Generalized Real-World Super-Resolution through Adversarial Robustness –Supplementary Material–

Angela Castillo*[1], María Escobar*[1], Juan C. Pérez[1,2], Andrés Romero[3], Radu Timofte[3], Luc Van Gool[3],
and Pablo Arbelaez[1]

[1]Center for Research and Formation in Artificial Intelligence, Universidad de los Andes, Colombia
[2]King Abdullah University of Science and Technology (KAUST), Saudi Arabia
[3]Computer Vision Lab, ETH Zürich, Switzerland
[1]{a.castillo13, mc.escobar11, jc.perez13, pa.arbelaez}@uniandes.edu.co
[3]roandres@ethz.ch, {radu.timofte, vangool}@vision.ee.ethz.ch

1. Additional quantitative results

We present non-reference quantitative comparisons for the state-of-the-art methods on the NTIRE$_{syn}$, AIM$_{syn}$, DPED$_{rw}$, and FACES$_{rw}$ datasets. Table 1 and Table 2 show that our RSR model outperforms state-of-the-art models on PIQE and is in a competitive range for BRISQUE and NIQE. This result confirms the qualitative finding reported on our main paper, that our RSR model creates HR images with a good perceptual quality.

Method	Training Dataset	**BRISQUE↓**			**NIQE↓**			**PIQE↓**		
		NTIRE$_{syn}$	AIM$_{syn}$	*Avg*	NTIRE$_{syn}$	AIM$_{syn}$	*Avg*	NTIRE$_{syn}$	AIM$_{syn}$	*Avg*
Bicubic	-	55.37	56.58	55.98	5.74	5.96	5.85	84.02	87.12	85.57
Impressionism [2]	NTIRE$_{syn}$	13.32	22.66	17.99	3.15	2.51	2.83	17.47	31.10	24.28
	AIM$_{syn}$	36.04	17.08	26.56	5.75	3.53	4.64	47.67	20.31	33.99
	DPED$_{rw}$	38.97	34.37	36.67	4.04	3.35	3.70	25.82	29.18	27.50
ESRGAN-FS [1]	NTIRE$_{syn}$	27.16	40.77	33.96	2.76	3.71	3.23	34.54	64.97	49.75
	AIM$_{syn}$	27.60	15.98	21.79	4.40	2.68	3.54	38.07	20.30	29.19
	DPED$_{rw}$	32.95	34.71	33.83	3.39	3.28	3.33	30.46	37.62	34.04
ESRGAN [4]	DIV2K	35.08	28.33	31.70	6.27	2.67	4.47	50.72	39.99	45.36
RSR (Ours)	DIV2K	16.47	19.24	17.86	3.65	3.37	3.51	19.77	17.68	18.72

Table 1. Non-reference metrics comparing our method and the state-of-the-art methods in NTIRE$_{syn}$ and AIM$_{syn}$ datasets. ↓ indicates lower is better. Red and blue colors highlight the best two scores.

2. Additional qualitative results

In Fig. 1 and Fig. 2 we include further qualitative results for the NTIRE$_{syn}$, AIM$_{syn}$, DPED$_{rw}$, and FACES$_{rw}$ datasets. Notice on Fig. 1 that our single robust model is able to enhance diverse types of images, including difficult textures like the squirrel's fur (second row) or the pattern in the military uniform (fourth row). Additionally, we further confirm that state-of-the-art models tend to underperform when evaluated on unseen datasets (red frames). Fig. 2 provides additional examples of the effectiveness of our model in removing noise. In particular, for the FACES$_{rw}$ dataset, our method removes the noise in the input image without creating more artifacts.

*equal contribution

Method	Training Dataset	**BRISQUE ↓**			**NIQE ↓**			**PIQE ↓**		
		DPED$_{rw}$	FACES$_{rw}$	*Avg*	DPED$_{rw}$	FACES$_{rw}$	*Avg*	DPED$_{rw}$	FACES$_{rw}$	*Avg*
Impressionism [2]	NTIRE$_{syn}$	22.55	52.77	37.66	2.83	5.41	4.12	12.21	46.23	29.22
	AIM$_{syn}$	21.33	52.51	36.92	4.12	5.80	4.96	25.09	34.85	29.97
	DPED$_{rw}$	23.35	20.43	21.89	4.13	3.21	3.67	14.03	16.63	15.33
ESRGAN-FS[1]	NTIRE$_{syn}$	46.12	64.16	55.14	4.83	5.82	5.33	48.66	92.00	70.33
	AIM$_{syn}$	15.51	47.54	31.53	3.75	5.78	4.77	14.99	28.91	21.95
	DPED$_{rw}$	22.94	48.56	35.75	3.21	5.05	4.13	12.06	34.22	23.14
RSR (Ours)	DIV2K	21.72	35.93	28.83	5.34	6.32	5.83	9.59	12.34	10.97

Table 2. Non-reference comparison for the DPED$_{rw}$ and FACES$_{rw}$ datasets. ↓ indicates that lower is better. Red and Blue highlights the best and the second best score, respectively.

3. Limitation of Existing Literature

As we extensively discuss in the main paper, methods that perform well on a dataset with a specific corruption fail on unseen artifacts. In this section, we qualitative highlight those findings. First, on Fig. 3 we find that the ESRGAN-FS model trained on AIM$_{syn}$ generates images with an unrealistic color intensity and hallucinates textures in the places where the input images are noisy. This color hallucination might happen because of the corruptions in the AIM$_{syn}$ dataset, as LR images have less intense colorization due to strong compression artifacts. Therefore, the model learns to counteract this corruption. In contrast, as out our single RSR model bypass learning on specific corruptions, we faithfully super-resolve the LR image without creating unrealistic colors. Next, Fig. 4 shows that the Impressionism model trained on NTIRE$_{syn}$ increases the JPEG compression artifacts present in AIM$_{syn}$ images, whereas our model is able to remove it.

