Rebuttal for "ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting"

Table 1. Quantitative comparison of different methods on the task of x4 bicubic super-resolution (64→ 256). We mark the number of sampling steps for each method by the format of "method-steps".

Methods	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MUSIQ↑	# Parameters (M)	Runtime (s)
IRSDE-100	24.48	0.602	0.304	0.513	45.382	137.2	5.927
DDRM-20	25.56	0.674	0.471	0.372	24.746	552.8	1.184
I2SB-20	26.76	0.730	0.206	0.489	53.936	552.8	1.832
F-ResShift-20	26.73	0.736	0.126	0.683	58.067	121.3	0.105

Table 2. Efficiency and performance comparison on the dataset of *ImageNet-Test* for the general x4 image super-resolution (64→ 256).

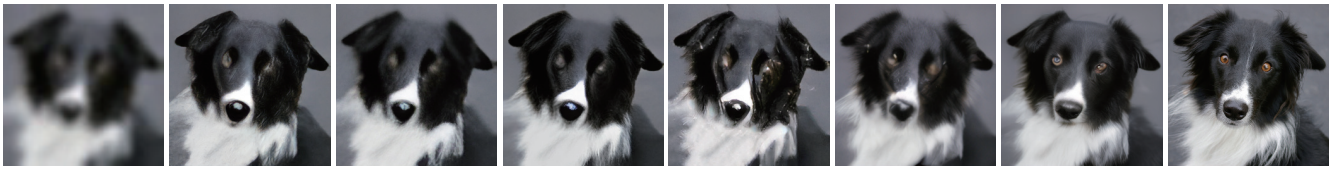
Metrics	BSRGAN	RealESRGAN	SwinIR	LDM-20	LDM-100	LDM-500	*F-ResShift*-20
PSNR↑	24.49	24.21	24.15	24.76	23.89	23.52	23.72
LPIPS↓	0.282	0.281	0.262	0.284	0.268	0.270	0.246
CLIPIQA↑	0.650	0.590	0.639	0.630	0.698	0.705	0.773
Runtime (s)	0.012	0.013	0.046	0.102	0.413	2.094	0.105
# Parameters (M)	16.70	16.70	28.01		113.6		121.3

Table 3. Quantitative comparison to recent SotA methods on the task of blind face restoration.

Methods	Metrics			
	PSNR↑	SSIM↑	LPIPS↓	FID ↓
GFPGAN	21.67	0.6167	0.4299	48.07
VQFR	21.43	0.5677	0.4393	48.21
CodeFormer	23.23	0.6515	0.3341	63.26
F-ResShift	23.26	0.6796	0.3480	49.95

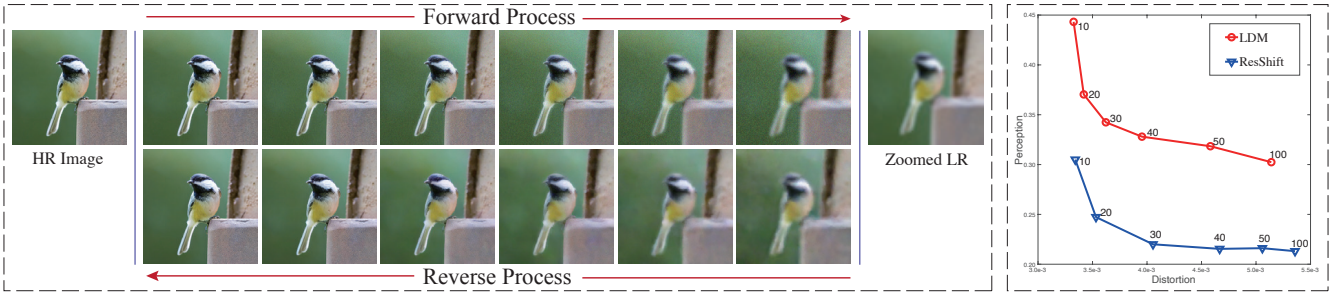
(a) LR Image (b) GFPGAN (c) VQFR (d) CodeFormer (e) F-*ResShift*

Figure 1. Qualitative comparison of different methods on one real-world example.



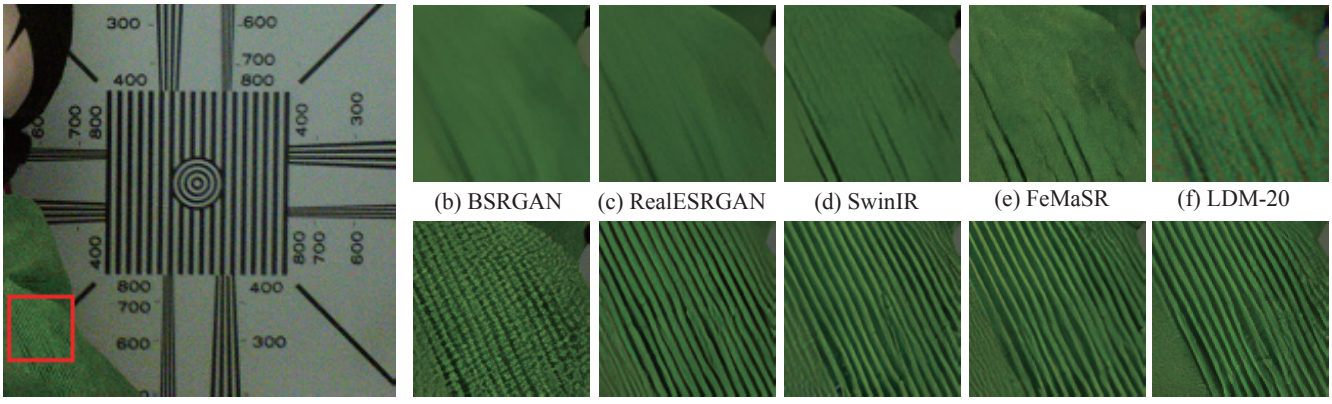
(a) Zoomed LR (b) BSRGAN (c) Real-ESRGAN (d) SwinIR (e) FeMaSR (f) LDM-20 (g) *F-ResShift* (h) Ground Truth

Figure 2. Qualitative comparison to various methods on the synthetic dataset of *ImageNet-Test*. Please zoom in for better view.



(a) An overview of *ResShift* (b) Perception-distortion Trade-off

Figure 3. (a) An overview of the proposed *ResShift*. We visualize its intermediate states at timesteps of 1, 5, 9, 13, 17, and 20 by setting $T = 20$. The results in the top row and bottom row are achieved in the original image space and the latent space of VQGAN, respectively. (b) Perception-distortion curves of *ResShift* and LDM **under different diffusion steps**. The vertical axis is the perceptual quality measured by $1 - c$, where c denotes CLIPIQA, and the horizontal axis represents the distortion measured by mean square-error (MSE).



(a) LR Image (b) BSRGAN (c) RealESRGAN (d) SwinIR (e) FeMaSR (f) LDM-20

(j) *F-ResShift* (h) *F-ResShift*+ (i) *F-ResShift*+ (j) *F-ResShift*+ (k) *F-ResShift*+

Figure 4. (a) LR image, (b)-(f) restored results by different competing methods, (j) recovered result by *F-ResShift* (trained with 500k iterations), (h)-(k) super-resolved results of *F-ResShift*+ (trained with 800k iterations) under multiple random seeds.