All-in-One Image Compression and Restoration
–Supplementary Material–

Huimin Zeng[1] Jiacheng Li[1] Ziqiang Zheng[2] Zhiwei Xiong[1,*]
[1]University of Science and Technology of China
[2]The Hong Kong University of Science and Technology

This supplementary document is organized as follows:

– Section 1 provides the rate-distortion (RD) performance of the Gaussian noise degradation setting, where the results are evaluated with MS-SSIM versus BPP.

– Section 2 includes the ablation studies that investigate the number of groups in C-GA, and the effectiveness of the adopted training scheme.

– Section 3 provides more qualitative comparisons on the weather degradation setting and Gaussian noise setting, including synthetic realistic weather-degraded images (Section 3.1), realistic weather-degraded images (Section 3.2), Gaussian noise-degraded images (Section 3.3) and clean images (Section 3.4).

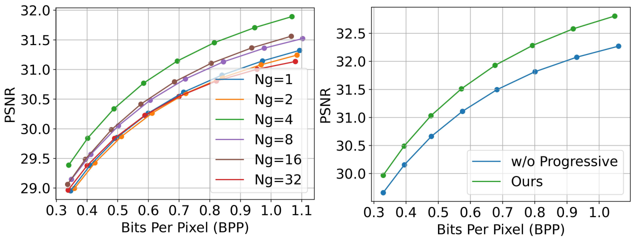
– Section 4 investigates the performance of cascaded solutions regarding the sequence of image restoration and image compression.

– Section 5 provides results of multiple downstream tasks to demonstrate the potential of the proposed method in real-world applications.

– Section 6 provides details of the experimental settings, including the detailed configurations of network architecture (Section 6.1), an overview of the adopted datasets (Section 6.2) and the training details (Section 6.3).

1. Rate-Distortion Performance

Gaussian noise degradation setting. The RD performance on the noisy Kodak dataset [6] is reported in Figure 2, where the inputs are degraded by both seen (*i.e.*, $\sigma = 15, 25, 50$) and unseen (*i.e.*, $\sigma = 35, 45, 55$) Gaussian noise. We evaluate the RD performance with MS-SSIM versus BPP. As shown in Figure 2, Ours-L shows superiority over all compared methods at all noise levels, while containing much lower model complexity and higher inference speed than the cascaded solutions (as outlined in Sec. 4.3). Moreover, despite the joint EVC* showing competitive performance at lower noise levels, its performance drops significantly with the increase in noise levels. Ours-S surpasses the joint EVC* by a large margin and achieves comparable performance with the well-preformed AirNet+EVC, while providing a 7.15× speedup and requiring only 10.91% of the FLOPs. These results highlight the superior performance and generalization ability of the proposed method.



(a) (b)

Figure 1. (a) Ablation study on the number of groups N_g in C-GA. (b) Ablation study on the effectiveness of progressive training strategy.

2. Ablation Studies

We construct a baseline model with the number of groups $N_g = 4$ in Sec. 4.5. In this section, we investigate the rationality of such a configuration, and further demonstrate the effectiveness of the adopted progressive training strategy. All ablation studies are conducted with Ours-S on the weather degradation setting, and evaluated on the RESIDE dataset [8].

Number of groups in C-GA. To identify the optimal configuration regarding the number of groups N_g, we assign various values (*i.e.*, 1, 2, 4, 8, 16 and 32) to N_g, then apply the specified N_g to all C-GA layers in the encoder and decoder across 4 stages. The RD performance comparison is reported in Figure 1(a). As can be seen, the configuration of $N_g = 4$ (depicted as the green curve) achieves the best RD performance. Therefore, we adopt the configuration of $N_g = 4$ in the proposed method.

Effectiveness of progressive training strategy. To evaluate the effectiveness of the progressive training strategy, we remove it and train the network for the same number of iterations (denoted as w/o Progressive). As shown by the

[*] Corresponding author (zwxiong@ustc.edu.cn).

