Low-Res Leads the Way:
Improving Generalization for Super-Resolution by Self-Supervised Learning

Supplementary Material

A. Experimental Details

A.1. Fine-tuning Details

Due to the different network architectures of different types of models, we trained different parts of the parameters for them. The rationale behind parameter selection for training is corroborated by empirical experiments detailed further in the text.

- For the training phases specific to BSRGAN, Real-ESRGAN+, and SwinIR-GAN, selective freezing of initial layers is implemented to concentrate training on the deeper parameters.
- In the case of FeMaSR, which is based on the VQ-GAN (Vector Quantized Generative Adversarial Network) structure, the focus is placed on the parameters of the VQGAN encoder.
- StableSR, which utilizes a pre-trained diffusion model, applies a controllable feature wrapping (CFW) module with an adjustable coefficient to refine the outputs of the diffusion model during the decoding process of the autoencoder. We choose to fine-tune the designed Collaborative Feature Weighting module and part of the encoder.

It usually takes 150 to 500 iterations to train. The time depends on the baseline network size, ranging from seconds to a few minutes. Our method can be fine-tuned either on individual images or on the entire test set assuming consistent degradation across the test set, which greatly reduces computational cost. Table 7 shows that our method takes only 8 minutes to fine-tune on the whole test set, much faster than others. Individual fine-tuning can improve the results if needed.

A.2. Testing Datasets

The validation of the effectiveness of our training method in real-world scenarios is conducted using real-world paired datasets, RealSR [4], and DRealSR [47]. These datasets are meticulously curated from various sensors to reflect different degradation characteristics inherent in each device. Furthermore, the datasets are segmented based on the capturing equipment. For RealSR, a 2× scale factor is employed, with separate subsets for Canon and Nikon. In the case of DRealSR, a 4× scale is applied across three subsets corresponding to Sony, Panasonic, and Olympus. To ensure a fair comparison with other models, we follow common settings employed by most methods. Each image is segmented into multiple smaller patches for performing 4× super-resolution, with the patch size for LR images being 128×128 and for HR images being 512×512.

	LPIPS↓	DISTS↓	PSNR↑	**fine-tuning time↓**
Ours ($d = 2048$)	**0.1629**	**0.1630**	**29.56**	**8 min**
ZSSR	0.2424	0.1889	29.14	19 hr
KernelGAN+ZSSR	0.3315	0.2774	23.52	72 hr
deep plug-and-play	0.2604	0.2524	29.44	1.4 hr
deep image prior	0.2091	0.2054	29.32	28 hr

Table 7. Our method can be fine-tuned on the entire test set assuming consistent degradation across the test set, which greatly reduces computational cost. Other methods need to be trained on each individual image.

A.3. Evaluation Metrics

Following prior research [16, 18], our study adopts a carefully curated set of perceptual metrics, ones that have shown a higher correlation with human perception, including Learned Perceptual Image Patch Similarity (LPIPS) [51], Deep Image Structure and Texture Similarity (DISTS) [13], and Normalized Laplacian Pyramid Distance (NLPD) [17]. LPIPS and DISTS have been empirically validated in [16, 18] as more closely aligned with human visual assessment than other metrics. We also include traditional metrics, such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [46], and most apparent distortion (MAD) [23].

A previous study [16, 18] investigated the correlation between human visual perception quality of images and various Image Quality Assessment (IQA) metrics, with the findings summarized in the Table 8. The experimental results reveal that MAD, LPIPS, and DISTS outperform traditional PSNR and SSIM across various aspects in the context of super-resolution evaluation. Specifically, MAD demonstrates superior accuracy in assessing traditional SR methods. On the other hand, both LPIPS and DISTS exhibit higher precision when evaluating GAN-based SR methods. In the overall comparison, DISTS emerges as the most effective metric for super-resolution assessment. These findings underscore the limitations of relying solely on conventional metrics such as PSNR and SSIM, emphasizing the importance of incorporating newer metrics like MAD, LPIPS, and DISTS for a more comprehensive and accurate evaluation of super-resolution techniques.

Method	SR Full	Traditional SR	PSNR. SR	GAN-based SR
PSNR	0.4099	0.4782	0.5462	0.2839
SSIM	0.5209	0.5856	0.6897	0.3388
MAD	0.5424	**0.6720**	0.7575	0.3494
LPIPS	0.5614	0.5487	0.6782	0.4882
DISTS	**0.6544**	0.6685	**0.7733**	**0.5527**

Table 8. The Spearman rank correlation coefficient (SRCC) between MOS (Mean Opinion Score) and various IQA (Image Quality Assessment) metrics across different distortion sub-types.

