RIDCP: Revitalizing Real Image Dehazing via High-Quality Codebook Priors
- Supplementary Materials -

Rui-Qi Wu[1] Zheng-Peng Duan[1] Chun-Le Guo[1*] Zhi Chai[2] Chongyi Li[3]
[1]VCIP, CS, Nankai University [2]Hisilicon Technologies Co. Ltd.
[3]S-Lab, Nanyang Technological University
{wuruiqi, adamduan0211}@mail.nankai.edu.cn, guochunle@nankai.edu.cn,
chaizhi2@huawei.com, chongyi.li@ntu.edu.sg

Figure 1. Dehazing results of data captured by us. The proposed RIDCP performs well on both daytime and nighttime.

Abstract

Our supplementary materials give more details of our RIDCP and more experiments results, which can be summarized as follows:

- *We provide the detailed architectures and the training objectives of the pre-trained VQGAN network and our RIDCP.*
- *We provide more visual results on RTTS to demonstrate the superior performance of the proposed RIDCP.*
- *We provide more qualitative comparisons to prove the effectiveness of HQPs and the proposed phenomenological degradation pipeline.*
- *We provide a video demo to show our RIDCP's potential in real video dehazing.*

1. Network Details

1.1. Detailed Architecture

Table 1 illustrates the detailed architecture of our RIDCP and the correspondence output size. Each encoder layer is consist of a down-sampling convolutional layer with a sliding stride of 2 and two residual layers [7]. Each decoder layer is consist of an up-sampling operation, a convolutional layer, and two residual layers [7].

1.2. Traning Objectives

VQGAN. Since the vector-quantized operation is non-differentiable, VQGAN is end-to-end trained by copying the gradients of $\mathbf{G}_{vq}$ to $\mathbf{E}_{vq}$ [1]. The optimization strategy can be divided into two parts, which are minimizing the loss after reconstruction and after feature matching, respectively. For the first part, the loss function can be formulated as:

$$\mathcal{L}_{rec} = ||x' - x||_1 + \mathcal{L}_{per} + \mathcal{L}_{adv}, \tag{1}$$

where $\mathcal{L}_{per}$ and $\mathcal{L}_{adv}$ are perceptual loss [8] and adversarial loss [9], respectively. And for codebook optimization, the loss function can be written as:

$$\begin{aligned}\mathcal{L}_{codebook} = ||sg(\hat{z}) - z^q||_2^2 + \beta||sg(z^q) - \hat{z}||_2^2 \\ +\gamma||CONV(z^q) - \phi(x)||_2^2,\end{aligned} \tag{2}$$

where $sg(\cdot)$ is the stop-gradient operation, and $\beta = 0.25, \gamma = 0.1$ respectively. The last term of $\mathcal{L}_{codebook}$ is a semantic guided regularization term follow [2], where $CONV$ is a simple convolutional layer, and ϕ is the pre-

Layers	Configurations	Output Size
Input	RGB Image	$h \times w \times 3$
Conv1	$c = 64 \quad k = 3$	$h \times w \times 64$
Enc1	$c = 128 \quad k = 3$	$\frac{h}{2} \times \frac{w}{2} \times 128$
Enc2	$c = 256 \quad k = 3$	$\frac{h}{4} \times \frac{w}{4} \times 256$
RSTB	$\begin{bmatrix} c = 256 \\ h = 8 \\ ws = 8 \end{bmatrix} \times 4$	$\frac{h}{4} \times \frac{w}{4} \times 256$
Conv2	$c = 512 \quad k = 1$	$\frac{h}{4} \times \frac{w}{4} \times 512$
Codebook	$c = 512 \quad K = 1024$	$\frac{h}{4} \times \frac{w}{4} \times 512$
Conv3	$c = 256 \quad k = 1$	$\frac{h}{4} \times \frac{w}{4} \times 256$
Dec1_vq	$c = 128 \quad k = 3$	$\frac{h}{2} \times \frac{w}{2} \times 128$
Dec2_vq	$c = 64 \quad k = 3$	$h \times w \times 64$
Dec1	$c = 128 \quad k = 3$	$\frac{h}{2} \times \frac{w}{2} \times 128$
Dec2	$c = 64 \quad k = 3$	$h \times w \times 64$
Conv4	$c = 3 \quad k = 3$	$h \times w \times 3$