Fig. 5 and 6 show the effect that Impressionism trained on DPED$_{rw}$ has on NTIRE$_{syn}$ and AIM$_{syn}$ respectively. The model hallucinates sharp details on incorrect parts of the images and creates unrealistic super-resolved images if the input includes texture. Furthermore, for NTIRE$_{syn}$ the model creates a sharper and more noticeable noise instead or removing it. Finally, in Fig. 7 we find that Impressionism trained on DPED$_{rw}$ transforms the noise of FACES$_{rw}$ images into very strong and unrealistic noise.

4. Additional Baseline

To assess the superior capacity of our method, we propose a new baseline. We aim to determine if it is possible to achieve better results if we denoise the real-world input image with a state-of-the-art method on adversarial attacks and then super-resolve the result with a pre-trained network (in this case, we use the pre-trained baseline). For the denoiser, we use the winning method [3] in the competition *NIPS 2017: Defense Against Adversarial Attack* which achieved remarkable results on denoising adversarially-attacked images. Table 3 shows that even if we use a method that is specialized in removing adversarial noise, it is not enough to successfully remove the noise to reach a good result after super-resolving the input image. Furthermore, our method outperforms this baseline, confirming the importance of the loss function to optimize, as explained in the main paper.

Method	**PSNR↑**			**SSIM↑**			**LPIPS↓**		
	NTIRE$_{syn}$	AIM$_{syn}$	*Avg*	NTIRE$_{syn}$	AIM$_{syn}$	*Avg*	NTIRE$_{syn}$	AIM$_{syn}$	*Avg*
Guided denoise [3]	11.99	12.10	12.04	0.22	0.23	0.22	0.60	0.59	0.60
RSR (Ours)	**24.31**	**21.99**	**23.15**	**0.65**	**0.60**	**0.62**	**0.23**	**0.37**	**0.30**

Table 3. Comparison of our method against a denoising method for adversarially-attacked images. ↑ and ↓ indicate higher is better and lower is better, respectively. Best results are presented in **bold**.

5. Mixture of types of noise

To further confirm the generalization capacity of our framework we explore the use of different type of noise for real-world SR. In particular, following [1], we model sensor noise as Gaussian noise with zero mean and a standard deviation of 8 and compression artifacts by converting the images to JPEG with a quality of 30. We performed two baseline experiments in which we train 18k iterations only with images modified with each type of degradation. Furthermore, we perform three mixed experiments in which each network iteration is trained with a different type of corruption: robust training and sensor noise, robust training and compression artifacts and finally the mixture of all three types of degradation.

Table 5. Non-reference comparison for 100 random images of the train set of NTIRE$_{syn}$ and AIM$_{syn}$. ↓ indicates that lower is better. Red highlights the best score.

Method	**NIQE ↓**		
	NTIRE$_{syn}$	AIM$_{syn}$	*Avg*
ESRGAN [4]	6.91	3.38	5.14
RSR (Ours)	4.65	4.90	4.77

Robust Training	Sensor Noise	Compression artifacts	**PSNR↑**			**SSIM↑**			**LPIPS↓**		
			NTIRE$_{syn}$	AIM$_{syn}$	Avg	NTIRE$_{syn}$	AIM$_{syn}$	Avg	NTIRE$_{syn}$	AIM$_{syn}$	Avg
✓			24.31	21.99	23.15	0.65	0.60	0.62	**0.23**	**0.37**	**0.30**
	✓		26.08	22.39	24.24	0.72	0.63	0.68	0.24	0.39	0.32
		✓	18.68	19.09	18.89	0.30	0.35	0.33	0.61	0.54	0.58
✓	✓		25.43	21.60	23.52	0.69	0.58	0.64	0.24	**0.37**	0.31
✓		✓	24.83	19.30	22.07	0.67	0.35	0.51	0.26	0.56	0.41
✓	✓	✓	25.46	19.59	22.53	0.69	0.38	0.54	0.24	0.38	0.31

Table 4. Comparison of different types of noise for training the SR model. Best results for LPIPS are presented in **bold**.

Table 4 shows that simulating sensor noise results in a better performance that training with simulated compression artifacts. This phenomenon might be explained by the fact that compression artifacts make stronger modifications to the original image, making the examples less useful for the network. Likewise, using compression artifacts in a mixture with robust training is detrimental for the generalization capability of the network. However, using a mixture of the three types of degradation results in an overall performance similar to our RSR. Thus, both sensor noise and compression artifacts work in a complementary way with our robust training scheme.

6. Additional Ablation Analysis

6.1. Visualization of adversarial attack hyper-parameters

As the core of our model is to introduce adversarial noise that can potentially resemble real-world artifacts, we visualize the optimized noise and the adversarial example for a training batch in each part of our ablation study. Fig. 8 shows that changing the loss function for adversarial optimization does not reflect a significant perceptual difference in the noise that is added to the images. However, the results in our main paper suggest that this imperceptible change helps in the improvement of average LPIPS for our model.

On the one hand, Fig. 9 illustrates that, the higher the ϵ used in the attack, the stronger the noise that is present in the adversarial examples. This stronger noise can be better appreciated on the fourth column in the adversarial examples of the figure. On the other hand, Fig. 10 depicts that there is not a perceptual variation in the noise of the adversarial example when we increase the iterations of the attack. This result is consistent with the quantitative results presented in our main paper where we found that increasing the iterations does not have an effect on the final performance of the model.