B. LR Reconstruction Network

B.1. Degradation Encoder

Following the methodology proposed by Liu et al. [35], our degradation encoder is constructed by integrating a pre-trained SR-GAN model [24] and downsampling layers. This collaborative framework aims to produce degradation embeddings, denoted as e, with a dimensionality of 512. The choice of a relatively small dimension for e ensures that the degradation embeddings do not encapsulate intrinsic image information but are sufficiently representative of content pertaining specifically to the degradation process. This design principle is crucial in isolating and preserving only the features relevant to degradation, avoiding contamination with the original image characteristics.

B.2. Reconstructor

In our methodology, we incorporate a modulation-demodulation-convolution strategy reminiscent of Instance Normalization as employed in StyleGAN2 [19]. This approach effectively utilizes the degradation embedding e to facilitate LR reconstruction when combined with the SR network's output I_{SR}. To delve deeper into the specifics of this strategy, during modulation, a style is learned from the provided degradation embedding e. The modulation operation scales each input feature map of the convolution using the acquired style, as denoted by the equation

$$w_{ijk} = s_i \cdot w_{ijk},$$

the variables w and w' represent the original and modulated weights, respectively. The scale factor, denoted as s_i, corresponds to the ith input feature map. The indices j and k are used to iterate over the output feature maps and spatial footprint of the convolution, respectively. This modulation process ensures that the convolutional features are adaptively adjusted based on the characteristics embedded in the degradation embedding. Following modulation, a demodulation step is executed to obtain the demodulated convolution weights, represented as

$$w''_{ijk} = \frac{w'_{ijk}}{\sqrt{\sum_{i,k} {w'_{ijk}}^2 + \epsilon}}.$$

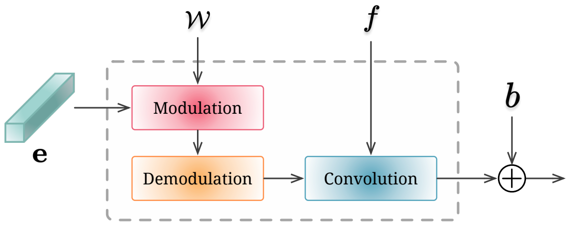

Figure 10. Modulation method used in our LR reconstructor.

The primary objective of demodulation is to restore the outputs to a unit standard deviation, providing stability and normalizing the feature representations. It is crucial to emphasize that this modulation-demodulation-convolution strategy facilitates the integration of degradation-specific information into the LR reconstruction process. The adaptability of the convolutional features based on the learned style ensures that the network can effectively reconstruct LR inputs, enhancing the overall performance of the SR framework.

C. More Experiment Results

C.1. The Effect of Fine-tuning Parameters for Different Network Architecture.

In our exploration of parameter fine-tuning, we observe that increasing the number of trained parameters of BSRGAN results in higher PSNR values. However, the LPIPS score reaches its optimal point at approximately 60% - 70% of the parameters, as depicted in Figure 11. Considering the limitation of PSNR, we prioritize the use of LPIPS as our reference. It's important to note that different network and testsets may yield varied conclusions. In our investigation into the impact of training parameters on the performance of the FeMaSR and SwinIR networks, the influence is shown in Figure 12 and Figure 13. Specifically, for the FeMaSR network, the optimal PSNR is achieved when training parameters constitute 86%, while the optimal LPIPS is obtained at 100%. In contrast, SwinIR attains the best PSNR and LPIPS values almost simultaneously at 100% of training parameters.

C.2. Ablation on Model Size

Table 9 delineates the efficacy of our proposed model across a spectrum of sizes, demonstrating that our method retains robust performance notwithstanding the model's capacity. From a comprehensive model with 12.9 million parameters to a compact version with merely 495 thousand parameters, our approach consistently outperforms the baseline.