Table 1. Architecture details of the RIDCP. c denotes the output channel number, k represents the kernel size, and K is the codebook size. h and ws are number of heads and window size respectively.

trained VGG19 [12]. Finally, the total loss of VQGAN is:

$$\mathcal{L}_{vq} = \mathcal{L}_{rec} + \mathcal{L}_{codebook}. \tag{3}$$

RIDCP. For encoder $\mathbf{E}$, we use pretrained VQGAN to teach it to find the correct code. Assuming that the input hazy image is x_h and the clear counterpart is x_{gt}, we can get features $\hat{z}_h = \mathbf{E}(x_h)$ and $z^q_{gt} = \mathcal{M}(\mathbf{E}_{vq}(x_{gt}))$. The loss function $\mathcal{L}_{\mathbf{E}}$ to optimize $\mathbf{E}$ can be formulated as:

$$\mathcal{L}_{\mathbf{E}} = ||\hat{z}_h - z^q_{gt}||^2_2 + \lambda_{style}||\Psi(\hat{z}_h - \Psi(z^q_{gt}))||^2_2 + \lambda_{adv} \sum_i -\mathbb{E}[D(\hat{z}^i_h)], \tag{4}$$

where Ψ is the Gram matrix calculation to build style loss [5] and D is the discriminator to supervise $\mathbf{E}$ adversarially. And x_{gt} is used for supervising $\mathbf{G}$, which can be written as:

$$\mathcal{L}_{\mathbf{G}} = ||y - x_{gt}||_1 + \lambda_{per}||\phi(y) - \phi(x_{gt})||^2_2, \tag{5}$$

where y is the output and ϕ the pretrained VGG16 [12]. Besides, the gradients of $\mathcal{L}_{\mathbf{G}}$ do not propagate backwards to $\mathbf{E}$.

2. Experiments Results

Since there is no ground-truth for real image dehazing tasks, quantitative metrics are difficult to reflect the true performance of the dehazing algorithms. Meanwhile, we provide extensive qualitative results in the section to further demonstrate the superior performance of the proposed RIDCP and the effectiveness of each key component.

2.1. More Visual Results

Figure 1 presents two dehazing cases on our own-captured data. Our dehazing method performs well on both daytime and nighttime scenes. Figure 2, 3 and 4 illustrate more visual comparisons with several state-of-the-art methods on RTTS [10] dataset. As we can see, the proposed RIDCP achieves satisfactory performance and maintains stable dehazing ability in scenes with dense haze and heavy color bias.

2.2. Ablation Study

We analyze the effectiveness of HQPs and the proposed phenomenological degradation pipeline. In Figure 5, we can observe that HQPs can help the network generate results with better brightness and lower color bias. Figure 6 shows the significant improvement in dehazing capability brought by our pipeline.

3. Broader Impacts

Our RIDCP performs well on real-world hazy scenes, which can possibly be applied to some industrial tasks like automatic driving and computational photography. Moreover, the proposed phenomenological degradation pipeline can also generally boost the performance of dehazing algorithms, which is beneficial for the development of real image dehazing. Thus, we believe our work will bring positive impacts on both academia and industry. As a typical low-level vision work, this paper will not bring negative impacts to society.