Finally, Fig. 11 shows that increasing the scale of the structured noise creates a more aggregated optimized noise. This result is also visualized on the adversarial examples. According to the quantitative results presented in our main paper, a scale of 1.5 gives the ideal trade-off between having a hard but realistic adversarial example.

6.2. Quantitative Ablation on Training Set

We perform a quantitative comparison between our RSR model and ESRGAN [4], our baseline model, on 100 random images from the training sets of NTIRE$_{syn}$ and AIM$_{syn}$. Since there is no HR ground-truth available for the training set, we use the non-reference quantitative metric NIQE. Table 5 shows that, on average, our model outperforms ESRGAN in perceptual quality on the training sets.

References

[1] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3599–3608. IEEE, 2019. 1, 2, 6

[2] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 466–467, 2020. 1, 2, 6

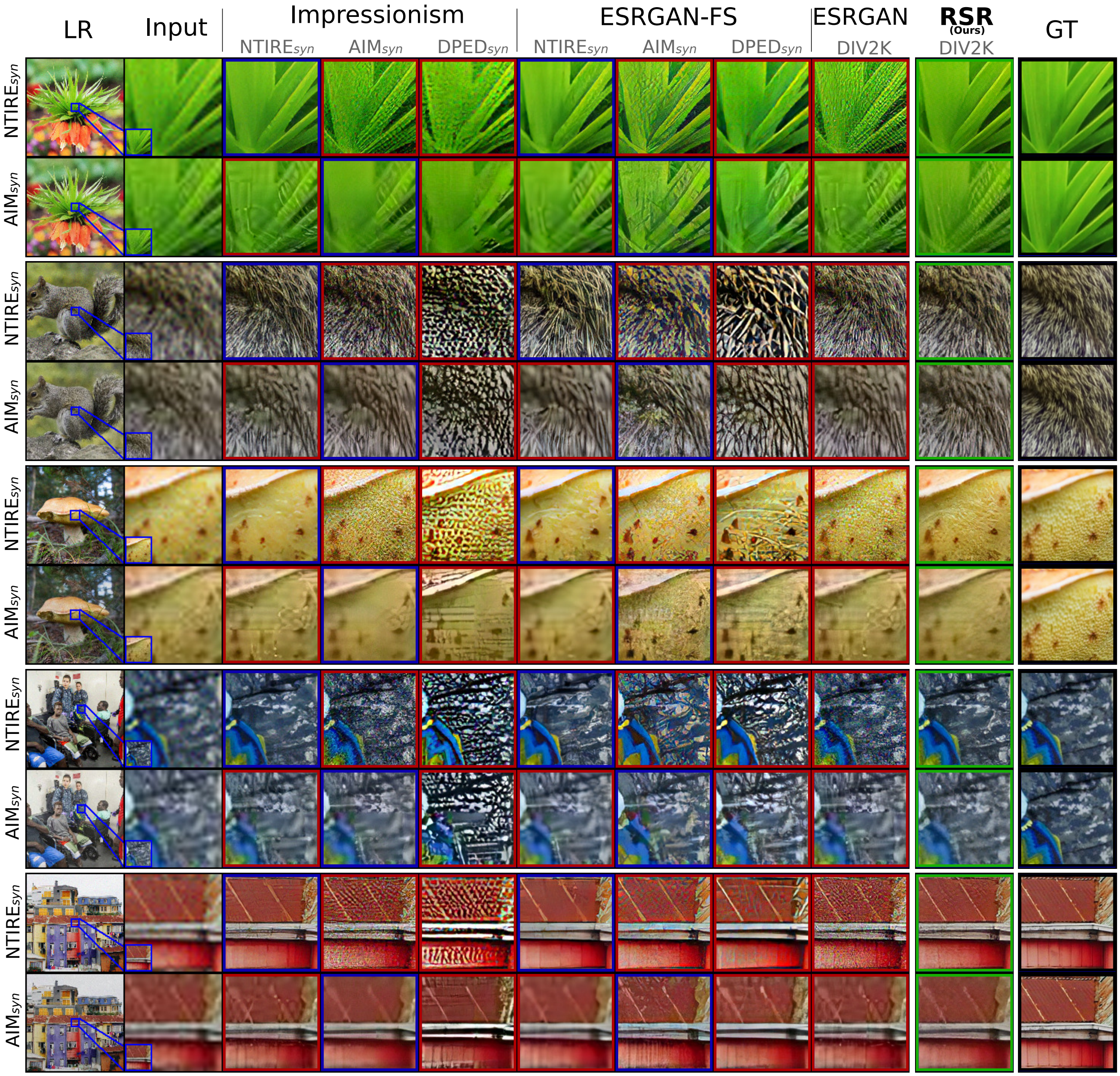

Figure 1. **Additional results on Synthetic images.** Comparison between our method and state-of-the-art methods, for two synthetic corruption datasets: NTIRE$_{syn}$ and AIM$_{syn}$. For reference, we show the bicubically upsampled input, the result of a supervised SISR method (ESRGAN [4]), and the ground-truth (GT). Blue frames denote training and validation on the same dataset. Red frames denote training and validation on different datasets. Green frames denote our method.