C.3. Performance on Different Degradation

Table 10 demonstrates the robustness of our 'LWay' method in handling various types of image degradations. It presents notable improvements in PSNR and reductions in LPIPS

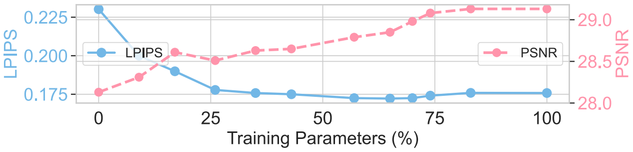

Figure 11. The performance curve for fine-tuning different percentages of parameters for BSRGAN.

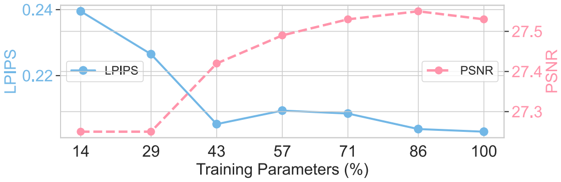

Figure 12. The performance curve for fine-tuning different percentages of parameters for FeMaSR.

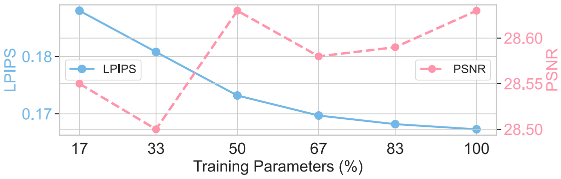

Figure 13. The performance curve for fine-tuning different percentages of parameters for SwinIR.

	Paras	FLOPS	PSNR↑	MAD↓	LPIPS↓	DISTS↓
baseline	-	-	28.13	118.48	0.2302	0.2102
+ LWay `(Large)`	12.9 M	589.4 G	28.85	104.71	0.1722	0.1772
+ LWay `(Medium)`	5.38 M	117.6 G	29.50	99.76	0.1798	0.1810
+ LWay `(Small)`	2.77 M	44.49 G	28.69	106.42	0.1837	0.1862
+ LWay `(Tiny)`	495 K	19.38 G	28.70	104.92	0.1808	0.1842

Table 9. The performence of different model size. LWay is not contingent on the parameter count of the LR reconstruction, demonstrating effectiveness even with a small parameter volume.

	real-world degradation		**synthetic**, blur 17×17		**synthetic**, blur 11×11		**synthetic**, JPEG $q = 15$	
	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS
baseline	28.13	0.2302	27.55	0.4065	27.5	0.3922	26.60	0.4240
+LWay (d=2048)	29.56	0.1629	28.39	0.2755	29.02	0.2265	26.85	0.3122

Table 10. Performance on different degradation. LWay improve image quality under a range of degradations.

for both real-world degradation and synthetic distortions such as blurring and JPEG compression, signifying our method's efficacy in maintaining image integrity across different degradation scenarios.

D. More Visual Results

D.1. LR Reconstruction Visualization

The visual outcomes of the LR reconstruction network are illustrated in Figure 14, encompassing HR, LR, and the reconstructed LR images. Notably, our network demonstrates the capability to restore LR images that closely approximate the ground truth LR by extracting a 512-dimensional degradation embedding solely from the LR input and subsequently integrating it with the HR image. This process demonstrates the effectiveness of our LR reconstruction approach in achieving visually compelling results. The showcased robustness of our LR reconstruction network is particularly noteworthy. Given that the transition from HR to LR is generally considered easier compared to the reverse process, our method exhibits a heightened degree of resilience with limited data. Leveraging only a finite dataset, our approach achieves a robust performance, underscoring its capacity to generalize and adapt well to diverse LR input scenarios.

D.2. More Visual Comparison

Figure 15 and Figure 16 provide additional comparisons of our proposed method with other state-of-the-art (sota) approaches. Our method excels in effectively restoring the texture and fine details of images. In contrast, DASR, DiffBIR, and StableSR tend to produce smoother results at the expense of losing texture details. ZSSR, on the other hand, exhibits limited restoration capabilities, resulting in less clear outcomes that are less faithful to the LR input. The results generated by LDM display inconsistencies in texture details compared to the ground truth. DARSR, while prone to failure and introducing significant color bias, and CAL_GAN both exhibit varying degrees of artifacts in their outputs. These visual comparisons underscore the superior performance of our proposed method in preserving intricate details and textures during the super-resolution process. The tendency of other methods to sacrifice fine details for smoother results, introduce artifacts, or inaccurately represent texture details highlights the unique strengths of our approach.

