Supplementary Materials to "Efficient and Degradation-Adaptive Network for Real-World Image Super-Resolution"

Jie Liang[1,2], **Hui Zeng**[2] and **Lei Zhang**[1,2,⋆]

[1]The HongKong Polytechnic University, [2]OPPO Research
{*liang27jie, cshzeng*}*@gmail.com*; *cslzhang@comp.polyu.edu.hk*

In this supplementary file, we provide: (1) the details of our training losses; (2) the detailed settings of our degradation modeling; (3) more sample images in our constructed dataset; (4) more qualitative comparisons and (5) more quantitative ablation studies to further validate the effectiveness of the proposed DASR.

1 Details of Training Losses

As discussed in Section 3.1 of the main paper, the total training loss is defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pixel}} + \lambda_1 \mathcal{L}_{\text{regression}} + \lambda_2 \mathcal{L}_{\text{perceptual}} + \lambda_3 \mathcal{L}_{\text{adversarial}},$$

where the regression loss $\mathcal{L}_{\text{regression}}$ has been provided in Eq. (1) of the main paper. For the other three losses, the settings are the same as in Real-ESRGAN. Specifically, the pixel loss is defined as the ℓ_1 distance $\mathcal{L}_{\text{pixel}} = \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_1$, where $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$ denote the super-resolved image and the ground-truth HR image, respectively. For the perceptual loss $\mathcal{L}_{\text{perceptual}}$, we first extract the $\{\text{conv}_1, \text{conv}_2, \text{conv}_3, \text{conv}_4, \text{conv}_5\}$ feature maps of $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$ by using the pre-trained VGG19 network [1], then calculate the weighted sum of the respective ℓ_1 distances between the feature maps of $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$ as the perceptual loss, where the weights are set to be [0.1, 0.1, 1, 1, 1]. For the adversarial loss $\mathcal{L}_{\text{adversarial}}$, the U-Net discriminator with spectral normalization is adopted.

2 Detailed Settings of Degradation Modeling

We report the detailed parameter settings of our degradation modeling in Table 1. We partition the whole degradation space S into 3 levels $[S_1, S_2, S_3]$, and randomly select one of them to generate the LR-HR image pairs during training with a balanced probability of $[0.3, 0.3, 0.4]$. For the blur operation, we use isotropic and anisotropic Gaussian kernels with a probability of $[0.65, 0.35]$, where we set $\sigma_1 = \sigma_2$ if isotropic blur kernel is specified. In the second degradation stage of S_3, following the practice in Real-ESRGAN, we skip the blur

⋆ Corresponding author.

⋆⋆ This work is supported by the Hong Kong RGC RIF grant (R5001-18) and the PolyU-OPPO Joint Innovation Lab.

operation with a probability of 0.2, and perform sinc kernel filtering with a probability of 0.8. We finally resize the image to the desired LR size, *i.e.*, 1/4 of the original size.

Table 1. Detailed parameter settings of the degradation sub-spaces $[S_1, S_2, S_3]$. Here, '-' indicates that the operation is not activated and the corresponding value in $\boldsymbol{v}$ is padded with 0; ['a', 'b', 'b'] denote the resize modes of [area, bilinear, bicubic]; ['G', 'P'] denote the noise types of [Gaussian, Poisson]; ω_c is the cutoff frequency of the sinc kernel; R-J and J-R indicate the different operating orders of resizing and JPEG compression; v_i denotes the i^{th} value of $\boldsymbol{v}$.