[3] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2018. 2

[4] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1, 3, 4, 5

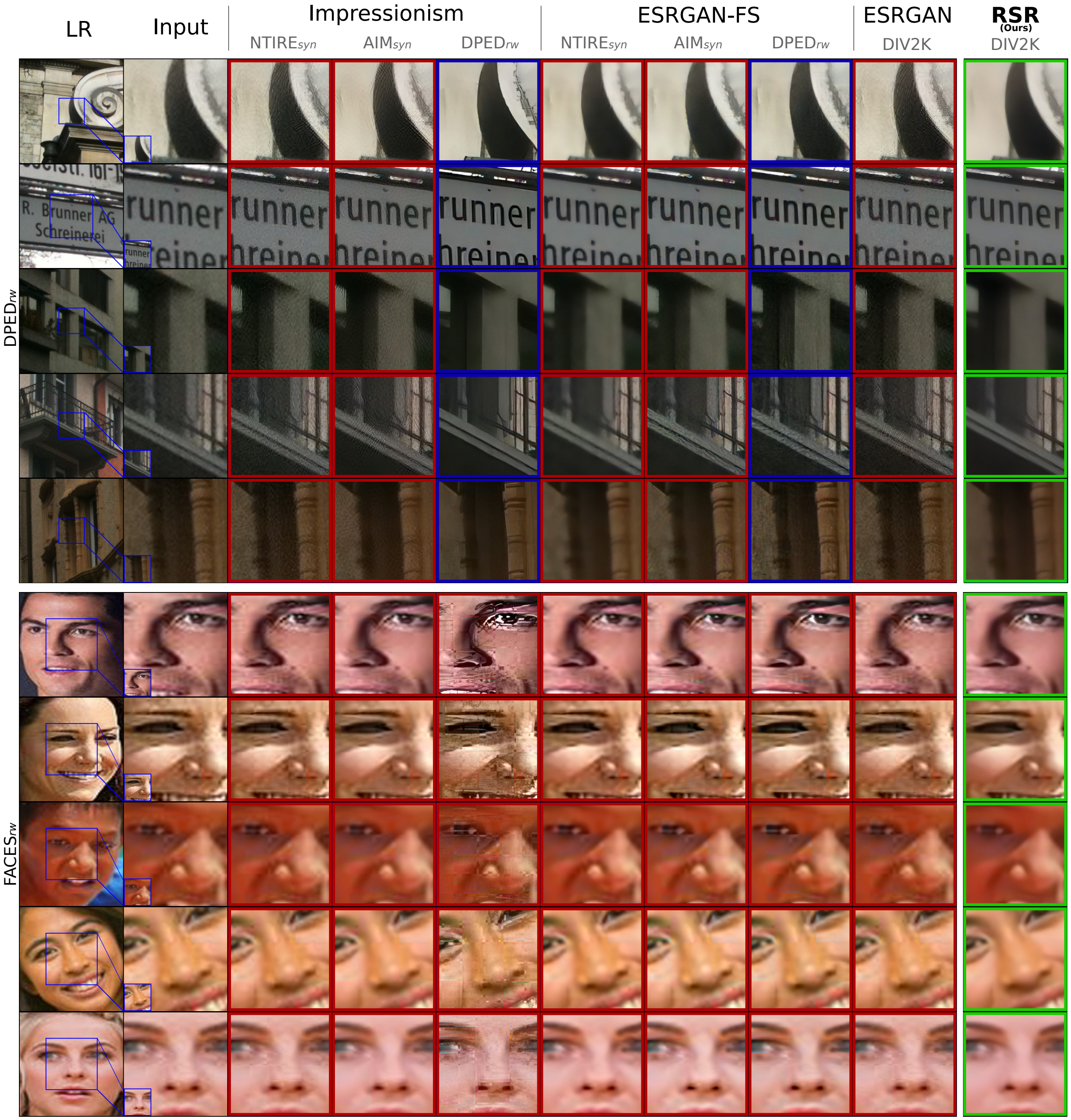

Figure 2. **Additional results on real-world images.** Comparison between our method and state-of-the-art methods, for two real-world datasets: DPED$_{rw}$ and FACES$_{rw}$. For reference, we show the bicubically upsampled input, the result of a supervised SISR method (ESRGAN [4]), and the ground-truth (GT). Blue frames denote training and validation on the same dataset. Red frames denote training and validation on different datasets. Green frames denote our method.

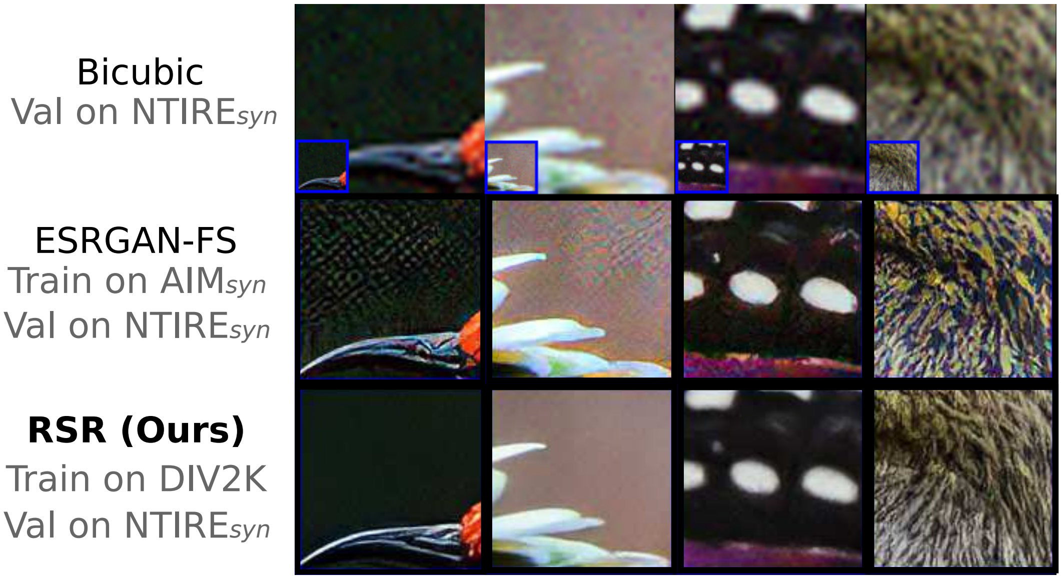

Figure 3. **Artifacts enforced on the NTIRE$_{syn}$ dataset** Qualitative comparison between the artifacts created by ESRGAN-FS [1] trained on AIM$_{syn}$ and our RSR method trained on DIV2K. For reference, we show the bicubically upsampled input. Note that ESRGAN-FS hallucinates the color intensity of the image.