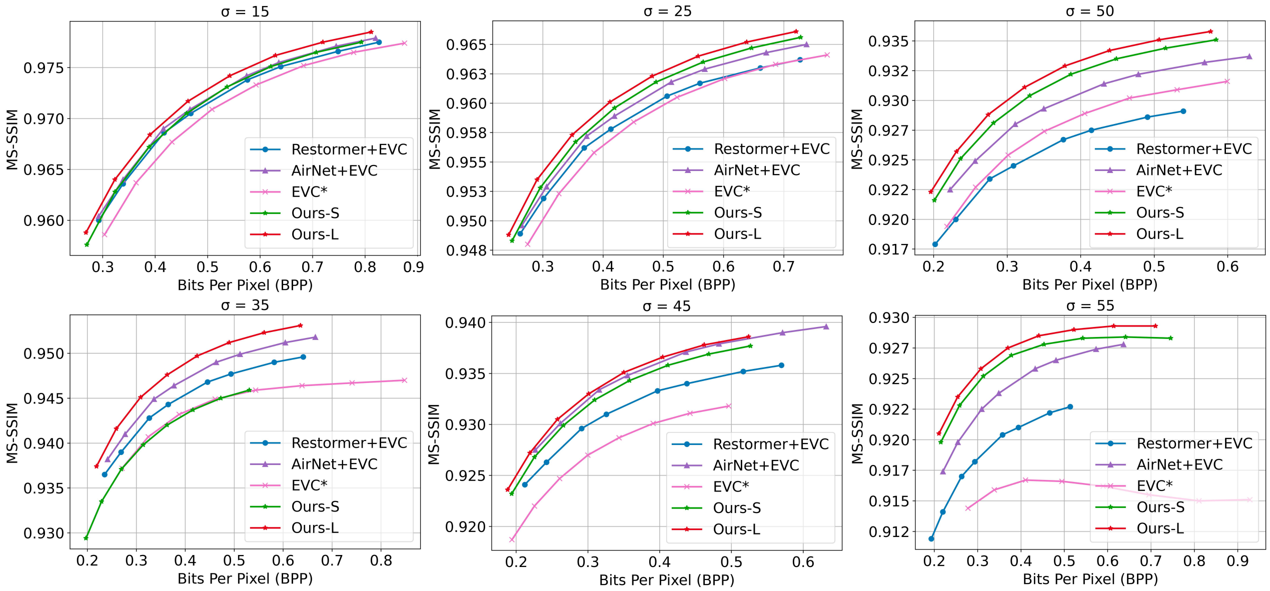


Figure 2. RD performance evaluation on the Kodak dataset [6], where inputs are corrupted by known levels (*i.e.*, 15, 25 and 50) and unknown levels (*i.e.*, 35, 45 and 55) of Gaussian noise. We evaluate the results with MS-SSIM.

blue curve in Figure 1(b), discarding the progressive training strategy results in a noticeable performance drop compared with the original design (green curve).

3. Qualitative Comparisons

3.1. Synthetic Weather-degraded Images

We provide qualitative comparisons on *synthetic* hazy, snowy and rainy images in Figure 3, Figure 6 and Figure 7, respectively. For each image, we provide the quantitative metrics of BPP, PSNR and MS-SSIM. As shown in Figure 3, cascaded solutions and the joint EVC* cannot fully rectify degradations and are likely to introduce color bias for the hazy inputs, such as the buildings in the 1st row. For the snowy results shown in Figure 6, cascaded solutions and joint EVC* fail to effectively eliminate the degradations and may introduce artifacts for degraded regions (*e.g.*, the ground region occluded by snow in the 1st row), while the joint EVC* additionally introduces noise. For rainy results depicted in Figure 7, cascaded methods struggle to distinguish the image content from rain streaks, which results in the loss of valid textures and blur, such as the roof in the 2nd row. The joint EVC* fails in removing the rain streaks and further introduces visually unpleasant noise (*e.g.*, the box in the 4th row). In contrast, our method effectively removes degradation and keeps accurate details with lower bit rates.

3.2. Realistic Weather-degraded Images

We provide more qualitative comparisons on *realistic* hazy, snowy and rainy images in Figure 8, Figure 9 and Figure 10, respectively. As can be seen from Figure 8, the joint EVC* and most cascaded methods struggle in generalizing to realistic hazy images, and may even introduce artifacts (*e.g.*, the results of SwinIR+EVC). Although the cascaded Restormer+EVC successfully eliminates the haze degradation, the results exhibit unnatural contrast and brightness (*e.g.*, the door in the 2nd row). In the snowy scenario depicted in Figure 9, the joint EVC* introduces additional noise and spends extra bits to preserve the degradations. In contrast, our method improves the contrast and effectively eliminates visible snow (*e.g.*, the building in the 1st row), thus outperforming the compared solutions. For the rainy images in Figure 10, the joint EVC* introduces texture distortion, while most cascaded methods fail to remove rain streaks (*e.g.*, the rainy case in the 1st row), and may amplify artifacts in the process of cascaded image restoration and compression (*e.g.*, the wall in the 2nd row). Despite SwinIR+EVC performing well in eliminating rain streaks, it removes valid image structures, such as the corner in the 1st case. In contrast, our method effectively removes rain streaks and preserves the background with lower bit rates.

3.3. Gaussian Noisy Images

Qualitative results of the Gaussian noise degradation setting are shown in Figure 11, where the noise level is set to $\sigma = 15$. As can be seen, although the cascaded methods seem to keep plausible textures, these textures are unreal and distorted (*e.g.*, the hair in the 1st row). Meanwhile, the joint EVC* and cascaded solutions tend to introduce over-smoothness (*e.g.*, the window in the 2nd row), leading to the loss of textures and details. The proposed method effectively eliminates noise degradation and preserves details, demonstrating its ability to handle various levels of noise