Level	Operation	Parameter	Stage 1 Range	Stage 1 v_i	Stage 2 Range	Stage 2 v_i
S_1	Blur	kernel size $[2m+1]$	$m \in [3, 10]$	v_1	-	-
		standard deviation σ_1	$[0.2, 0.8]$	v_2	-	-
		standard deviation σ_2	$[0.2, 0.8]$	v_3	-	-
		rotation degree θ	$[-\pi, \pi]$	v_4	-	-
	Resize	[up, down, keep]	$[0.1, 0.2, 0.7]$	-	-	-
		scale factor	$[0.85, 1.2]$	v_{11}	-	-
		resize mode	['a', 'b', 'b']	$v_{12} \sim v_{14}$	-	-
	Noise	type	['G', 'P']	v_{21}, v_{22}	-	-
		sigma of Gaussian	$[1, 10]$	v_{19}	-	-
		scale of Poisson	$[0.05, 0.5]$	v_{19}	-	-
		gray probability	0.4	v_{20}	-	-
	JPEG	quality factor	$[90, 95]$	v_{27}	-	-
		mode of final resize	['a', 'b', 'b']	$v_{31} \sim v_{33}$	-	-
S_2	Blur	kernel size $[2m+1]$	$m \in [3, 10]$	v_1	-	-
		standard deviation σ_1	$[0.2, 1.5]$	v_2	-	-
		standard deviation σ_2	$[0.2, 1.5]$	v_3	-	-
		rotation degree θ	$[-\pi, \pi]$	v_4	-	-
	Resize	[up, down, keep]	$[0.3, 0.4, 0.3]$	-	-	-
		scale factor	$[0.5, 1.2]$	v_{11}	-	-
		resize mode	['a', 'b', 'b']	$v_{12} \sim v_{14}$	-	-
	Noise	type	['G', 'P']	v_{21}, v_{22}	-	-
		sigma of Gaussian	$[1, 20]$	v_{19}	-	-
		scale of Poisson	$[0.05, 1.5]$	v_{19}	-	-
		gray probability	0.4	v_{20}	-	-
	JPEG	quality factor	$[50, 95]$	v_{27}	-	-
		mode of final resize	['a', 'b', 'b']	$v_{31} \sim v_{33}$	-	-
S_3	Blur	kernel size $[2m+1]$	$m \in [3, 10]$	v_1	$m \in [3, 10]$	v_5
		standard deviation σ_1	$[0.2, 3]$	v_2	$[0.2, 1.5]$	v_6
		standard deviation σ_2	$[0.2, 3]$	v_3	$[0.2, 1.5]$	v_7
		rotation degree θ	$[-\pi, \pi]$	v_4	$[-\pi, \pi]$	v_8
		sinc kernel size $[2m+1]$	-	-	$m \in [3, 10]$	v_9
		ω_c of sinc kernel	-	-	$[\pi/3, \pi]$	v_{10}
	Resize	[up, down, keep]	$[0.2, 0.7, 0.1]$	-	$[0.3, 0.4, 0.3]$	-
		scale factor	$[0.15, 1.5]$	v_{11}	$[0.3, 1.2]$	v_{15}
		resize mode	['a', 'b', 'b']	$v_{12} \sim v_{14}$	['a', 'b', 'b']	$v_{16} \sim v_{18}$
	Noise	type	['G', 'P']	v_{21}, v_{22}	['G', 'P']	v_{25}, v_{26}
		sigma of Gaussian	$[1, 30]$	v_{19}	$[1, 25]$	v_{23}
		scale of Poisson	$[0.05, 3]$	v_{19}	$[0.05, 2.5]$	v_{23}
		gray probability	0.4	v_{20}	0.4	v_{24}
	JPEG	quality factor	$[30, 95]$	v_{27}	$[30, 95]$	v_{28}
		operating order	-	-	R-J or J-R	v_{29}, v_{30}
		mode of final resize	-	-	['a', 'b', 'b']	$v_{31} \sim v_{33}$

For those operations that have more than one mode, *e.g.*, the resize mode, we use a one-hot vector to indicate the choice of mode in $\boldsymbol{v}$. For other parameters, we normalize each of them by $v' = (v - v_{\min})/(v_{\max} - v_{\min})$, where $v, v', v_{\min}$ and $v_{\max}$ indicate the original value, the normalized value, the minimum and maximum values of the parameter, respectively.