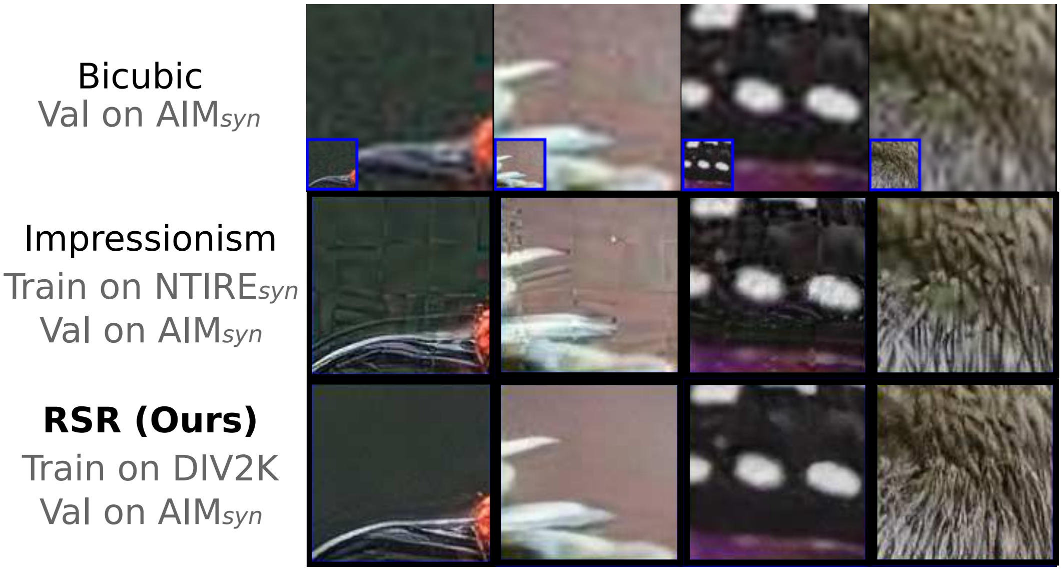

Figure 4. **Artifacts enforced on the AIM$_{syn}$ dataset** Qualitative comparison between the artifacts created by Impressionism [2] trained on NTIRE$_{syn}$ and our RSR method trained on DIV2K. For reference, we show the bicubically upsampled input. Note that Impressionism increases the JPEG noise that AIM$_{syn}$ includes instead of removing it.

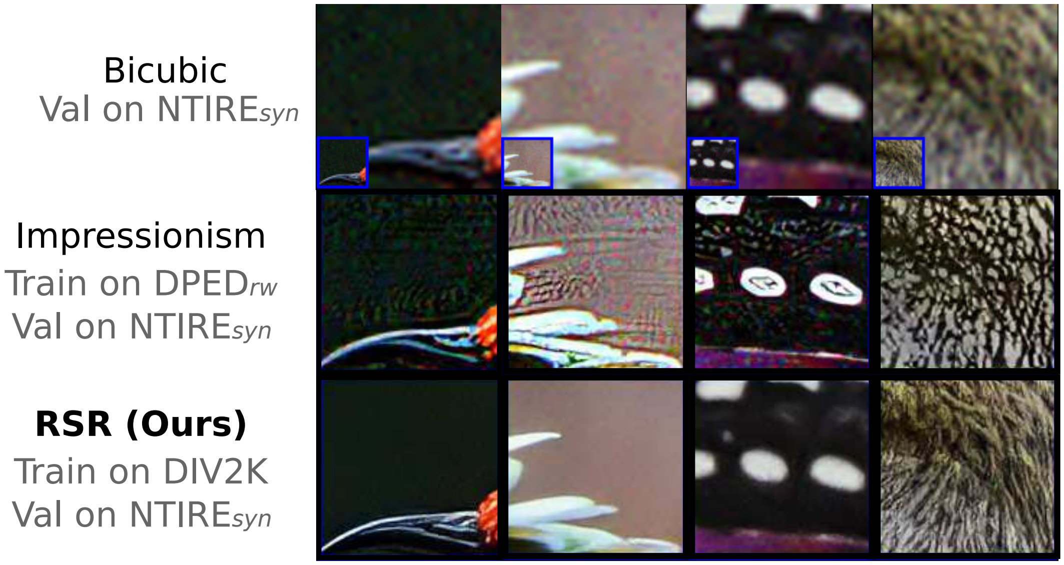

Figure 5. **Artifacts enforced on the NTIRE$_{syn}$ dataset with training on DPED$_{rw}$** Qualitative comparison between the artifacts created by Impressionism trained on DPED$_{rw}$ and our RSR method trained on DIV2K. For reference, we show the bicubically upsampled input. Note that Impressionism hallucinates sharp details where it should remove noise.