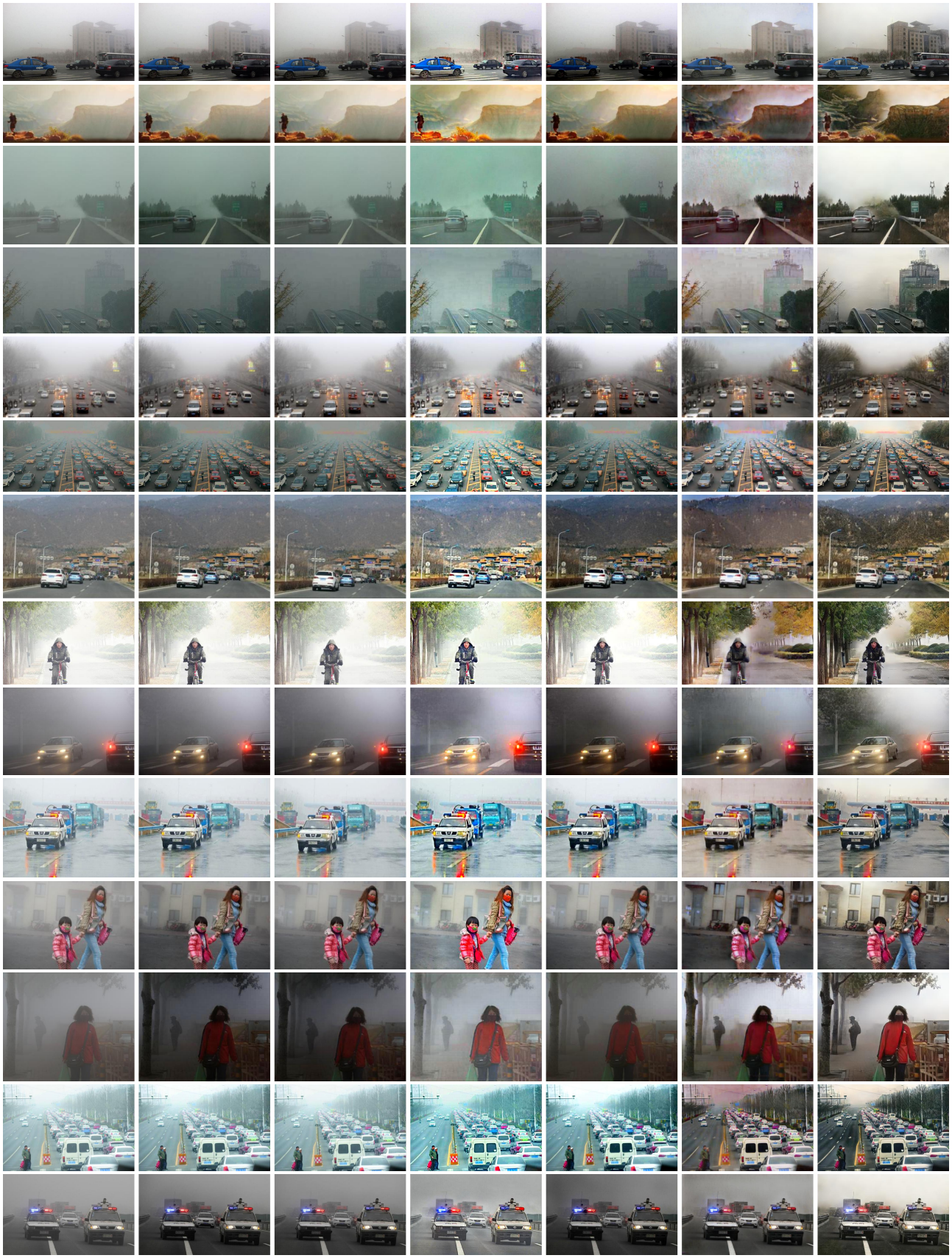
References

[1] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 1

[2] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pages 1329–1338, 2022. 1

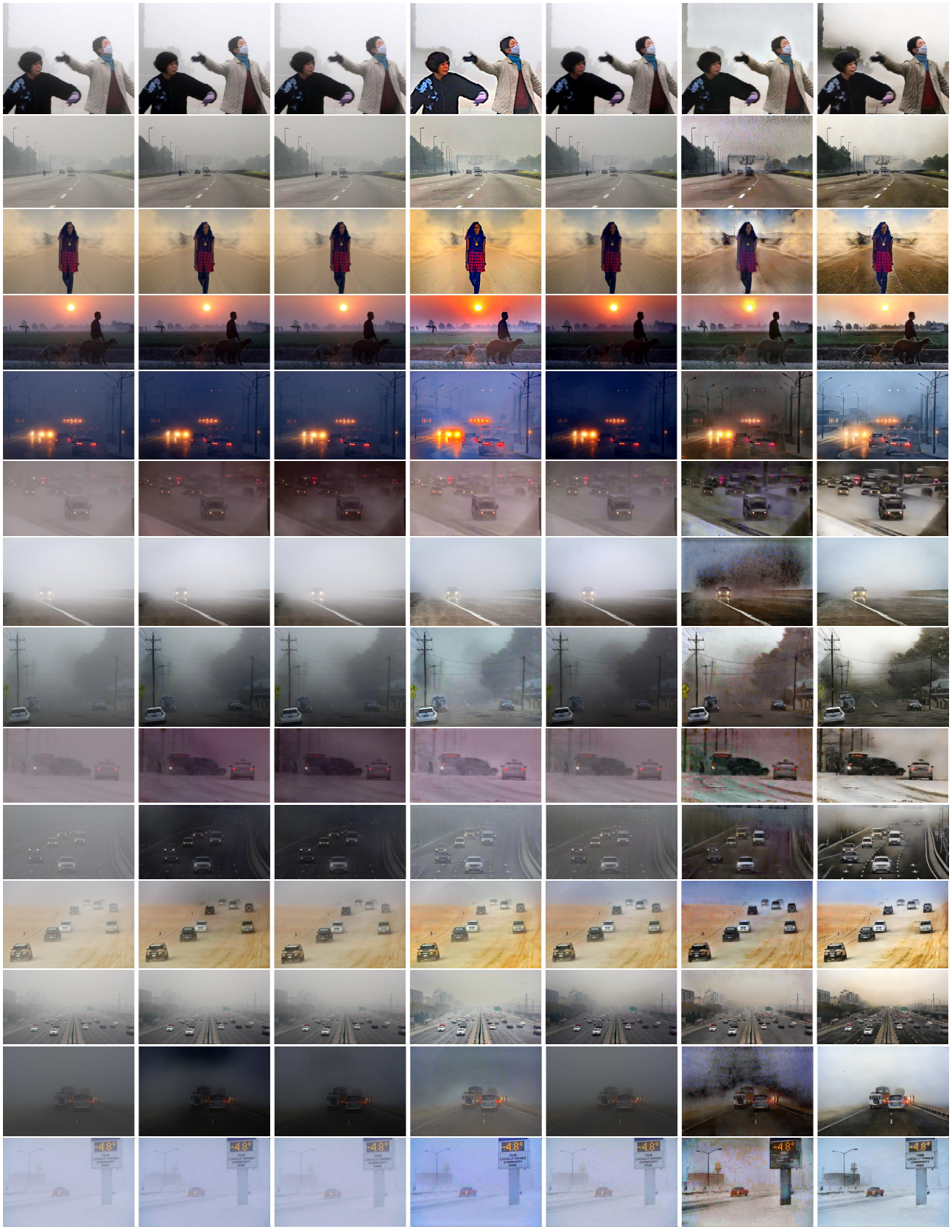
[3] Zeyuan Chen, Yangchao Wang, Yang Yang, and Dong Liu. Psd: Principled synthetic-to-real dehazing guided by physical priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7180–7189, 2021. 3, 4, 5

