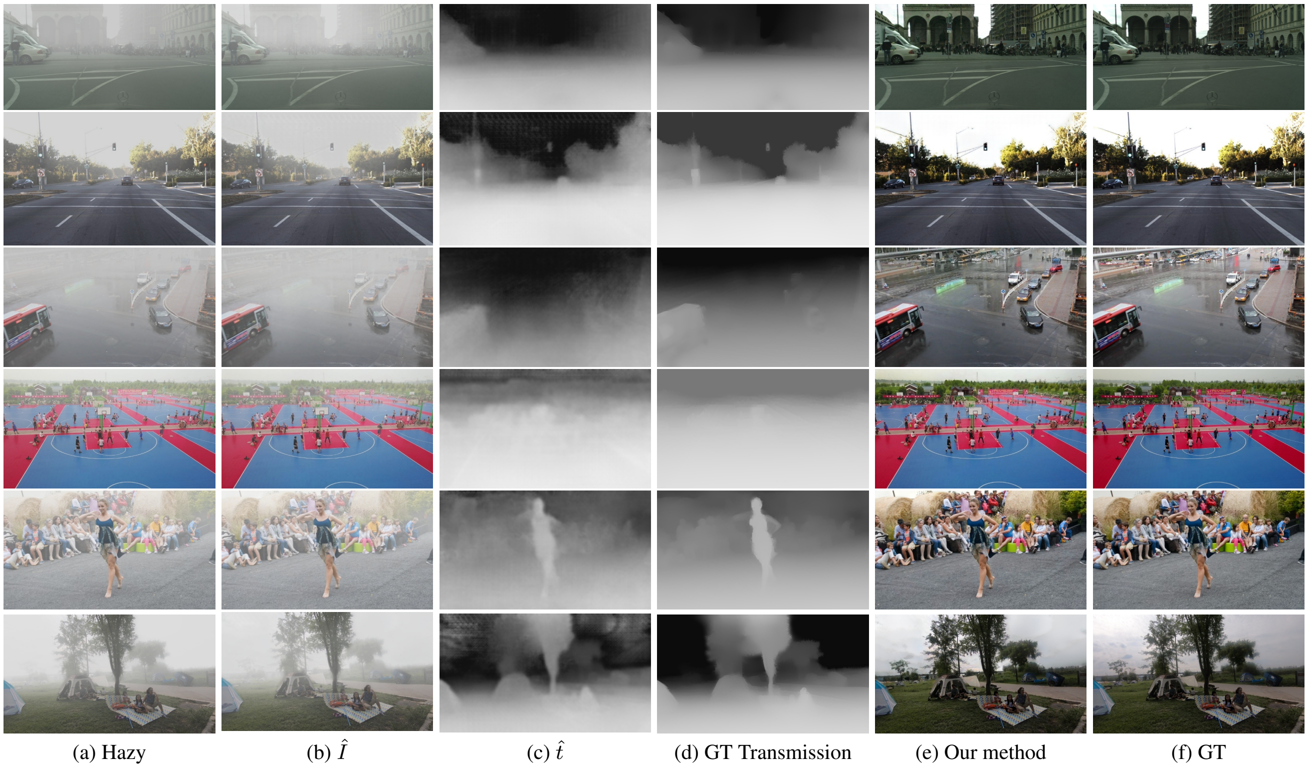
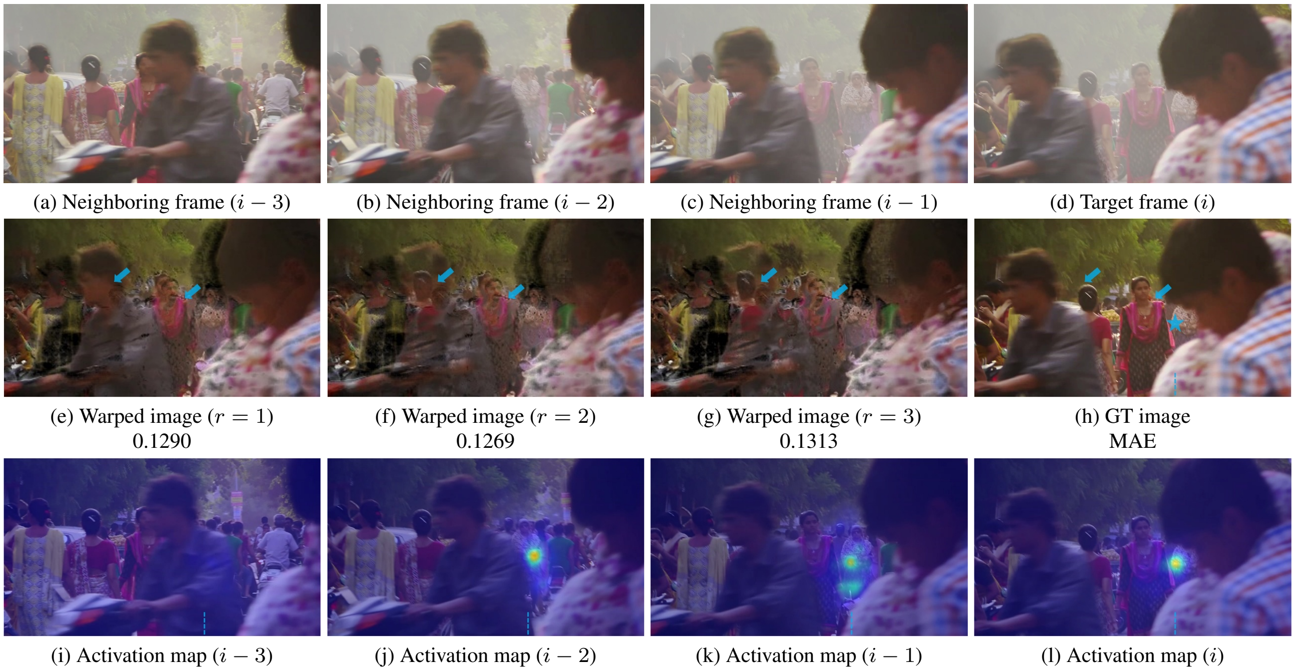

