Supplementary Material of Semi-Supervised Video Desnowing Network via Temporal Decoupling Experts and Distribution-Driven Contrastive Regularization

Hongtao Wu[1], Yijun Yang[1], Angelica I. Aviles-Rivero[3], Jingjing Ren[1], Sixiang Chen[1], Haoyu Chen[1], and Lei Zhu[1,2]

[1] The Hong Kong University of Science and Technology (Guangzhou), China
[2] The Hong Kong University of Science and Technology, Hong Kong SAR, China
leizhu@ust.hk
[3] University of Cambridge, UK

In this supplementary material, we present network complexity (Section 1), extra visual demonstration (Section 2) and definition of loss functions (Section 3). In addition, a video demo is provided to showcase the effectiveness of our method in the supplementary video.

1 Network Complexity

1.1 Model Complexity and Parameters Comparison

We report the model complexity and parameters comparison in Table 1. The GFLOPs and Runtime are calculated by inferring a video clip of three frames with a resolution of 256×256. We demonstrate the effectiveness of our SemiVDN by achieving state-of-the-art results on the synthetic and real-world video snow removal datasets while maintaining a comparatively minimal computational expense. In the presented tabular data, our method obtained the best restoration performance results compared to other comparative methods and achieved the third fastest speed. Compared to state-of-the-art method SVDNet [3], our SemiVDN has 3.26× fewer FLOPs and runs 2.39× faster. Moreover, our model boosts a 1.66 dB improvement in PSNR compared to the fastest method S2VD [12] , and achieving an inference speed of 19.19 FPS. Thanks to our temporal decoupling experts, our approach can adaptively decompose the backbone feature from the temporal dimension, avoiding the typically employed slow sorting or top-k operations prevalent in MoE-based methodologies [4, 7].

2 More Visual Demonstration

In this section, we show more visual results to demonstrate the effectiveness of the proposed method.

Table 1: Model complexity and parameters comparisons between our network and other methods. Bolded and underlined values indicate the best and the second-best performance, respectively.

Method	PSNR↑	SSIM↑	LPIPS↓	GFLOPs	Runtime(s)	Parameters
TransWeather [10]	23.11	0.8543	0.2086	**18.39**	<u>0.0217</u>	37.68M
WeatherDiffusion [8]	22.01	0.8621	0.1539	745.2	332.17	113.68M
ViWS-Net [11]	24.43	0.8922	0.1142	53.11	0.4002	57.68M
MPRNet [14]	24.27	0.896	0.1266	445.65	0.5057	<u>3.64M</u>
Restormer [13]	24.34	0.8929	0.1164	422.97	0.1864	26.10M
BasicVSR++ [1]	22.64	0.8618	0.1868	907.13	0.2264	6.22M
S2VD [12]	24.02	0.8761	0.1513	<u>30.68</u>	**0.0041**	**0.525M**
SVDNet [3]	<u>25.06</u>	<u>0.9210</u>	<u>0.0842</u>	511.95	0.1248	14.78M
Ours	**25.68**	**0.9254**	**0.0785**	157.13	0.0521	33.92M

2.1 Visual Comparisons Against State-of-the-Art Methods

Figure 1 and Figure 2 demonstrate more visual comparisons between the results generated by our methods and the other compared methods on the RVSD dataset and real-world dataset, respectively. The results on the synthetic dataset demonstrate that our SemiVDN effectively eliminates diverse snow patterns, particularly large snow accumulations. Also, our approach preserves the most natural color compared to alternative comparative methods. Moreover, the results on the real-world dataset prove that our proposed model has better video desnowing ability on real-world scenarios. More video desnowing results can be found in the supplementary video.

2.2 More Results Across Real-World Scenarios

Figure 3 illustrates the results of our model across diverse scenarios. The top row showcases several snowfall scenes observed within forest environments, which are commonly accompanied by the presence of haze and snowflakes of varying particle sizes. The second row displays multiple scenes of daily human activities, including streets, parks and complex lighting at night. The third row exhibits snowfall scenes extracted from movies. In the aforementioned array of scenes, our approach effectively removal snow and haze while preserving the natural coloration. This result is attributable to our proposed semi-supervised video snow removal method, which incorporates both synthetic and real images during the training process and employs contrast learning to address distribution shift.

Fig. 1: Visually comparing video desnowing results produced by our network and state-of-the-art methods. Input frame comes from the RVSD dataset. The proposed method generates high-quality desnowing results with more accurate detail and texture recovery.

