Learning to Restore Hazy Video: A New Real-World Dataset and A New Method Supplementary Material

Xinyi Zhang[1*] Hang Dong[2,3*†] Jinshan Pan[4] Chao Zhu[3] Ying Tai[1] Chengjie Wang[1]
Jilin Li[1] Feiyue Huang[1] Fei Wang[3]
[1] Tencent Youtu Lab [2] ByteDance Intelligent Creation Lab
[3] College of Artificial Intelligence, Xi'an Jiaotong University
[4] Nanjing University of Science and Technology

Overview

In this supplemental material, we provide more details about scene layout and optical parameters of the camera in Sec. A. Then, the details of the Confidence Guided Pre-Dehazing (CGPD) and restoration modules are provided in Sec. B. To better illustrate the effectiveness of the Improved Deformable Alignment (IDA) module, we show the partial cost volume can help the deformable module to estimate the flow-like offset in Sec. C. Some typical frames of the REVIDE and REVIDE-SYN datasets are presented in Sec. D to further demonstrate the high fidelity of the collected haze and the drawbacks of the synthetic haze. Finally, more quantitative and qualitative comparisons on the REVIDE dataset and real-world videos are provided in Sec. E.

A. More Details of Scene Layout and Camera Settings

Scene layout. As mentioned in the manuscript, the furnishing styles of scenes are diverse which can be grouped into 4 categories: the eastern style, western style, modern style, and laboratory style. In addition, the types of the selected room are also manifold, e.g office, study room, small library, living room, dining room, kitchen, bedroom, veranda, etc. To further enrich these scenes, colorful objects are also arranged to these scenes.

According to the layout of the collected scenes, we choose one trajectory with suitable rotational direction from eight pre-designed trajectories. For each scene, we also add some positional disturbance to the chosen trajectory to enrich the movement of the camera.

Optical parameters of camera. Before the acquisition, we also carefully adjust the optical parameters of the camera for obtaining high-quality frames. The focal length of the camera is set in a range of 2 m to 5 m and the exposure time is fixed at 0.025s to avoid the out-of-focus blurs and motion blurs respectively. To get a suitable exposure value (EV=0) and alleviate camera noise, small ISO and F-number are preferred. Finally, a standard gray card is used to correct the white balance of the camera, which can minimize the chromatism of the collected frames.

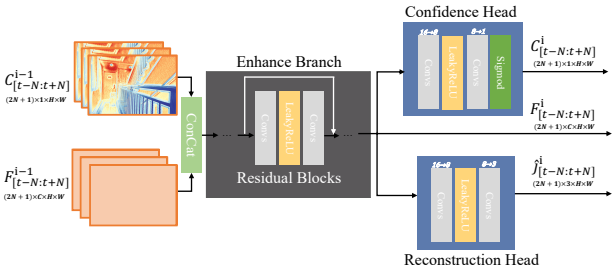


Figure A. **Detials of the Confidence Guided Pre-Dehazing (CGPD) module.**

B. More Implementation Details

As shown in Fig. A, the enhance branch in the CGPD module is built with three residual blocks [12]. Both the confidence and dehazing heads consist of two convolutional layers, except that the confidence head ends with a sigmoid activation layer.

The reconstruction module in the proposed method is built on the MSBDN [5] with some modifications. To reduce the parameters, we remove the DFF modules and change the numbers of residual blocks in the decoder and encoder modules to 2. Finally, the input channel of the first convolutional layer is set to 16 according to the fused features F_t^{Fused} from the Multi-Feature Fusion (MFF) module.

C. More Analysis on IDA Module

Due to the large displacement and changeful haze among the neighboring frames in the REVIDE dataset, aligning the features from the neighboring frames is more difficult than other video datasets. To obtain more robust aligned features, an Improved Deformable Alignment (IDA) mod-

*These authors contributed equally to this work.
†Corresponding author.