(a) Hazy (b) $\hat{I}$ (c) $\hat{t}$ (d) GT Transmission (e) Our method (f) GT

Figure 15. Visualization of intermediate component predictions by our method on HazeWorld. $\hat{I}, \hat{t}$ are the reconstructed hazy image and estimated transmission map, respectively.

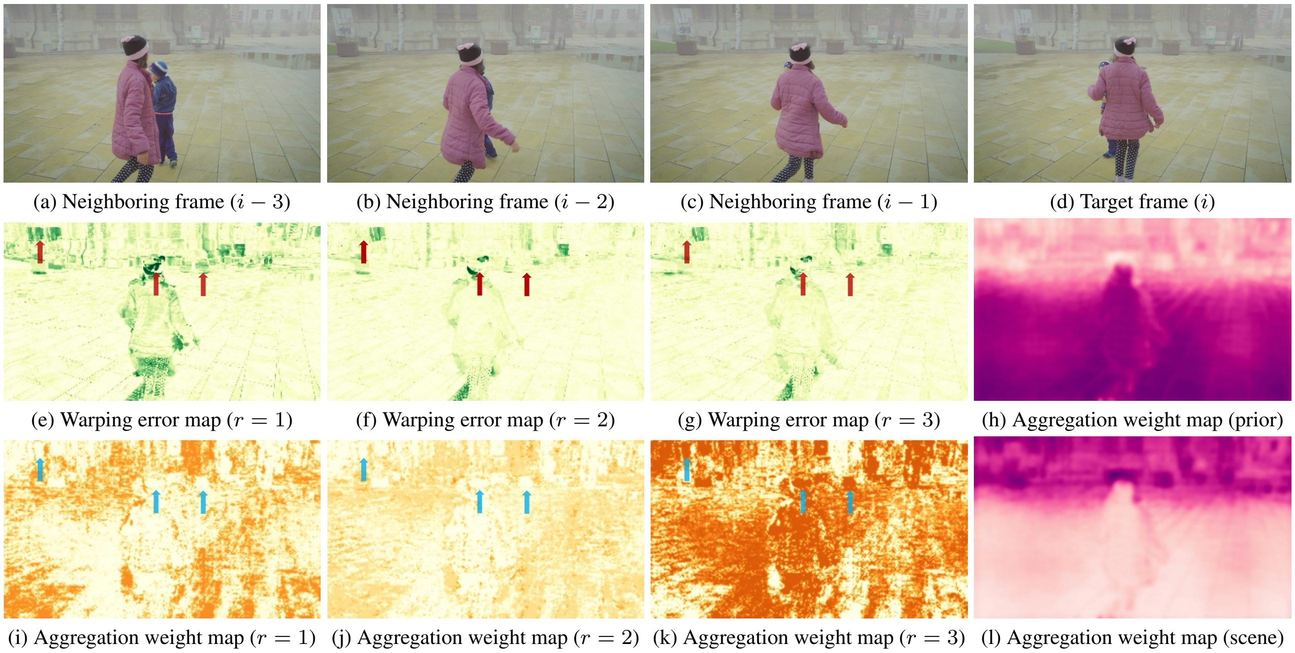


(a) Neighboring frame $(i-3)$ (b) Neighboring frame $(i-2)$ (c) Neighboring frame $(i-1)$ (d) Target frame (i)

(e) Warped image $(r=1)$ 0.1290 (f) Warped image $(r=2)$ 0.1269 (g) Warped image $(r=3)$ 0.1313 (h) GT image MAE

(i) Activation map $(i-3)$ (j) Activation map $(i-2)$ (k) Activation map $(i-1)$ (l) Activation map (i)

Figure 16. **Analysis of multi-range alignment.** We visualize the warped images to showcase the alignment quality and activation maps to show the regions of interest. (**e-g**) The warped images are obtained by the space-time sampling operation, which takes the ground truth neighboring frames, and the learned space-time flows from our model as the inputs. Note that some regions cannot be well aligned if only considering one adjacent frame due to occlusion, *e.g.*, women's heads, as indicated by blue arrows. MAE stands for mean absolute error, which measures the deviations between the warped image and the ground truth image (the lower, the better). (**i-l**) The activation maps are visualized for different adjacent frames given the query region denoted by a blue star in the target frame (**h**).



(a) Neighboring frame $(i-3)$ (b) Neighboring frame $(i-2)$ (c) Neighboring frame $(i-1)$ (d) Target frame (i)

(e) Warping error map $(r=1)$ (f) Warping error map $(r=2)$ (g) Warping error map $(r=3)$ (h) Aggregation weight map (prior)

(i) Aggregation weight map $(r=1)$ (j) Aggregation weight map $(r=2)$ (k) Aggregation weight map $(r=3)$ (l) Aggregation weight map (scene)

Figure 17. **Analysis of guided multi-range aggregation.** We visualize the warping error and aggregation weight maps. (**e-g**) The warping error maps are computed between the warped images and the ground truth image. (**i-k**) The model pays attention to different range features considering the alignment qualities, as indicated by the red/blue arrows. (**h, l**) The aggregation weights from the prior and the scene perspectives are assigned to different regions concerning the transmission. The darker colors denote the larger error/weight values.

References

[1] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *CVPR*, 2020. 2