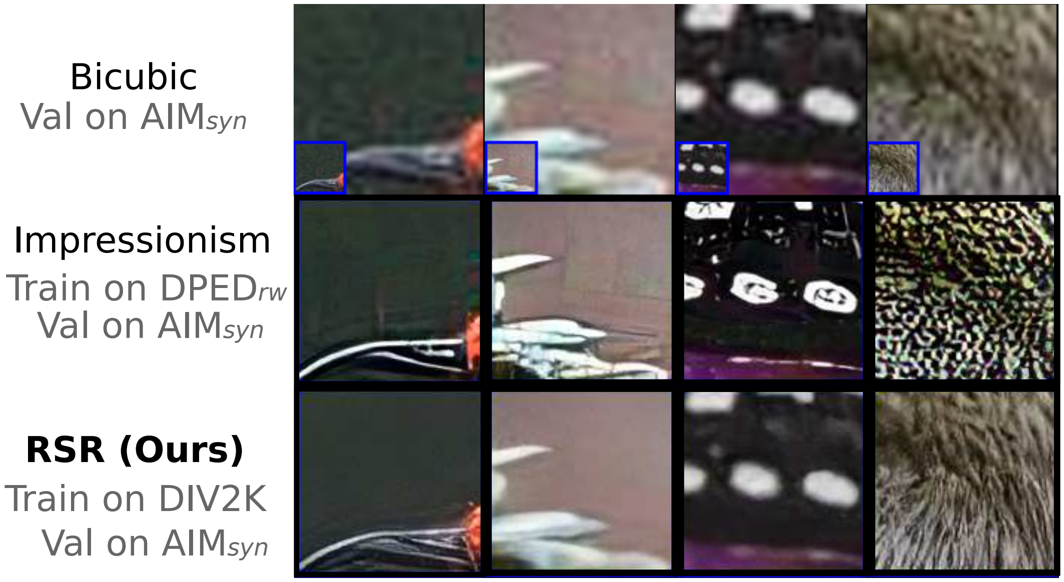

Figure 6. **Artifacts enforced on the AIM$_{syn}$ dataset with training on DPED$_{rw}$** Qualitative comparison between the artifacts created by Impressionism trained on DPED$_{rw}$ and our RSR method trained on DIV2K. For reference, we show the bicubically upsampled input. Note that Impressionism hallucinates sharp details on the image.

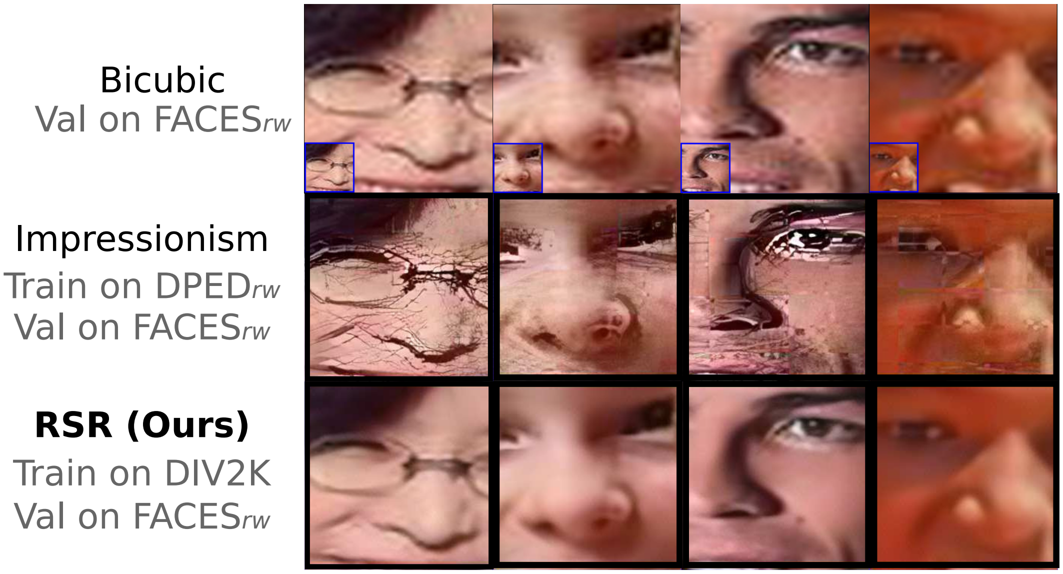

Figure 7. **Artifacts enforced on the FACES$_{rw}$ dataset with training on DPED$_{rw}$** Qualitative comparison between the artifacts created by Impressionism trained on DPED$_{rw}$ and our RSR method trained on DIV2K. For reference, we show the bicubically upsampled input. Note that Impressionism creates unrealistic super-resolved images in comparison with our results.