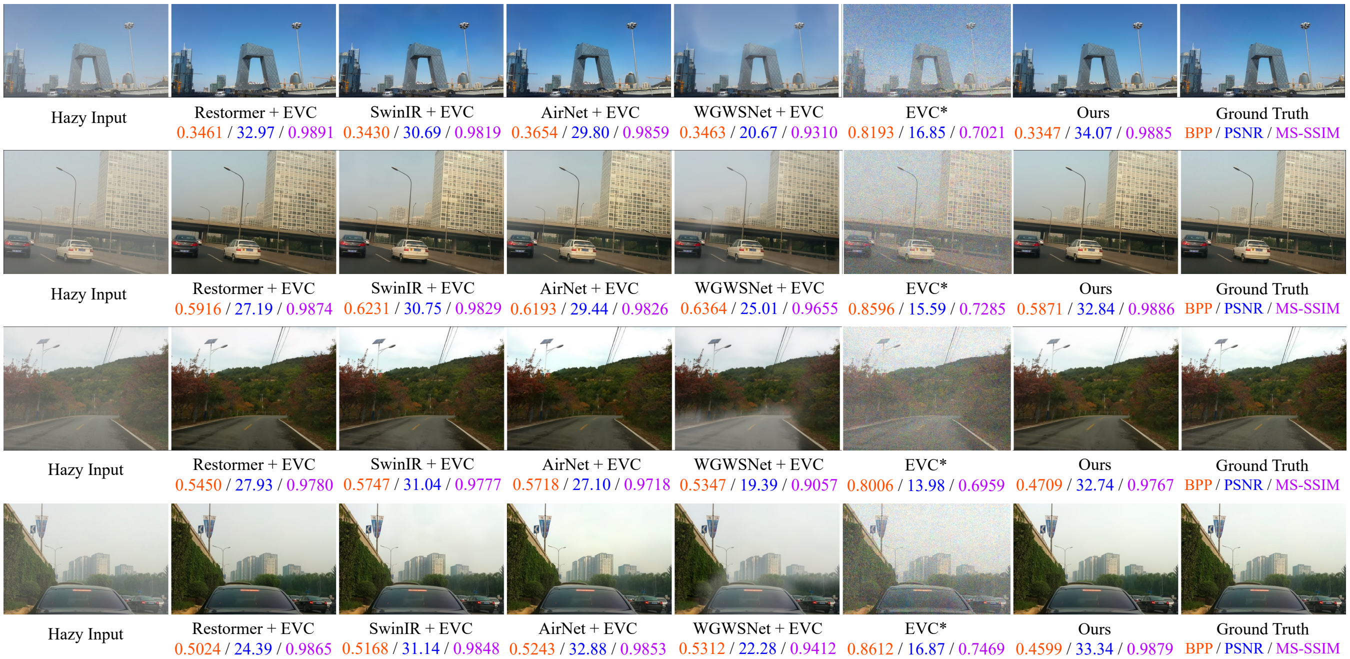


Figure 3. Qualitative comparisons on *synthetic* hazy images, where cascaded solutions are denoted referred to as *restoration + compression*, and Ours denotes the results of Ours-L. For each image, we include metrics of BPP/PSNR/MS-SSIM.

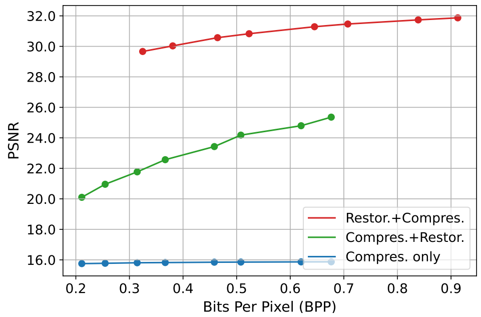


Figure 4. Discussion regarding the sequence of image restoration and compression, where Restor. and Compres. denote restoration and compression, respectively. We evaluate the RD performance with PSNR.

and finer details with a unified framework.

3.4. Clean Images

We provide qualitative comparisons on clean images in Figure 5. As can be seen, despite the proposed method showing a slight drop in quantitative performance compared to the clean-image-specific EVC (Fig. 7[1]), the visual differences are negligible (*e.g.*, the door and flower). When dealing with intricate details, the proposed method even provides more visually pleasing results (*e.g.*, the hair in the 3rd row). However, in challenging scenarios, such as the water ripples in the 4th row, both EVC and our method struggle to deliver high-fidelity results, which occasionally leads to a loss of texture in other regions (*e.g.*, the sky in the last row), since most of the bits are spent to preserve the details of water surface.

4. Sequence of Cascaded Solutions

For the cascaded solutions, we further discuss the sequence of image restoration and image compression, denoted as *restoration+compression* and *compression+restoration*, respectively. We adopt Restormer [16] and EVC [4] for image restoration and image compression, respectively. The performance of EVC [4] on degraded images (denoted as *compression only*) is provided for reference. As illustrated in Figure 4, *compression only* underperforms on degraded images due to its tendency to faithfully preserve degraded inputs. Compared with the *restoration+compression*, *compression+restoration* yields inferior rate-distortion performance, which may result from the degradation mismatch between the compressed results and the subsequent image restoration model. The sequence of *restoration+compression* shows an overall promising performance in improving the quality of inputs and reducing the size of images. Therefore, we compare our models with the *restoration+compression* solution in Sec. 4.2.

5. Real-world Applications

In this section, we devote the compressed results to multiple downstream tasks, *i.e.*, Object Detection (OD) and

[1] To differentiate from this supplementary material, we use abbreviations to denote sections, tables, and figures in the paper (*i.e.*, "Sec." for sections, "Tab." for tables, and "Fig." for figures).

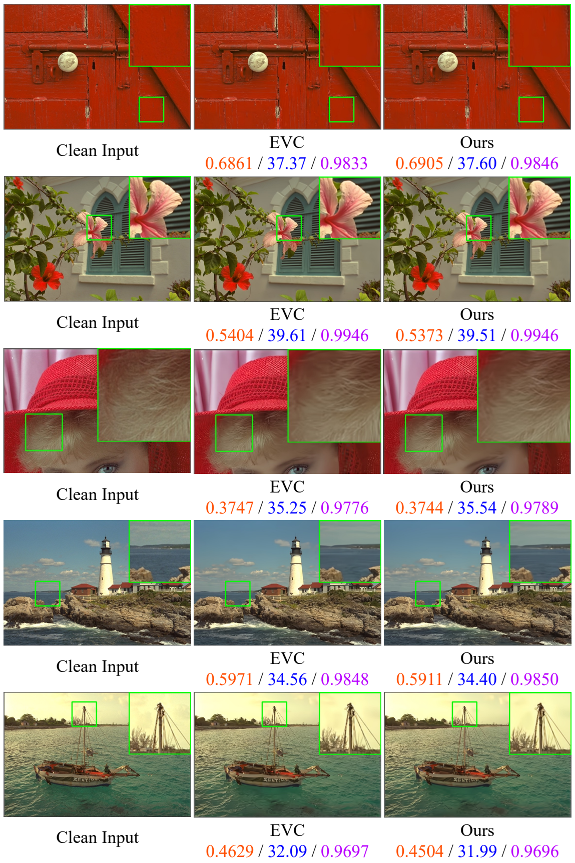


Figure 5. Qualitative comparisons on clean images, where metrics of BPP/PSNR/MS-SSIM are reported for each image.