3 More Sample Images

In Fig. 1, we provide more sample images with different degradation levels in our datasets, as well as the ground-truth HR images. As can be seen from the figure, those images can cover a wide range of real-world degradations. The balanced sampling from the three levels during training improves the generalization capacity of our DASR to real-world images with different degradations.

4 More Qualitative Comparisons

In Fig. 2, we provide more qualitative comparisons of competing methods on real-world images, while in Figs. 3, 4, 5 and 6, we provide more qualitative comparisons of competing methods on datasets with bicubic, Level-I, Level-II and Level-III degradations, respectively. Our models are trained by using the images in DIV2K, Flickr2K, and OutdoorScene-Training datasets. To further validate the generalization capability of DASR to different image contents, the visual comparisons in Figs. 3, 4, 5 and 6 also include images from the Urban100 dataset by using the same degrading strategy as in our main paper. From those figures, consistent observations to our main paper can be made. Our DASR can generate more realistic structures and details on different degradations, benefiting from its degradation-adaptive strategy and the joint training and adaptive mixture of multiple experts.

5 More Ablation Studies

We provide more quantitative ablation study results in this section.

In Table 2, we directly merge 5 SRResNet models into one large model, and compare its results against our DASR on Level-I and II datasets. As shown in the table, the merged model achieves comparable performance yet consumes much more cost (*e.g.*, about 3 times the latency and 6 times the #FLOPs), which validates the efficiency of the proposed method.

In Table 3, we plug DASR into the EDSR backbone and report the results on Level-I and II datasets. One can see the clear improvements of DASR over EDSR baseline in both PSNR and LPIPS, demonstrating the effectiveness of the proposed method.

In Table 4, we report more quantitative ablation results of DASR. The observations are consistent with Fig. 4 in the main paper. In specific, we first evaluate the selection of N, *i.e.*, the number of expert models. As shown in the

Table 2. Quantitative comparison between DASR ($N = 5$) and the large model directly merged from 5 models on datasets of Level-I and II. The merged model achieves comparable performance yet consumes much more cost, which validates the effectiveness of the proposed method.

Metrics	PSNR	LPIPS	Latency	#FLOPs	#Params	#Memory
Merged-I	27.79	0.1713	451ms	1258G	7.60M	2377M
DASR-I	27.84	0.1707	142ms	184G	8.07M	2452M
Merged-II	27.52	0.2111	451ms	1258G	7.60M	2377M
DASR-II	27.58	0.2126	142ms	184G	8.07M	2452M

Table 3. Quantitative comparison of Level-I and II datasets by plugging DASR into the EDSR backbone. The improvements of DASR over EDSR baseline in both PSNR and LPIPS demonstrate the effectiveness of the proposed method.

Metrics	PSNR	LPIPS	Latency	#FLOPs	#Params	#Memory
EDSR-I	27.79	0.1834	105ms	130G	1.52M	2169M
DASR-I	27.94	0.1736	134ms	148G	8.07M	2262M
EDSR-II	27.53	0.2284	105ms	130G	1.52M	2169M
DASR-II	27.71	0.2162	134ms	148G	8.07M	2262M