[2] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, 2022. 2, 3, 4, 6, 7, 8, 9, 10

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 3

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 4

[5] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *CVPR*, 2020. 3, 6, 7, 8, 9, 10

[6] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020. 2, 3

[7] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *CVPR*, 2022. 2, 3, 6, 7, 9, 10

[8] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *TPAMI*, 2010. 6, 7

[9] Cong Huang, Jiahao Li, Bin Li, Dong Liu, and Yan Lu. Neural compression-based feature learning for video restoration. In *CVPR*, 2022. 2, 8

[10] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. 3

[11] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *CVPR*, 2020. 2

[12] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *CVPR*, 2021. 2

[13] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018. 2, 3

[14] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. End-to-end united video dehazing and detection. In *AAAI*, 2018. 6, 7

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 4

[16] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 2

[17] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPRW*, 2019. 2, 3

[18] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 2

[19] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 2, 3

[20] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *AAAI*, 2020. 6, 7

[21] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, 2017. 2

[22] Wenqi Ren, Jingang Zhang, Xiangyu Xu, Lin Ma, Xiaochun Cao, Gaofeng Meng, and Wei Liu. Deep video dehazing with semantic segmentation. *TIP*, 2018. 6, 7

[23] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 2018. 2

[24] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *TIP*, 2023. 3

[25] Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *CVPR*, 2020. 6, 7

[26] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019. 3, 6, 7, 9, 10

[27] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 2020. 2, 3

[28] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *CVPR*, 2021. 2, 6, 7

[29] Xinyi Zhang, Hang Dong, Jinshan Pan, Chao Zhu, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Fei Wang. Learning to restore hazy video: A new real-world dataset and a new method. In *CVPR*, 2021. 2, 3, 8

[30] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2

[31] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *TPAMI*, 2021. 2, 3