Method	EVC	Restormer+EVC	AirNet+EVC	Ours-S	Ours-L
mAP ↑	43.21	52.15	<u>54.02</u>	53.93	**54.93**
Recall ↑	0.44	0.51	0.51	<u>0.52</u>	**0.54**
δ_1 ↑	0.859	0.880	0.879	<u>0.936</u>	**0.939**
AbsRel ↓	0.132	0.131	0.125	<u>0.087</u>	**0.083**
RMSE ↓	0.540	0.371	0.383	<u>0.302</u>	**0.292**

Table 1. Results on the task of OD and MDE, where the best and second best results are highlighted with **bold** and underline.

Monocular Depth Estimation (MDE), to evaluate the potential of the proposed method in real applications (*e.g.*, autonomous driving). We adopt the pre-trained Swin Transformer [11] for Object Detection (OD) and Depth Anything [15] for Monocular Depth Estimation (MDE) on the compressed results of RESIDE dataset [8]. To demonstrate the improvement introduced by compared methods and the proposed method, we provide the results of EVC [4] (tailored for clean images) as a reference. We compare with the well-performing cascaded solutions Restormer+EVC and AirNet+EVC. As shown in Table 1, the proposed Ours-L introduces superior improvement over other methods, while Ours-S also achieves competitive performance and surpasses almost all the cascaded methods. The significant improvement over EVC and cascaded methods shows the effectiveness of our method in improving the performance of OD and MDE on degraded images, demonstrating its potential for practical scenarios.

6. Experimental Settings

6.1. Network Architecture

Each stage in the encoder and decoder consists of 4 hybrid-attention transformer blocks. The number of groups N_g in channel-wise group attention (C-GA) is set to 4. For the spatially decoupled attention (S-DA), we set the kernel sizes K_v and K_h of depth-wise convolution to 5. For the entropy model, we adopt the dual spatial prior configuration [9]. In the comparison of attention variants, to keep similar computational complexity, we set the number of MDTA and SWTA to 2-3-3-4 and 1-1-1-1 across the four stages, respectively.

6.2. Dataset

Weather degradation setting. This setting includes weather-related degradations, *i.e.*, haze, snow and rain. For the synthetic images, the Rain1400 dataset [3] contains 12,600 pairs of rainy-clean images for training and 1,400 for testing, with rain streaks of different levels included. The RESIDE dataset [8] comprises the ITS dataset (72,135 images) for training and the OTS dataset (500 images) for testing. The CSD dataset [1] includes 8,000 snowy images for training and 2,000 images for testing. By convention [2,13], we randomly select 5,000 images from each dataset, and merge them for training. Testing splits of these datasets are adopted for quantitative and qualitative evaluation. For the realistic images, six indoor scenes from the REVIDE dataset [17] (with four different styles) are used for evaluation. Snow100K [10] offers 1,329 realistic snowy images for evaluation, which differs a lot from the synthetic snowy scenario. Based on SPA [14], SPA+ [18] removes images with repetitive backgrounds and further densifies the rain streaks.

Gaussian noise degradation setting. We adopt the testing split of Open Images [7] for training, which consists of 125,436 high-quality images. The Kodak [6] dataset provides 24 high-quality images for evaluation.

6.3. Training Details.

During training, to guarantee the versatility of the proposed method for both clean and degraded images, we randomly select clean images as input with a probability of 0.2. For each input image, it is randomly augmented with cropping, horizontal flip, and vertical flip. We adopt the Adam optimizer [5] with $\beta1 = 0.9$, $\beta2 = 0.999$. The initial learning rate is set to 1×10^{-4} and adjusted with the Cosine Annealing scheme [12]. For the progressive training