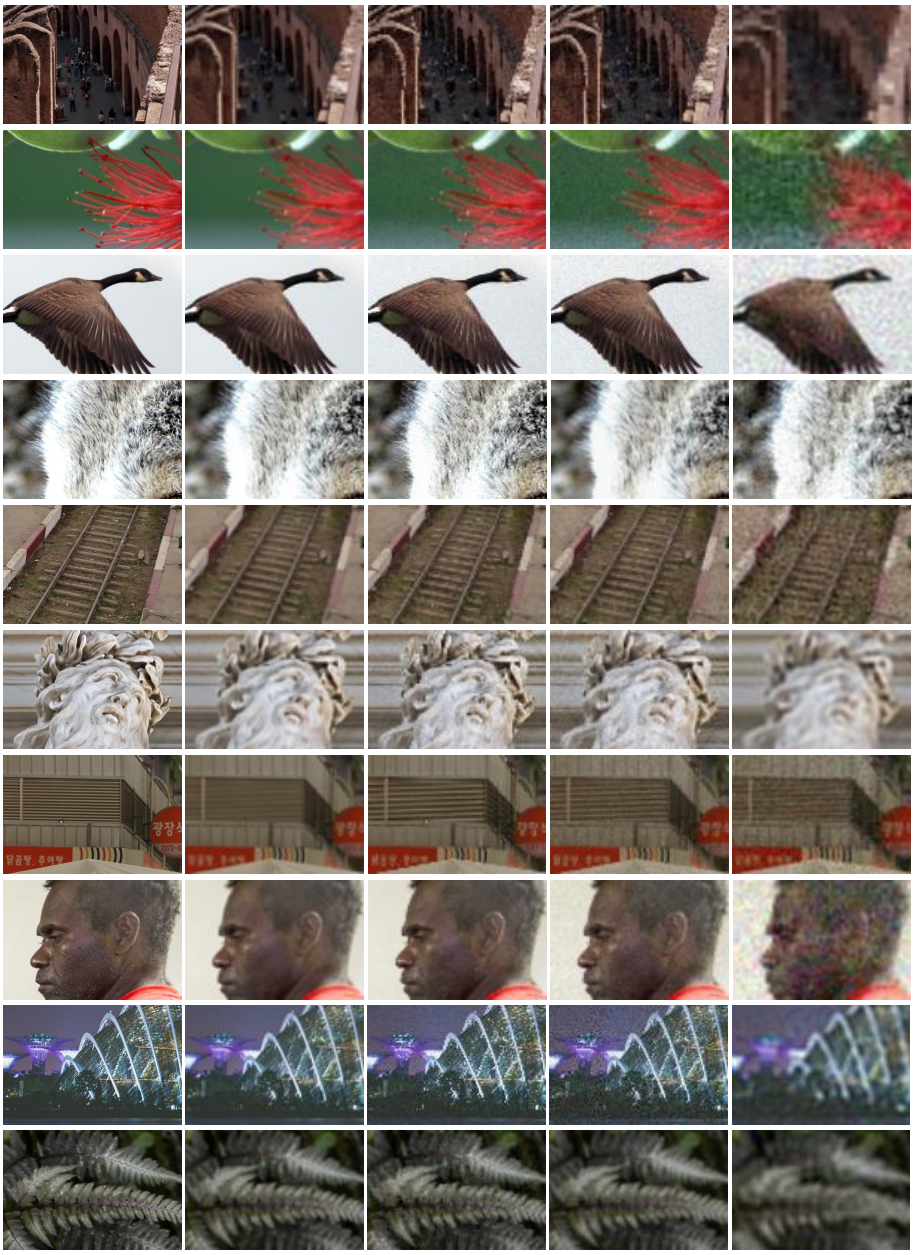
table, $N = 3$ may be insufficient to achieve good fidelity and perceptual quality, while $N = 9$ shows similar performance to $N = 5$ in Table 1 and 2 in the main paper. The columns 'w/sig' and 'f-fuse' evaluate the effectiveness of our model design, 'w/sig' denotes adding a sigmoid layer to the weighting module $\mathcal{A}$ and 'f-fuse' indicates the feature fusion strategy in traditional MoE. Neither of them shows good perceptual quality according to the LPIPS. The column 'multiply' performs dynamic convolution with a single expert by learning a mapping matrix and multiplying it to the parameters, while the column 'DCD' is another dynamic strategy as in [2]. Both strategies can hardly induce satisfactory results.

Table 4. Quantitative ablation results on the dataset of Level-I, which show consistent observations with Fig. 4 in the main paper.

Methods	N=3	N=9	w/ sig	f-fuse	multiply	DCD	EDSR
PSNR	27.76	27.82	27.91	27.88	27.71	27.02	27.94
LPIPS	0.1723	0.1698	0.1854	0.1843	0.1715	0.2135	0.1736

References

1. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1
2. Yunsheng Li, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Ye Yu, Lu Yuan, Zicheng Liu, Mei Chen, and Nuno Vasconcelos. Revisiting dynamic convolution via matrix decomposition. In *ICLR*, 2021. 4



(a) HR (b) Bicubic (c) Level-I (d) Level-II (e) Level-III

Fig. 1. More sample images with different levels of degradations in our constructed datasets, as well as the ground-truth HR images. Level-I, -II, and -III represent the samples whose degradations belong to S_1, S_2, and S_3, respectively.