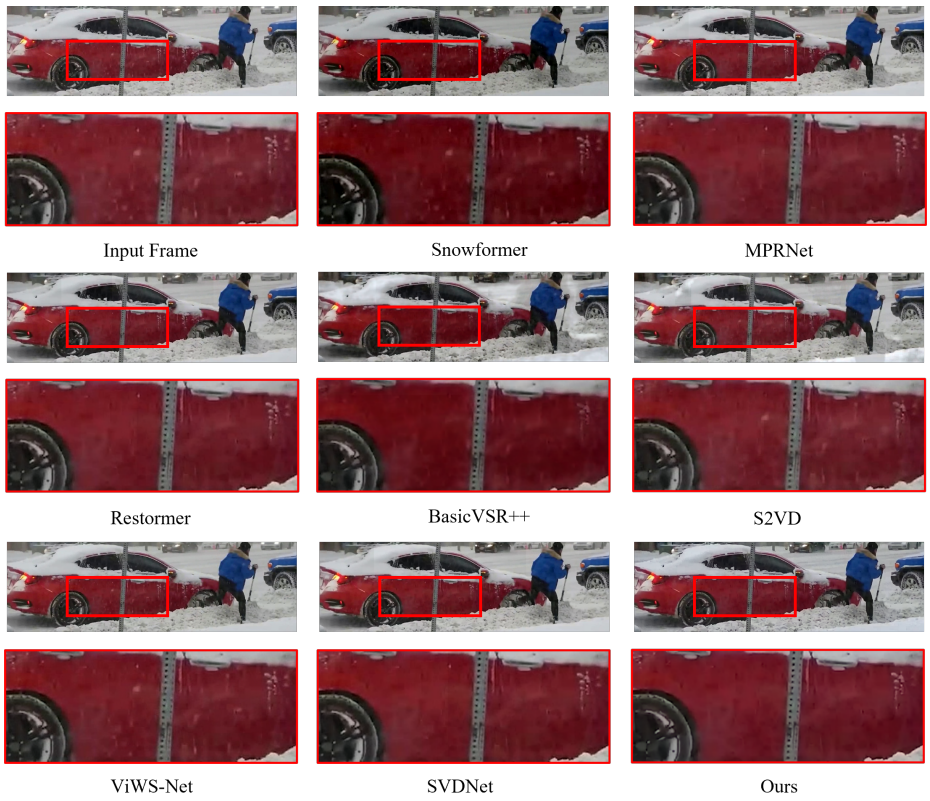

Fig. 2: Visually comparing video desnowing results produced by our network and state-of-the-art methods. Input frame comes from the real-world dataset. Our method can successfully remove most snow particles with various degradation scales and obtain visual pleasant background recovery results.

Fig. 3: More results of our proposed semi-supervised method in real snowy video samples.

3 Definition of Loss Functions

3.1 Supervised Loss

Pixel-wise Loss. We first use the Charbonnier loss [2] as our basic restoration loss:

$$\mathcal{L}_{pixel} = \sqrt{\left(J_t^{gt} - J_t^{syn}\right)^2 + \chi^2}\,, \tag{1}$$

where J_t^{gt} is the corresponding ground-truth of the input snowy video sample, and J_t^{syn} is the corresponding student network's prediction results. The hyper-parameter χ is set to 10^{-6},

Perceptual Loss. We also add a perceptual loss that measures the discrepancy between the features of prediction and the ground truth. We extract features from the $3^{rd}, 8^{th}$ and 15^{th} layers of the pretrained VGG-16 to calculate the perceptual loss. The perceptual loss is formulated as follows:

$$\mathcal{L}_{perceptual} = \mathcal{L}_{MSE}\left(VGG_{3,8,15}(J_t^{gt}), VGG_{3,8,15}(J_t^{syn})\right)\,. \tag{2}$$

Focal Frequency Loss. We introduce the Focal Frequency Loss to focus the model on the response of different regions in the frequency spectrum to varying artifacts in the image, aiming to reduce artifacts and enhance image restoration quality. The Focal Frequency Loss is formulated as:

$$\mathcal{L}_{Frequency} = \frac{1}{HW}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1}\omega_f(u,v)\left|F_{J_t^{gt}}(u,v) - F_{J_t^{syn}}(u,v)\right|^2\,, \tag{3}$$

where the image size is $H \times W$; (u, v) represents the coordinate of a spatial frequency on the frequency spectrum; the matrix element $\omega_f(u, v)$ is the weight for the spatial frequency at (u, v). The complex frequency value $F(u, v)$ denotes the 2D discrete Fourier transform as follows:

$$F(u,v) = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1} f(x,y)\cdot e^{-i2\pi\left(\frac{ux}{H}+\frac{vy}{W}\right)}\,, \tag{4}$$

where (x, y) symbolizes the coordinate of an image pixel in the spatial domain; e and i represent Euler's number and the imaginary unit, espectively. Further elaboration and comprehensive details can be found in [6].