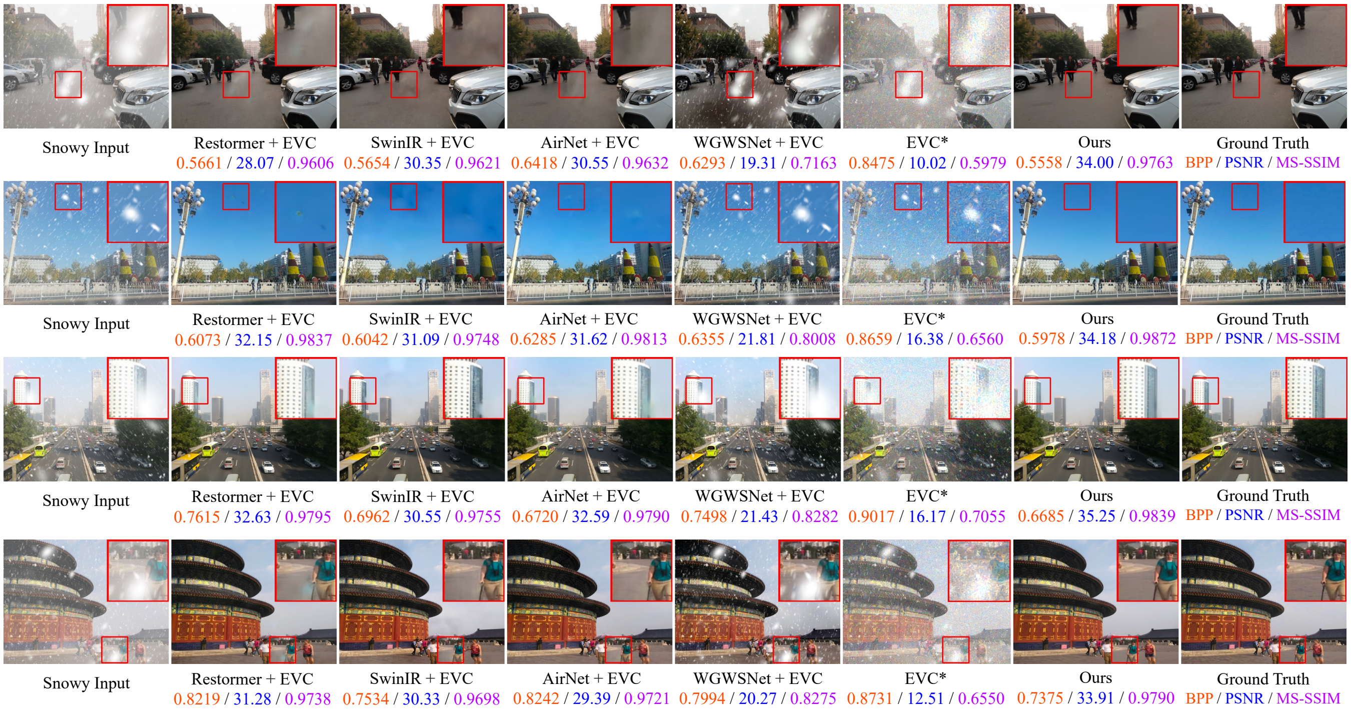


Figure 6. Qualitative comparisons on *synthetic* snowy images, where cascaded solutions are referred to as *restoration + compression*, and Ours denotes the results of Ours-L. For each image, we include metrics of BPP/PSNR/MS-SSIM.

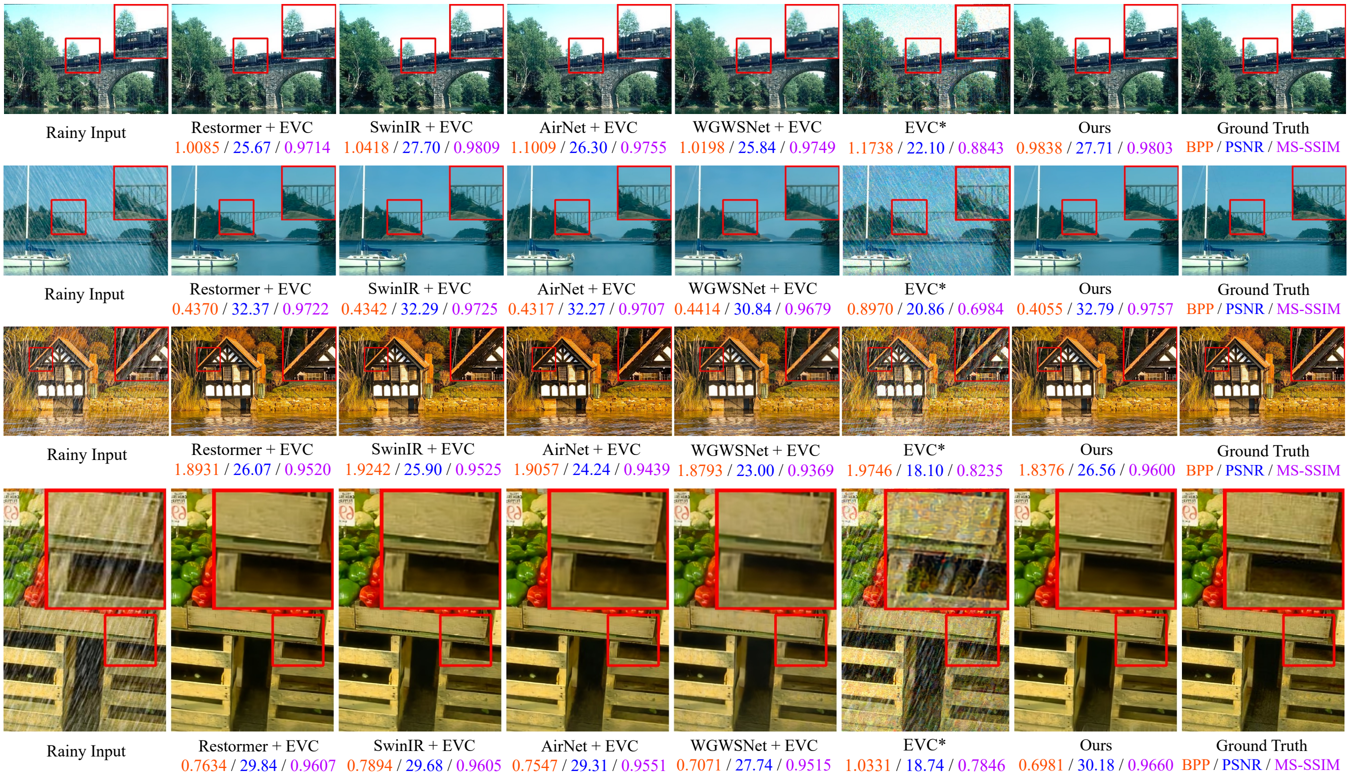


Figure 7. Qualitative comparisons on *synthetic* rainy images, where cascaded solutions are referred to as *restoration + compression*, and Ours denotes the results of Ours-L. For each image, we include metrics of BPP/PSNR/MS-SSIM.