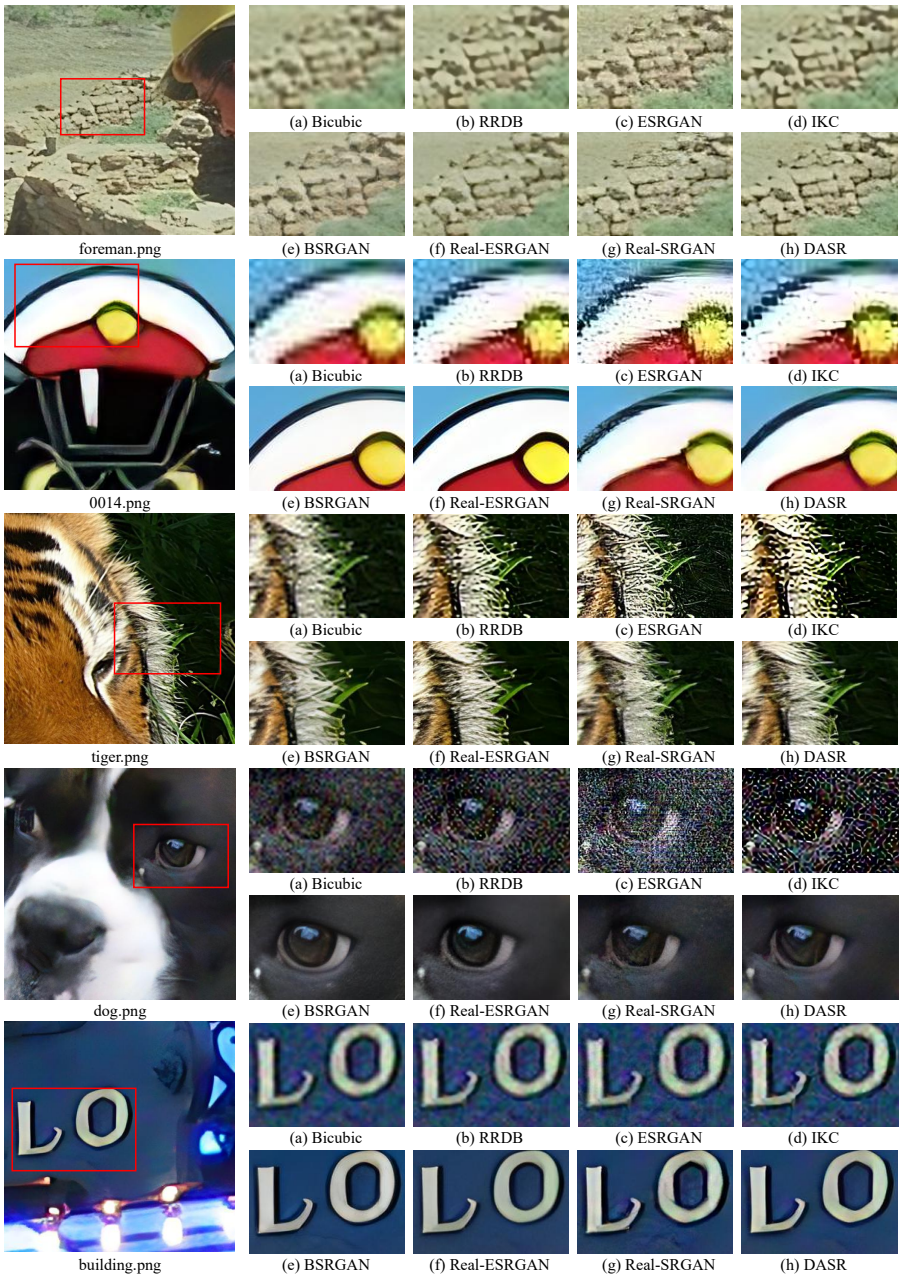


Fig. 2. More qualitative comparison of competing methods on **real-world** images. The results of (b-f) are generated by using the officially released models, while the output of (g) is obtained by re-training the SRResNet backbone with our proposed degradation model. Better zoom in for details.