3.2 Unsupervised Loss

Pixel-Level Loss. For semi-supervised learning, we firstly apply the Charbonnier loss to ensure the student network's prediction results J_t^{stu} are consistent with the teacher network's prediction results J_t^{tea} in the pixel level, which can be defined as:

$$\mathcal{L}_{pixel}^{'} = \sqrt{\left(J_t^{tea} - J_t^{stu}\right)^2 + \chi^2}\,. \tag{5}$$

Feature-Level Loss. We further exploit the feature representation of images to constrain the alignment process of unlabelled data in the feature domain. Specifically, we employ contrastive learning to ensure J_t^{stu} is pulled closer to J_t^{tea} and pushed far away from the strongly augmented unlabeled degraded images $\hat{I}_t^{real}$ obtained by the pre-trained VGG feature extractor. The perceptual contrastive loss can be formulated as:

$$\mathcal{L}_{cl} = \sum_{r=1}^{R} \omega_r \cdot \frac{\mathcal{L}_{\mathrm{L}_1}\left(G_r(J_t^{tea}), G_r(J_t^{stu})\right)}{\mathcal{L}_{\mathrm{L}_1}\left(G_r(\hat{I}_t^{real}), G_r(J_t^{stu})\right)}, \tag{6}$$

where $\{G_r \mid r \in [1, R]\}$ extracts the $r-th$ hidden features from the fixed pre-trained VGG-19 model. ω_r is a weight coefficient.

Prior Losses. Furthermore, we employ the dark channel prior (DCP) [5] loss $\mathcal{L}_{DCP}$ and the total variation loss [9] $\mathcal{L}_{TV}$ as the prior losses. The total variation loss is an ℓ_1-regularization gradient prior, which can be expressed as:

$$\mathcal{L}_{TV} = \left\|\partial_{hor} J_t^{stu}\right\|_1 + \left\|\partial_{ver} J_t^{stu}\right\|_1, \tag{7}$$

where ∂_{hor} denotes the horizontal gradient operators, and ∂_{ver} represents the vertical gradient operators.

We also apply the dark channel prior loss to ensure that the dark channel of the predicted images are in consistence with similar statistical characteristics of the clear images:

$$\mathcal{L}_{DCP} = \left\| \min_{y \in N_L(x)} \left[\min_{c \in \{r,g,b\}} (J_t^{stu})^c(y) \right] \right\|_1, \tag{8}$$

where x and y are pixel coordinates of image J_t^{stu}, $(J_t^{stu})^c$ represents $c-th$ color channel of J_t^{stu}, and $N_L(x)$ denotes the local neighborhood centered at x.

References

1. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5972–5981 (2022)
2. Charbonnier, P., Blanc-Feraud, L., Aubert, G., Barlaud, M.: Two deterministic half-quadratic regularization algorithms for computed imaging. In: Proceedings of 1st international conference on image processing. vol. 2, pp. 168–172. IEEE (1994)
3. Chen, H., Ren, J., Gu, J., Wu, H., Lu, X., Cai, H., Zhu, L.: Snow removal in video: A new dataset and a novel method. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13211–13222 (2023)
4. Chen, X., Li, H., Li, M., Pan, J.: Learning a sparse transformer network for effective image deraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5896–5905 (2023)
5. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE transactions on pattern analysis and machine intelligence **33**(12), 2341–2353 (2010)
6. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13919–13929 (2021)
7. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668 (2020)
8. Özdenizci, O., Legenstein, R.: Restoring vision in adverse weather conditions with patch-based denoising diffusion models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
9. Shao, Y., Li, L., Ren, W., Gao, C., Sang, N.: Domain adaptation for image dehazing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2808–2817 (2020)
10. Valanarasu, J.M.J., Yasarla, R., Patel, V.M.: Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2353–2363 (2022)
11. Yang, Y., Aviles-Rivero, A.I., Fu, H., Liu, Y., Wang, W., Zhu, L.: Video adverse-weather-component suppression network via weather messenger and adversarial backpropagation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13200–13210 (2023)
12. Yue, Z., Xie, J., Zhao, Q., Meng, D.: Semi-supervised video deraining with dynamical rain generator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 642–652 (2021)
13. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739 (2022)
14. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14821–14831 (2021)