strategy, we train the network with the patch size of 256, 320 and 384 for 250K, 100K and 50K iterations, respectively. To conduct a fast evaluation in the ablation studies, the baseline model that investigates the number of channels

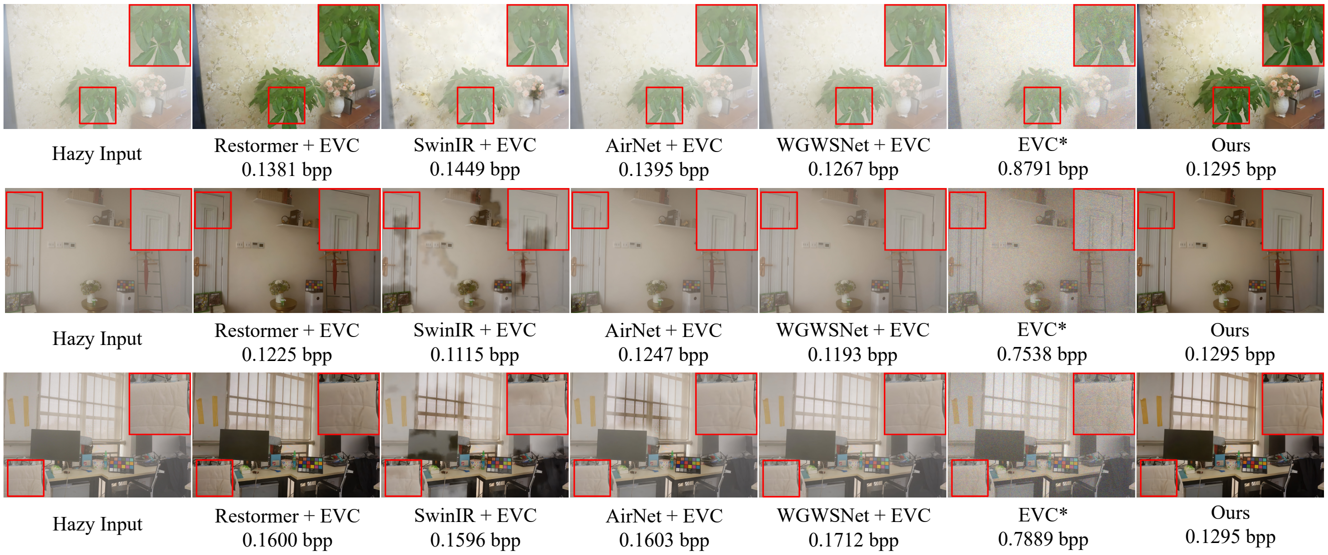


Figure 8. Qualitative comparisons on *realistic* hazy images, where cascaded solutions are denoted referred to as *restoration* + *compression*, and Ours denotes the results of Ours-L. We include BPP for each image.

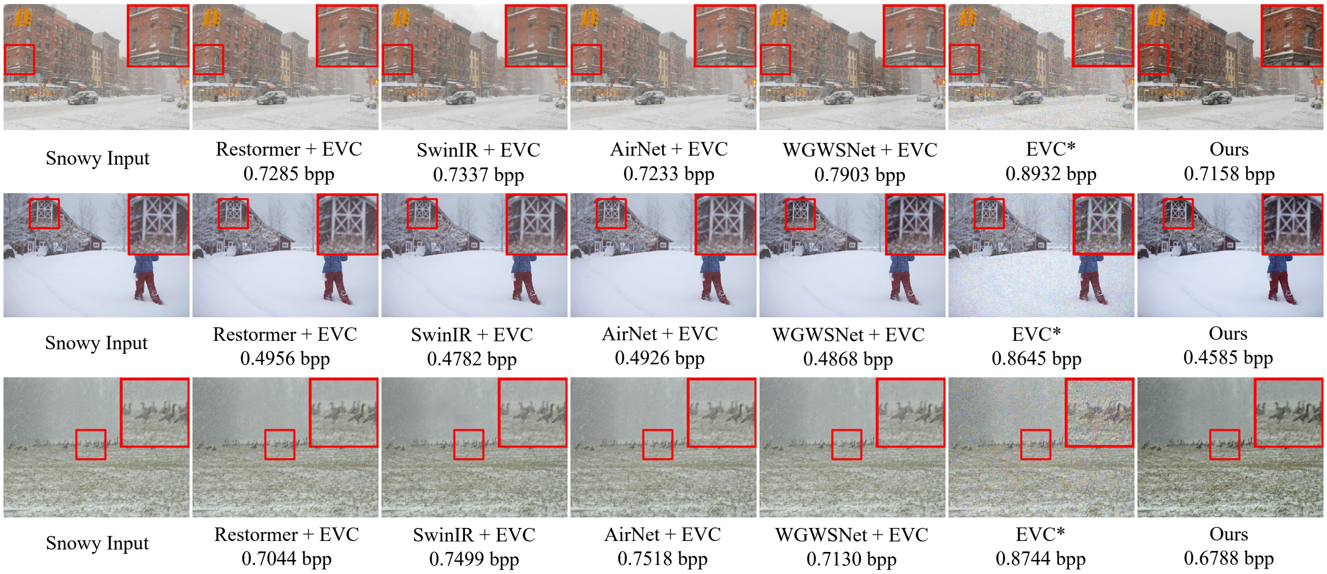


Figure 9. Qualitative comparisons on *realistic* snowy images, where cascaded solutions are denoted referred to as *restoration* + *compression*, and Ours denotes the results of Ours-L. We include BPP for each image.