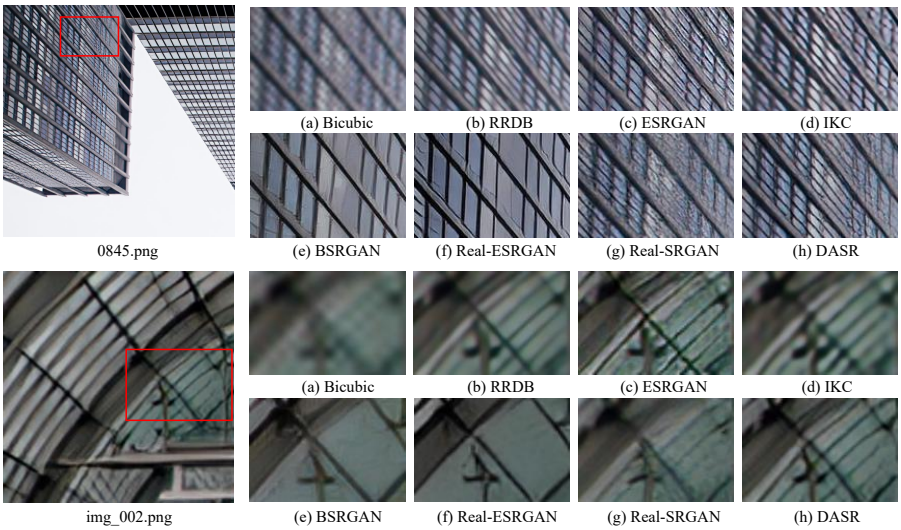


Fig. 3. More qualitative comparison of competing methods on images with **bicubic** downsampling. The results of (b-f) are generated by using the officially released models, while the output of (g) is obtained by re-training the SRResNet backbone with our proposed degradation model. Better zoom in for details.

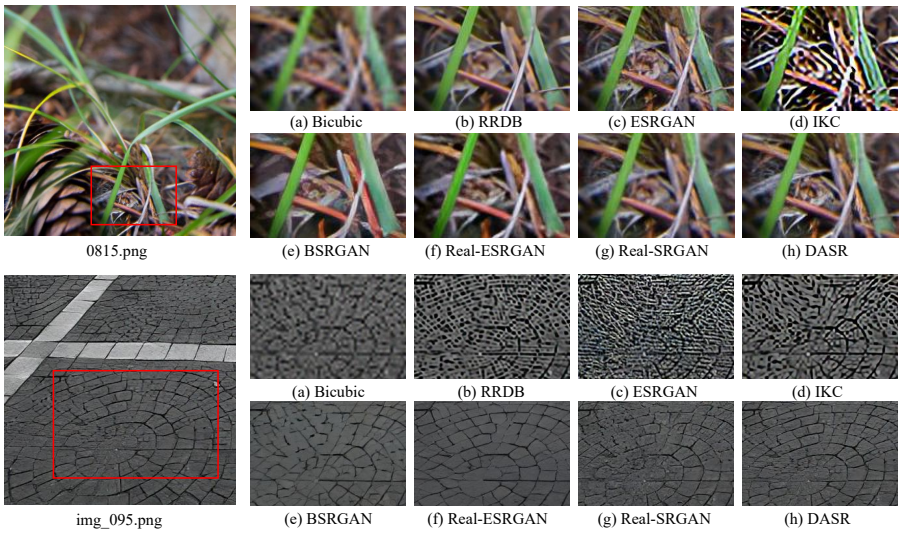


Fig. 4. More qualitative comparison of competing methods on images with degradation of **Level-I**. The results of (b-f) are generated by using the officially released models, while the output of (g) is obtained by re-training the SRResNet backbone with our proposed degradation model. Better zoom in for details.