ule is proposed in the manuscript by introducing the partial cost volume [11] to the PCD module [12]. To prove that the partial cost volume can boost the performance of the deformable alignment, we re-train the proposed method with the IDA and PCD modules on the REVIDE dataset respectively. Then, We plot the mean values of the learned offsets for all the deformable convolutions (DCNs) in the IDA and PCD modules during the training process. As shown in Fig. B, the mean value of the pixel shift among the training frames (the red dotted line) is about 72 *. However, the learned offsets of all the four DCNs in the PCD module are less than 1 pixel (the dashed line), which is far less than the pixel shift. On the contrary, the learned offsets of the IDA module (the solid line) are more close to the optical flows in a pyramid structure: the first three learned offsets in the pyramid structure (IDA_Ln) are ascending as the spatial sizes of the offsets increase and the mean values of the first level (IDA_L1) are close to the pixel shift. It is also noted that the final learned offsets (IDA_Cas) are relatively low since the cascading DCN aims to refine the coarsely aligned features to the sub-pixel accuracy. Therefore, the experiments show that the introduction of the partial cost volume can help the deformable module obtain the flow-like offsets, and thus improving the alignment.

In addition, we also apply the proposed IDA module to the EDVR [12] and find that it boosts the performance of EDVR on the REVIDE by a margin of 0.48 dB †, which demonstrates that the proposed IDA module can be generalized to other architectures with deformable alignment modules.

D. Typical Scenes of Proposed and Synthetic Dehazing Datasets

To further demonstrate the high fidelity of the collected hazy scenes and the drawbacks of the synthetic hazy scenes, we also present some typical frames in the REVIDE and REVIDE-SYN datasets according to the prediction results and MOS.

As shown in Fig. C and Fig. D, we present 3 typical synthetic hazy scenes with low fidelity (classified as synthetic scenes and get the most negative scores) and 3 typical collected hazy scenes with high fidelity (classified as real-world scenes and get the most positive scores). According to Fig. C, the synthetic hazy scenes often suffers from the unnatural distribution of synthetic haze (Fig. C (a)), extremely low color saturation (Fig. C (b)), and inconsistent color temperature between scene and synthetic haze (Fig. C (c)). On the other hand, the collected hazy scenes in the REVIDE dataset can address the limitations in the synthetic hazy scenes and

*The mean value of pixel shift is calculated by averaging the optical flows between adjacent haze-free frames by the PWC-Net [11].

†The PSNR of EDVR is 21.22 dB and the PSNR of EDVR-IDA is 21.70 dB.

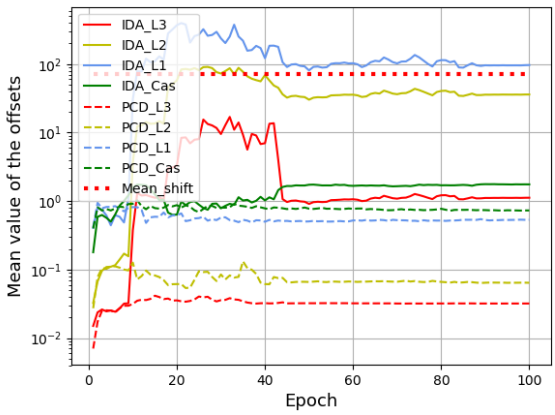


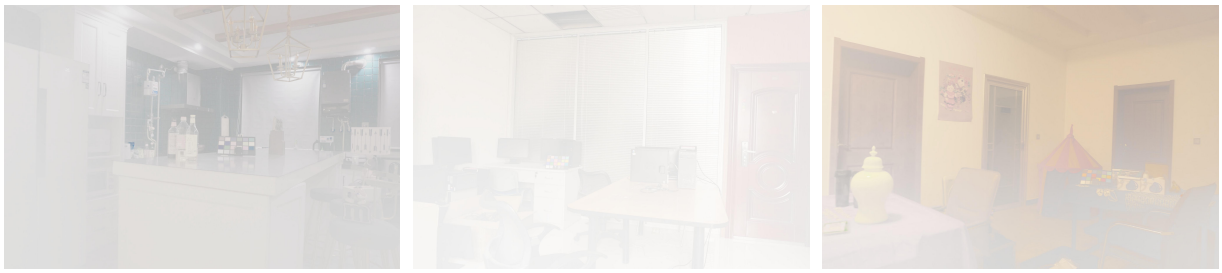
Figure B. **Mean values of the offsets during the training process.** IDA_Ln and PCD_Ln ($n \in \{1, 2, 3\}$) denote the mean values of the learned offsets for the DCNs at the $n - th$ level of the IDA and PCD modules. IDA_Cas and PCD_Cas denote the mean values of the learned offsets for the cascading DCNs of the IDA and PCD modules. The red dotted line denotes the mean value of the pixel shift among the training frames.

contain realistic haze with active light sources (Fig. D (a)), natural dense haze layer (Fig. D (b)), and high fidelity haze that is consistency with the color temperature of background scene (Fig. D (c)).