N_g, the experiment that verifies the effectiveness of S-DA, the model disposing of spatial decoupling design, and the models composed by different attention variants are trained for 300K iterations. To investigate the effectiveness of the progressive training strategy, we train the complete model for 400K iterations under the conditions of with and without the progressive training strategy.

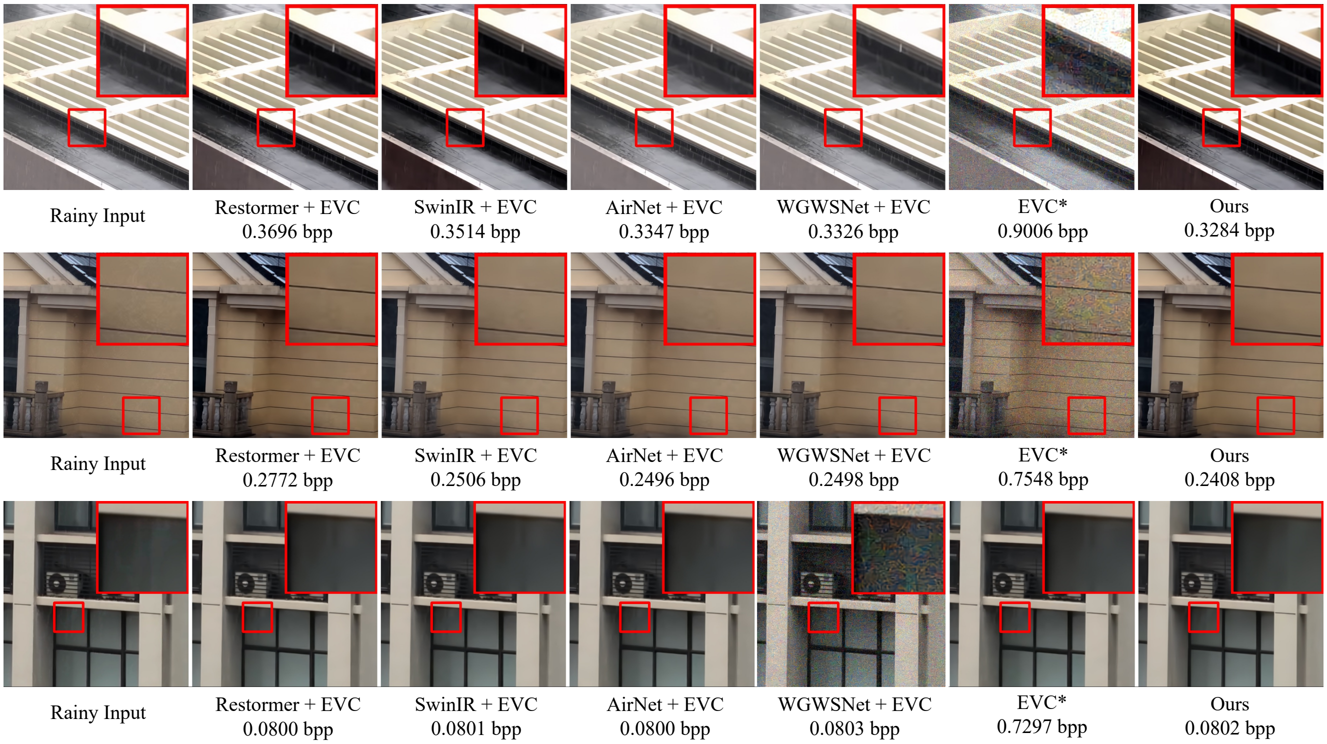
References

[1] Wei-Ting Chen, Hao-Yu Fang, Cheng-Lin Hsieh, Cheng-Che Tsai, I Chen, Jian-Jiun Ding, Sy-Yen Kuo, et al. All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In *ICCV*, pages 4196–4205, 2021. 4

[2] Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen Kuo. Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In *CVPR*, pages 17653–17662, 2022. 4

[3] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *CVPR*, pages 3855–3863, 2017. 4

[4] Wang Guo-Hua, Jiahao Li, Bin Li, and Yan Lu. Evc: Towards real-time neural image compression with mask decay. In *ICLR*, 2023. 3, 4