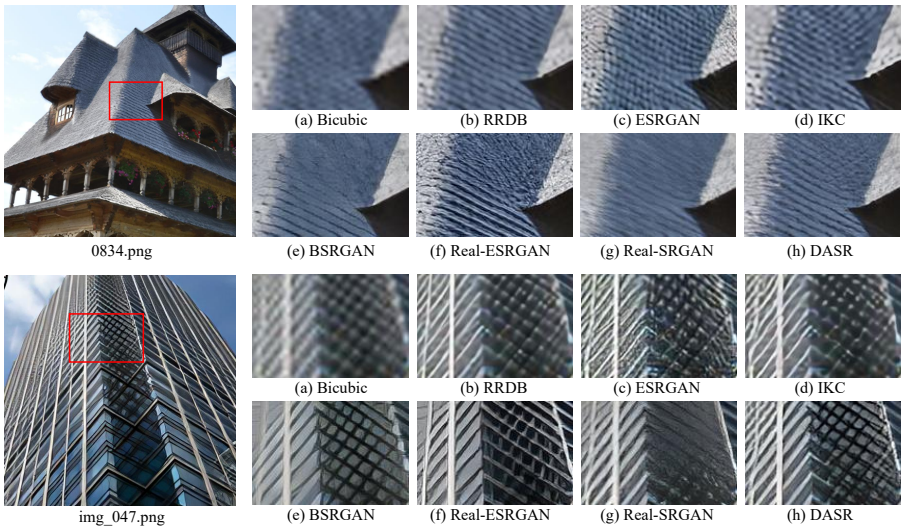


Fig. 5. More qualitative comparison of competing methods on images with degradation of **Level-II**. The results of (b-f) are generated by using the officially released models, while the output of (g) is obtained by re-training the SRResNet backbone with our proposed degradation model. Better zoom in for details.

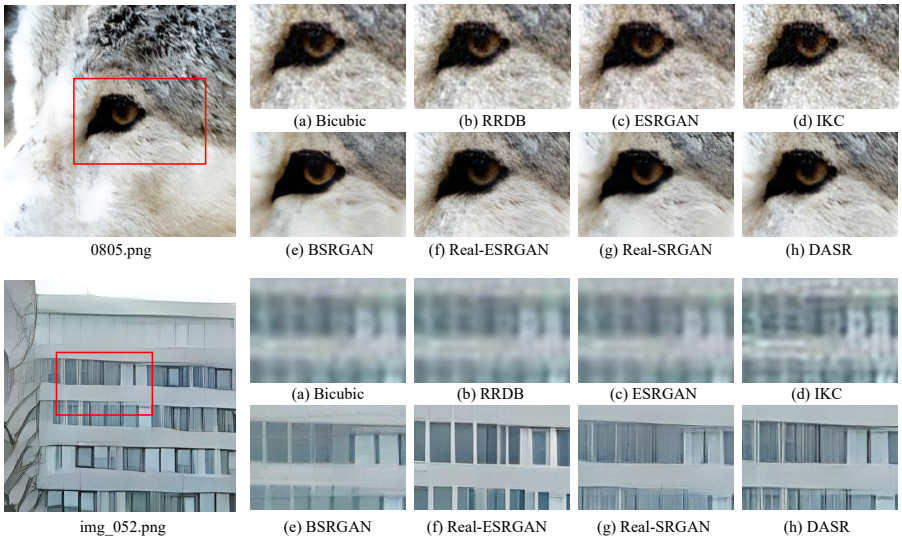


Fig. 6. More qualitative comparison of competing methods on images with degradation of **Level-III**. The results of (b-f) are generated by using the officially released models, while the output of (g) is obtained by re-training the SRResNet backbone with our proposed degradation model. Better zoom in for details.