[4] 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*



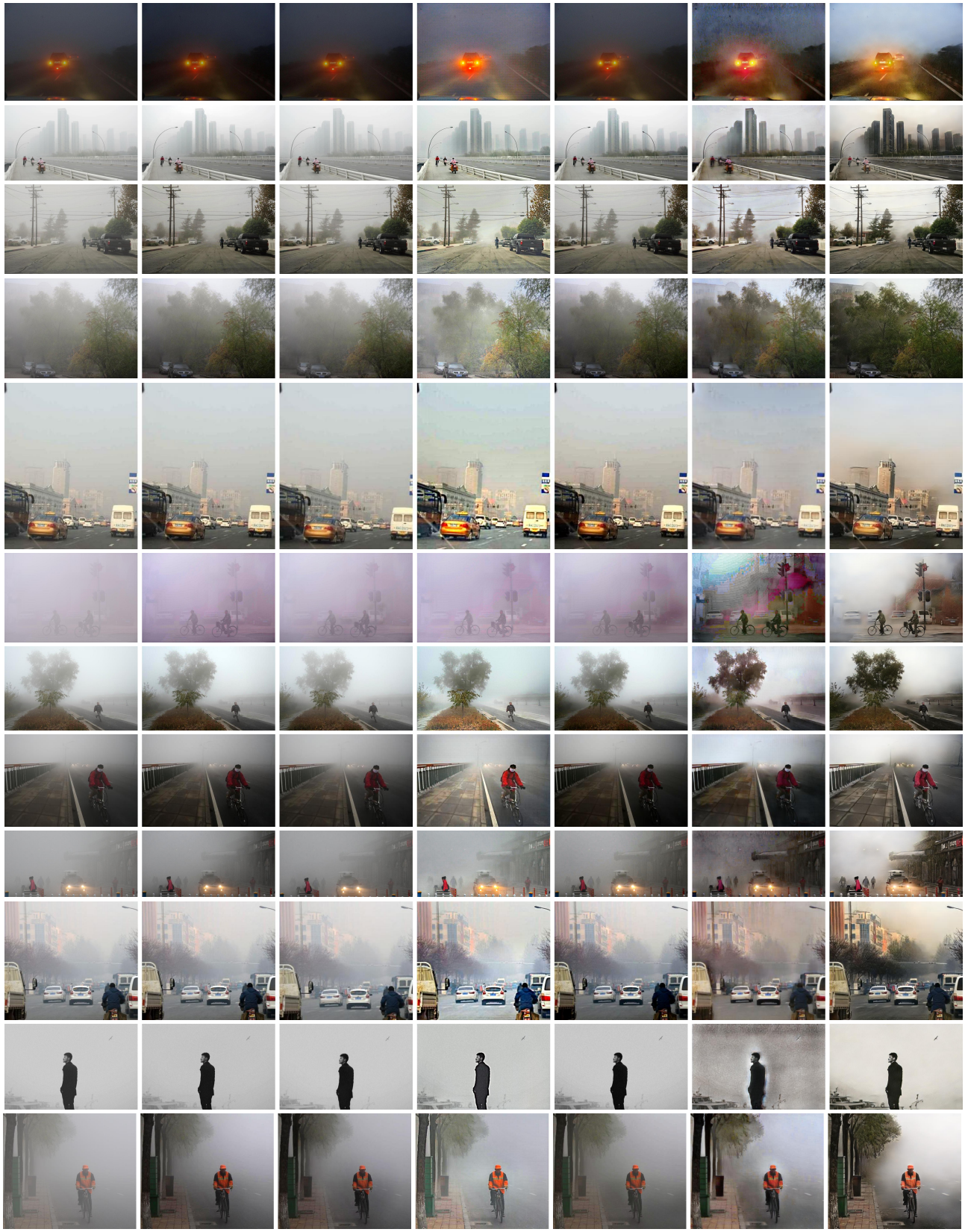
(a) Hazy image (b) MSBDN [4] (c) Dehamer [6] (d) PSD [3] (e) D4 [13] (f) DAD [11] (g) RIDCP