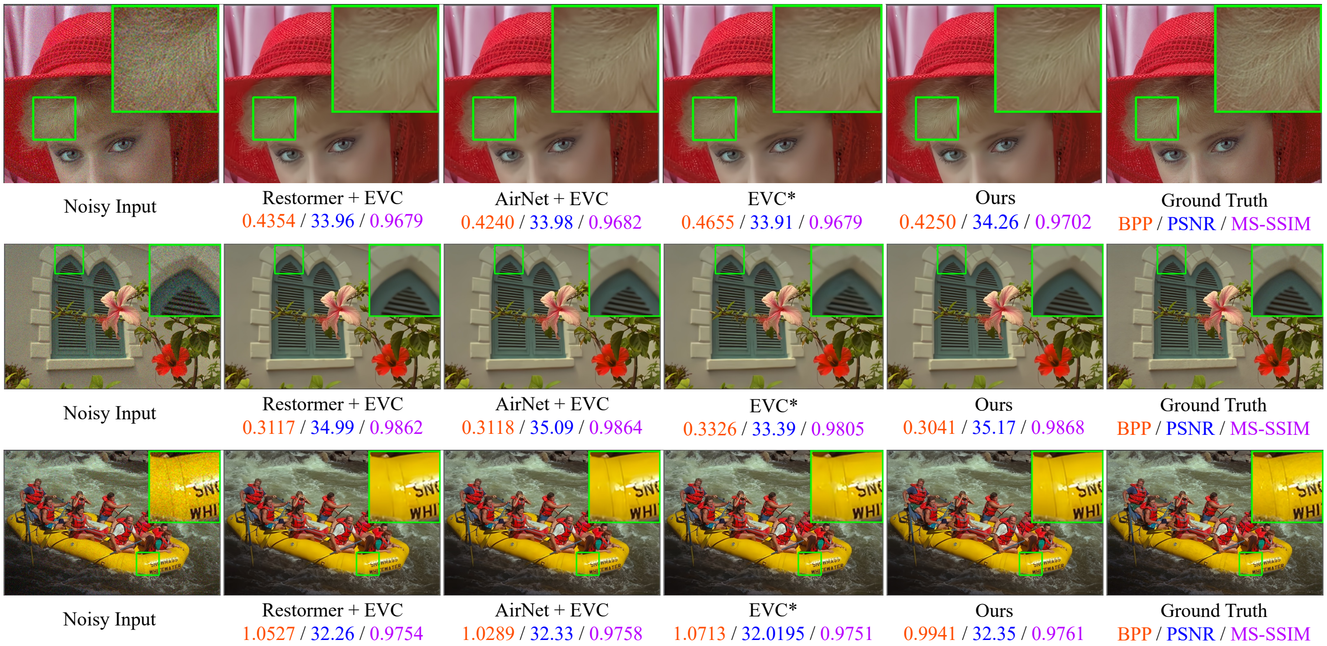


Rainy Input | Restormer + EVC 0.3696 bpp | SwinIR + EVC 0.3514 bpp | AirNet + EVC 0.3347 bpp | WGWSNet + EVC 0.3326 bpp | EVC* 0.9006 bpp | Ours 0.3284 bpp

Rainy Input | Restormer + EVC 0.2772 bpp | SwinIR + EVC 0.2506 bpp | AirNet + EVC 0.2496 bpp | WGWSNet + EVC 0.2498 bpp | EVC* 0.7548 bpp | Ours 0.2408 bpp

Rainy Input | Restormer + EVC 0.0800 bpp | SwinIR + EVC 0.0801 bpp | AirNet + EVC 0.0800 bpp | WGWSNet + EVC 0.0803 bpp | EVC* 0.7297 bpp | Ours 0.0802 bpp

Figure 10. Qualitative comparisons on *realistic* rainy images, where cascaded solutions are denoted referred to as *restoration + compression*, and Ours denotes the results of Ours-L. We include BPP for each image.



Noisy Input | Restormer + EVC 0.4354 / 33.96 / 0.9679 | AirNet + EVC 0.4240 / 33.98 / 0.9682 | EVC* 0.4655 / 33.91 / 0.9679 | Ours 0.4250 / 34.26 / 0.9702 | Ground Truth BPP / PSNR / MS-SSIM

Noisy Input | Restormer + EVC 0.3117 / 34.99 / 0.9862 | AirNet + EVC 0.3118 / 35.09 / 0.9864 | EVC* 0.3326 / 33.39 / 0.9805 | Ours 0.3041 / 35.17 / 0.9868 | Ground Truth BPP / PSNR / MS-SSIM

Noisy Input | Restormer + EVC 1.0527 / 32.26 / 0.9754 | AirNet + EVC 1.0289 / 32.33 / 0.9758 | EVC* 1.0713 / 32.0195 / 0.9751 | Ours 0.9941 / 32.35 / 0.9761 | Ground Truth BPP / PSNR / MS-SSIM

Figure 11. Qualitative comparisons on Gaussian noise-degraded images, where the noise level of input is set to $\sigma = 15$. Results of cascaded solutions are denoted as *restoration+compression*. We report metrics of BPP/PSNR/MS-SSIM for each image.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[6] Eastman Kodak. Kodak lossless true color image suite (photocd pcd0992). *URL http://r0k. us/graphics/kodak*, 6, 1993. 1, 2, 4

[7] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Ui-

jlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017. 4

[8] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *TIP*, 28(1):492–505, 2018. 1, 4

[9] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *MM*, pages 1503–1511, 2022. 4

[10] Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *TIP*, 27(6):3064–3073, 2018. 4

[11] 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*, pages 10012–10022, 2021. 4

[12] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 4

[13] Dongwon Park, Byung Hyun Lee, and Se Young Chun. All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations. In *CVPR*, pages 5815–5824. IEEE, 2023. 4

[14] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *CVPR*, pages 12270–12279, 2019. 4

[15] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024. 4

[16] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022. 3

[17] 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*, pages 9239–9248, 2021. 4

[18] Yurui Zhu, Tianyu Wang, Xueyang Fu, Xuanyu Yang, Xin Guo, Jifeng Dai, Yu Qiao, and Xiaowei Hu. Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions. In *CVPR*, pages 21747–21758, 2023. 4