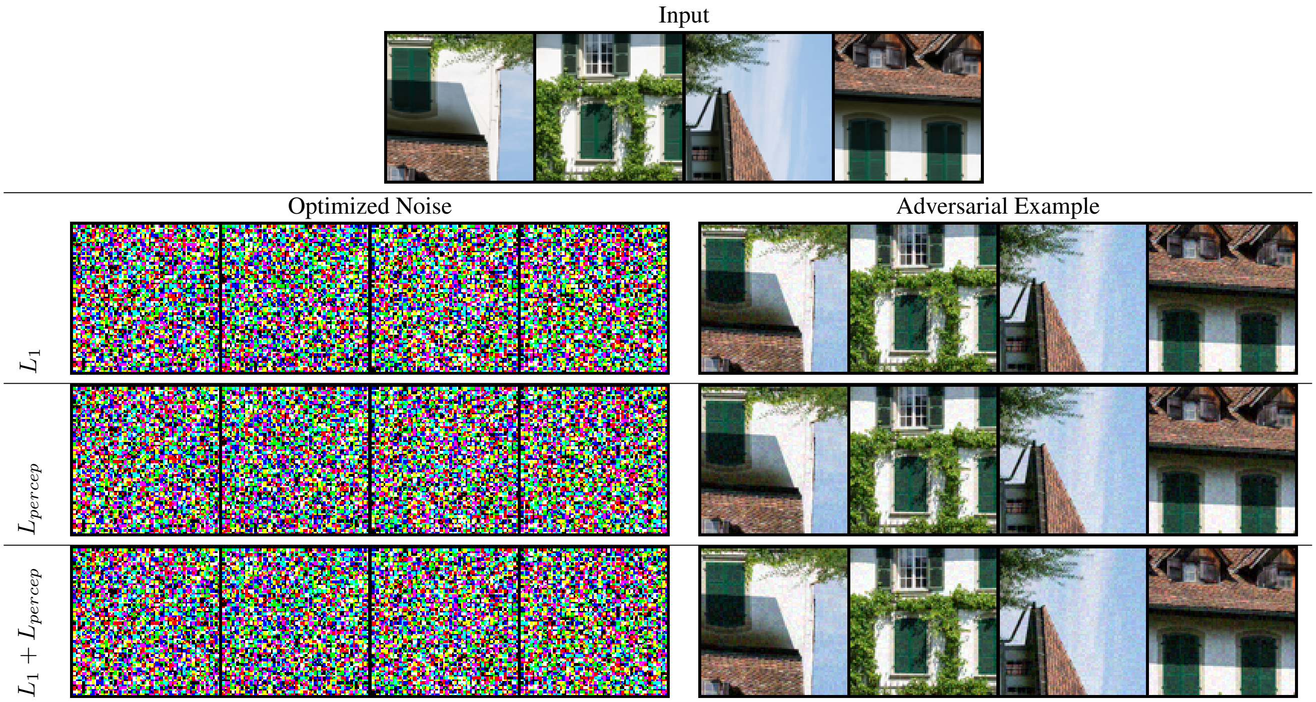

Figure 8. **Qualitative visualization of the loss ablation.** Optimized noise and adversarial examples for the loss function used in the adversarial attack's ablations.

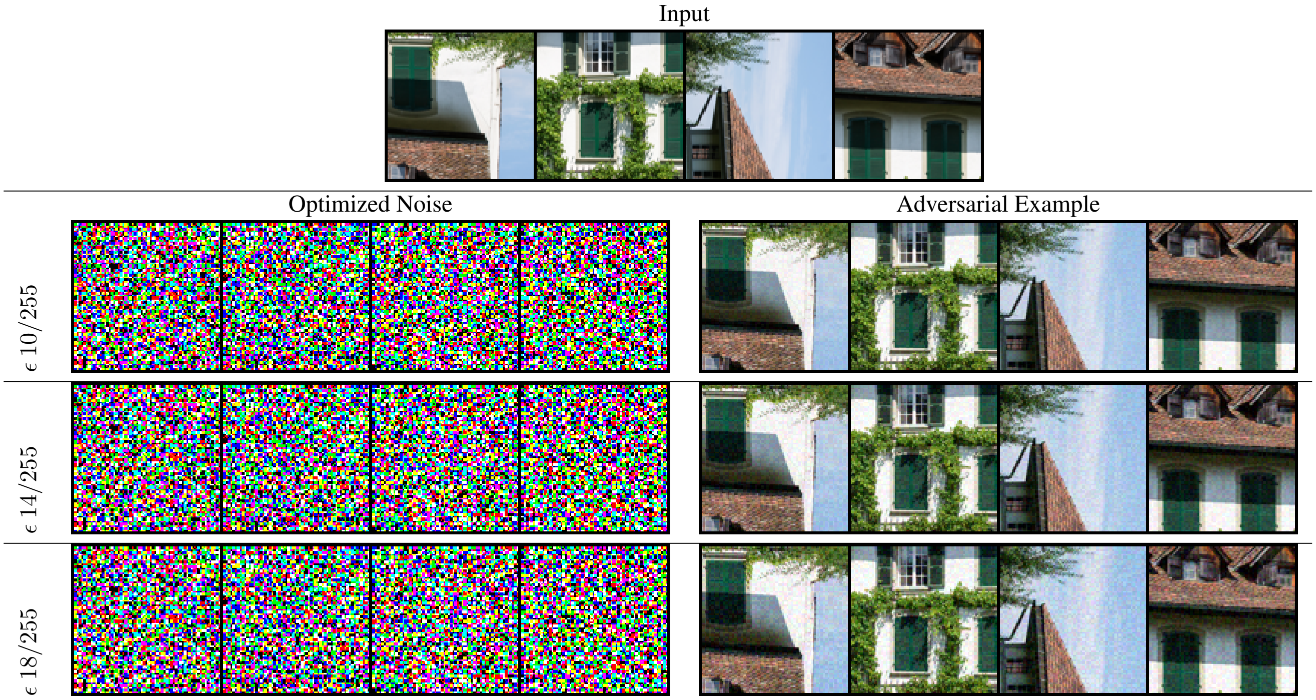

Figure 9. **Qualitative visualization of the ϵ ablation.** Optimized noise and adversarial examples for different ϵ in the adversarial attack's ablations. Note that, the higher the ϵ, the stronger the noise in the adversarial examples.