E. More Results on REVIDE and Real-World Videos

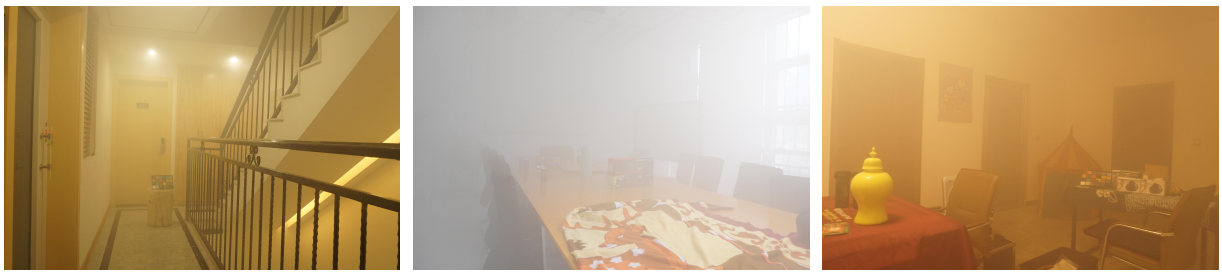
To evaluate state-of-the-art dehazing methods on the REVIDE dataset, we present more quantitative results in the supplemental material. The evaluated method includes: a traditional image dehazing algorithm (NLD [1]), six deep image dehazing algorithms (GFN_dehazing[9], PFFNet [8], GCANet[2], DA_dehazing[10], HardGAN[4]), and one deep video restoration method (WDVR [6]). All the deep learning-based algorithms are re-trained on the training set of the REVIDE dataset. As shown in Tab. A, the architectures with large receptive of field (PFFNet [8] and GCANet [2]) achieve better performance, which validates the motivation of choosing MSBDN [5] as the reconstruction module in the manuscript. We also present the full video results of the DCP [7], EDVR [12], and CG-IDN in Fig. E. The proposed CG-IDN obtains better dehazing results in the whole video.

To evaluate the performance of the proposed method on the real-world videos, we use the semantic segmentation results as the metric to evaluate the perceptual quality of the dehazed videos. Specifically, we collect a real-world hazy video with a smartphone and restore the video by the EDVR [12] and CG-IDN. Then, we use Deeplab-V3 [3] with ResNeSt [13] backbone (trained on the ADE20K dataset [14]) to obtain the semantic segmentation results of the hazy video and two dehazed videos from the EDVR and CG-IDN. By observing the semantic segmentation results,



(a) Unnatural distribution haze (b) Extremely low color saturation (c) Inconsistent temperature color

Figure C. **Typical synthetic hazy scenes with low fidelity**.



(a) Realistic haze with active light sources (b) Natural dense haze layer (c) Consistence temperature color

Figure D. **Typical collected hazy scenes with high fidelity**.

Table A. **More quantitative evaluations on the REVIDE dehazing datasets.** **Red texts** and blue texts indicate the best and the second-best performance respectively.

Methods		NLD [1]	GFN_dehazing [9]	PFFNet [8]	GCANet [2]	DA_dehazing [10]	HardGAN [4]	WDVR [6]	CG-IDN (Ours)
Trained on Real.	PSNR	14.22	17.52	21.59	20.7484	17.23	20.09	17.95	**23.21**
	SSIM	0.7338	0.8027	0.8611	0.8301	0.8390	0.8511	0.7597	**0.8836**