Figure 2. More visual comparisons on RTTS. **Zoom-in for best view.**



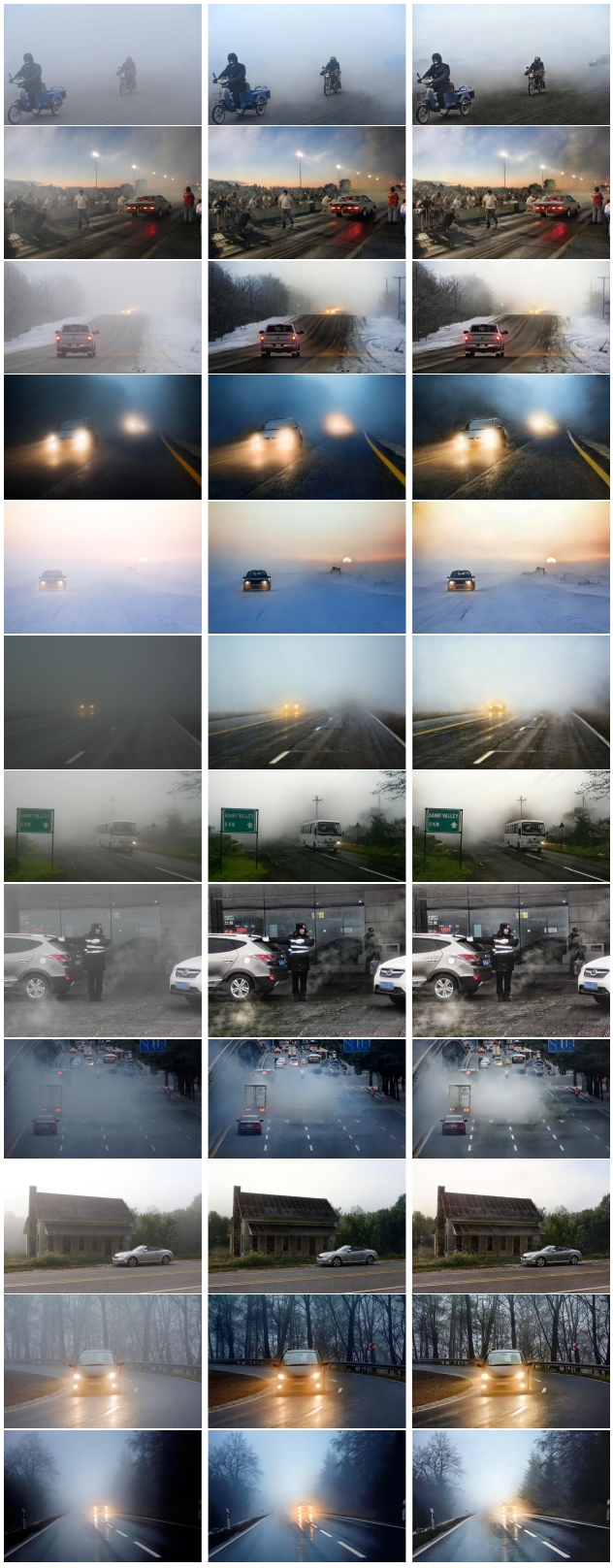
(a) Hazy image (b) MSBDN [4] (c) Dehamer [6] (d) PSD [3] (e) D4 [13] (f) DAD [11] (g) RIDCP

Figure 3. More visual comparisons on RTTS. **Zoom-in for best view.**



(a) Hazy image (b) MSBDN [4] (c) Dehamer [6] (d) PSD [3] (e) D4 [13] (f) DAD [11] (g) RIDCP

Figure 4. More visual comparisons on RTTS. **Zoom-in for best view.**



(a) Hazy image (b) w/o HQPs (c) Full model

Figure 5. Ablation results on HQPs.

(a) Hazy image (b) OTS (c) Haze4K (d) Our pipeline

Figure 6. Ablation results on the proposed phenomenological degradation pipeline.

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2157–2167, 2020. 3, 4, 5

[5] Muhammad Waleed Gondal, Bernhard Schölkopf, and Michael Hirsch. The unreasonable effectiveness of texture transfer for single image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 80–97. Springer, 2018. 2

[6] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5812–5820, 2022. 3, 4, 5

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1

[9] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017. 1

[10] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing (TIP)*, 28(1):492–505, 2019. 2

[11] 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 (CVPR)*, pages 2808–2817, 2020. 3, 4, 5

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations (ICLR)*, 2015. 2

[13] Yang Yang, Chaoyue Wang, Risheng Liu, Lin Zhang, Xiaojie Guo, and Dacheng Tao. Self-augmented unpaired image dehazing via density and depth decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2037–2046, 2022. 3, 4, 5