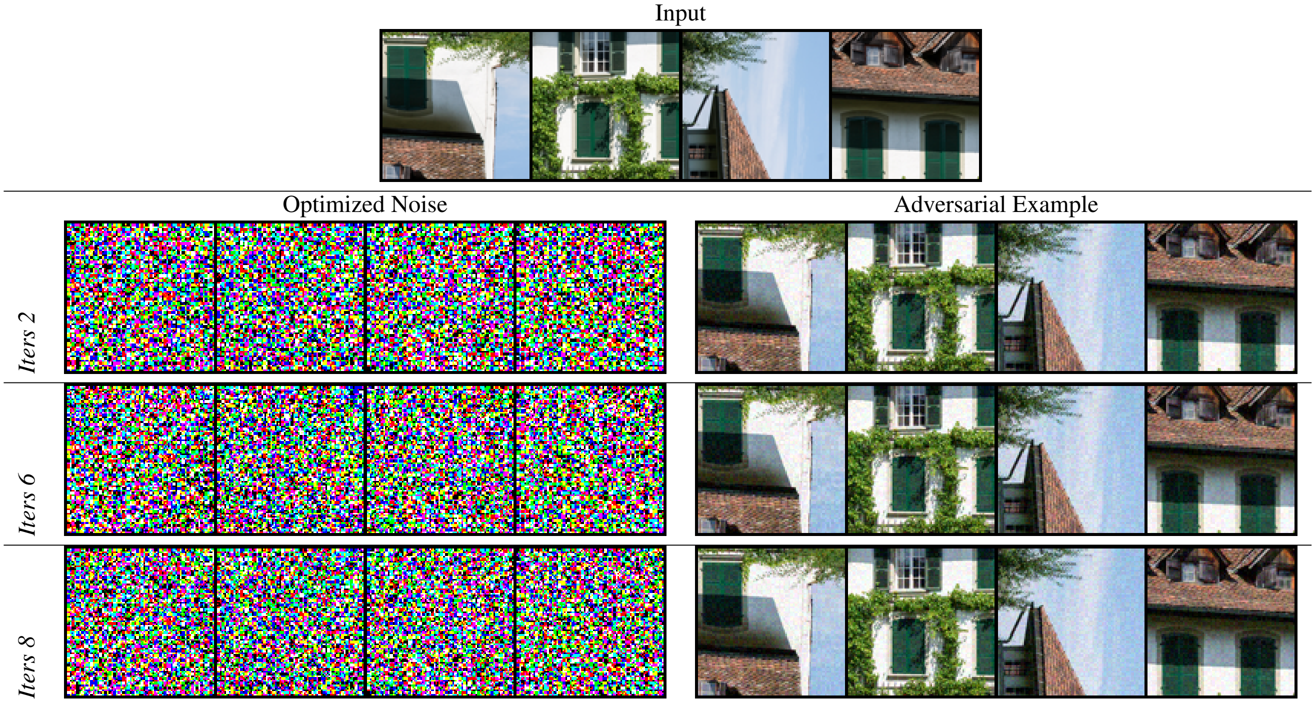

Figure 10. **Qualitative visualization of the iterations ablation.** Optimized noise and adversarial examples for different iterations in the adversarial attack's ablations.

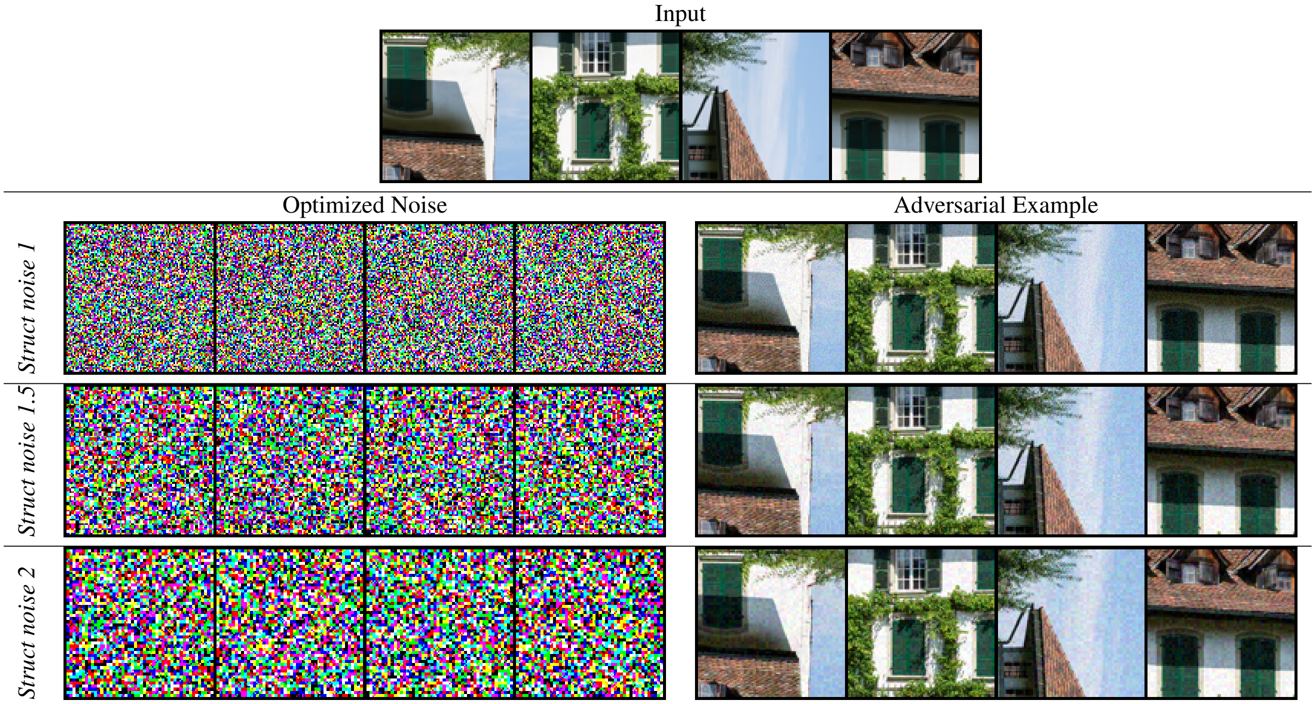

Figure 11. **Qualitative visualization of the scale of structured noise.** Optimized noise and adversarial examples for the different scales of structured noise in the adversarial attack's ablations. Note that, the higher the scale, the more aggregated the noise in the adversarial examples.