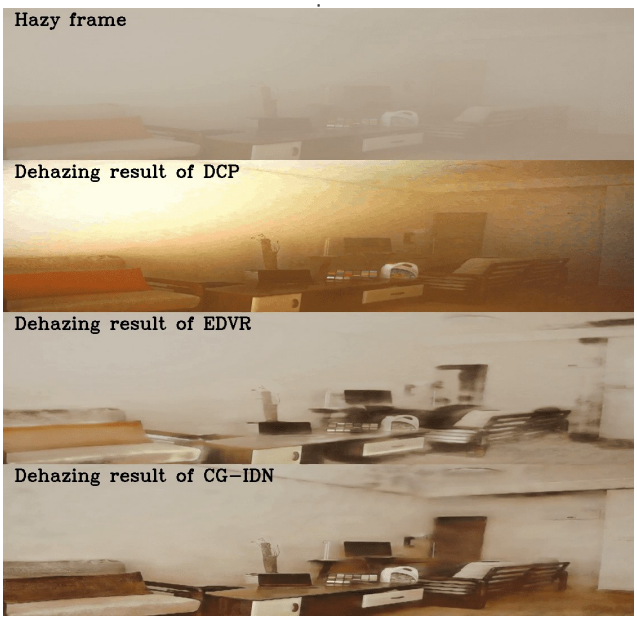


Figure E. **Full video results of the DCP [7], EDVR [12], and CG-IDN**.

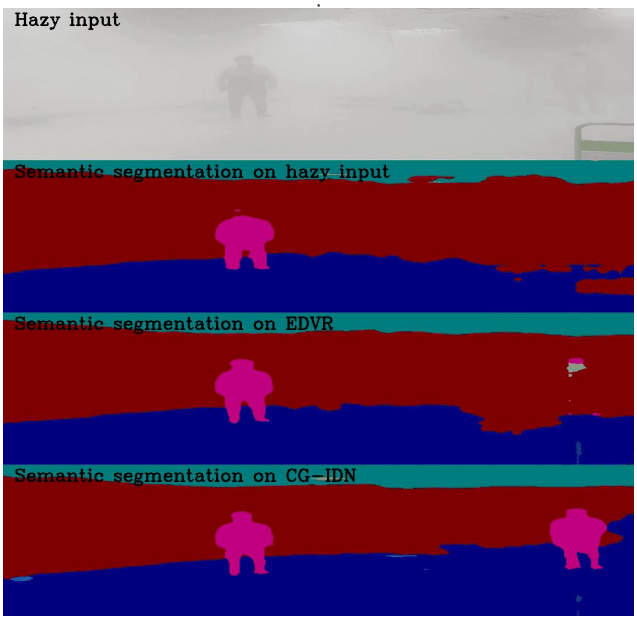


Figure F. **Full video results of semantic segmentation.** The last three rows present the semantic segmentation results on the hazy video and dehazed videos from the EDVR and CG-IDN.

Figure G. **Video results on real-world hazy scenes.** Please view this figure using the Adobe Acrobat Reader.

we can evaluate the perceptual qualities of different dehazing algorithms. As shown in Fig. F, our method can work well on videos with extremely dense hazes and help the semantic segmentation algorithm to recognize all the two persons.

In spite of the absence of outdoor scenes, our dataset can generalize to video dehazing in the wild since that most collected indoor hazy scenes (80.1%) can be distinguished as real-world scenes by the classifier (Table 4). In the meantime, all three sets (ROS, CIS, SIS) receive similar fidelity scores from 50 experienced researchers (Figure 7), which can verify the credibility of the predictions by the classification network. Moreover, we also evaluate the models in the manuscripts on real-world outdoor hazy videos with dynamic scenes to show the generalization of the proposed REVIDE dataset. The Fig. G shows that our method trained on the REVIDE dataset obtains better results with fewer color distortions and temporal flickers.

References

[1] Dana Berman, Shai Avidan, et al. Non-local image dehazing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1674–1682, 2016. 2, 3

[2] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1375–1383, 2019. 2, 3

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2

[4] Qili Deng, Ziling Huang, Chung-Chi Tsai, and Chia-Wen Lin. Hardgan: A haze-aware representation distillation gan for single image dehazing. In *European Conference on Computer Vision*, pages 722–738. Springer, 2020. 2, 3

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2157–2167, 2020. 1, 2

[6] Yuchen Fan, Jiahui Yu, Ding Liu, and Thomas S Huang. An empirical investigation of efficient spatio-temporal modeling in video restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 3

[7] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2011. 2, 3

[8] Kangfu Mei, Aiwen Jiang, Juncheng Li, and Mingwen Wang. Progressive feature fusion network for realistic image dehazing. In *Asian Conference on Computer Vision*, pages 203–215, 2018. 2, 3

[9] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3253–3261, 2018. 2, 3

[10] Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, and Nong Sang. Domain adaptation for image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2808–2817, 2020. 2, 3

[11] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2

[12] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2, 3

[13] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020. 2

[14] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2