E. Discussion

Differences to optimization-based methods. Optimization-based methods, relying on pre-defined degradation models or downsampling operators, have limited capabilities in handling complex degradation. They are also time-consuming and difficult with large data. In contrast, our approach incorporates a more general and robust degradation modeling. Moreover, our method marries the benefits of supervised and unsupervised training, outperforming optimization-based methods that only use test images.

Differences to KernelGAN. KernelGAN's discriminator only makes binary judgments (0/1), while LWay uses pixel-level regression to better capture the distribution. Moreover, local KernelGAN's kernels have limited information and robustness in real-world, while our embedding has richer ex-

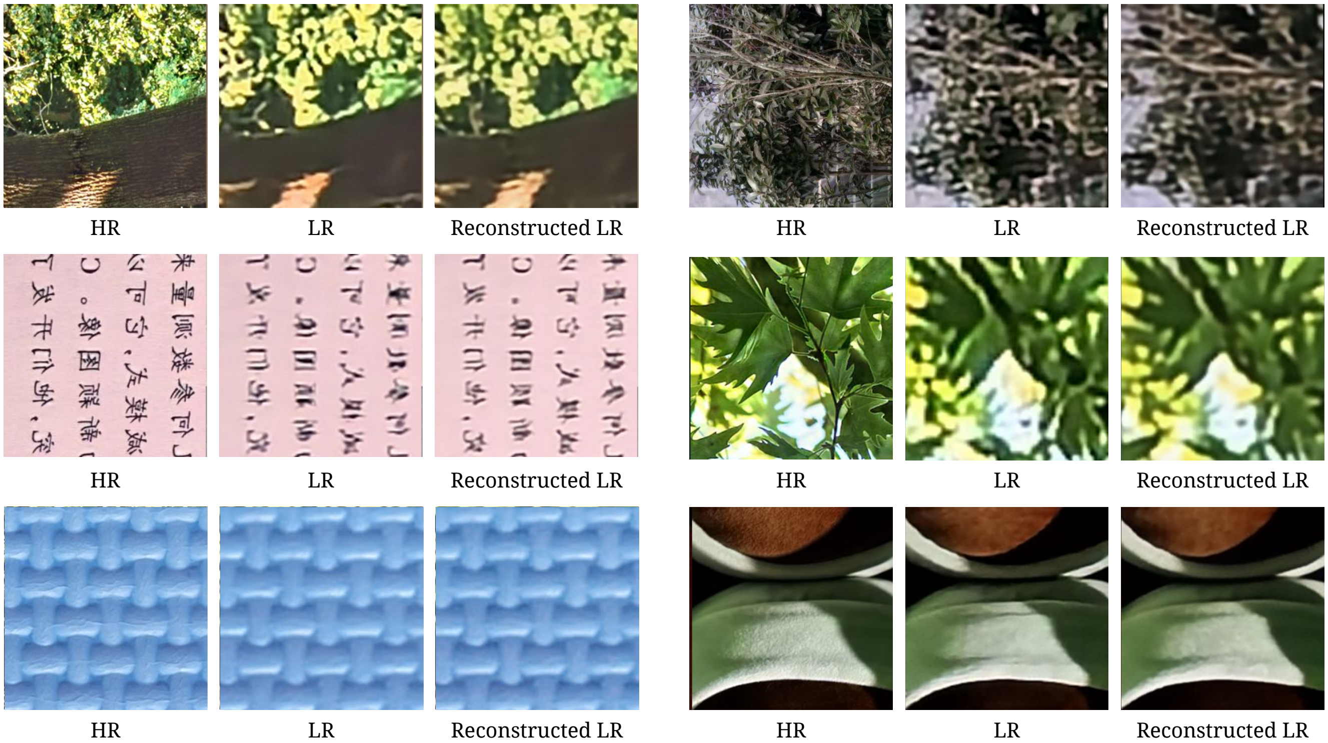

Figure 14. Visual results of LR reconstruction. For instance, given an input LR image and HR image, the degradation encoder encodes a 512-dimension degradation embedding e, the reconstructor utilizes e and the HR image to reconstruct the estimated LR image.

ternal priors rather than relying on solely learning test images and is robust as demonstrated by validation.

F. Limitation

The proposed architecture excels in extracting and restoring information from low-resolution (LR) images, especially when they contain discernible texture details. It is within these conditions that our method showcases its maximum effectiveness. However, a limitation might arise when the LR images themselves lack texture details, impeding the model's capability to execute effective restoration.

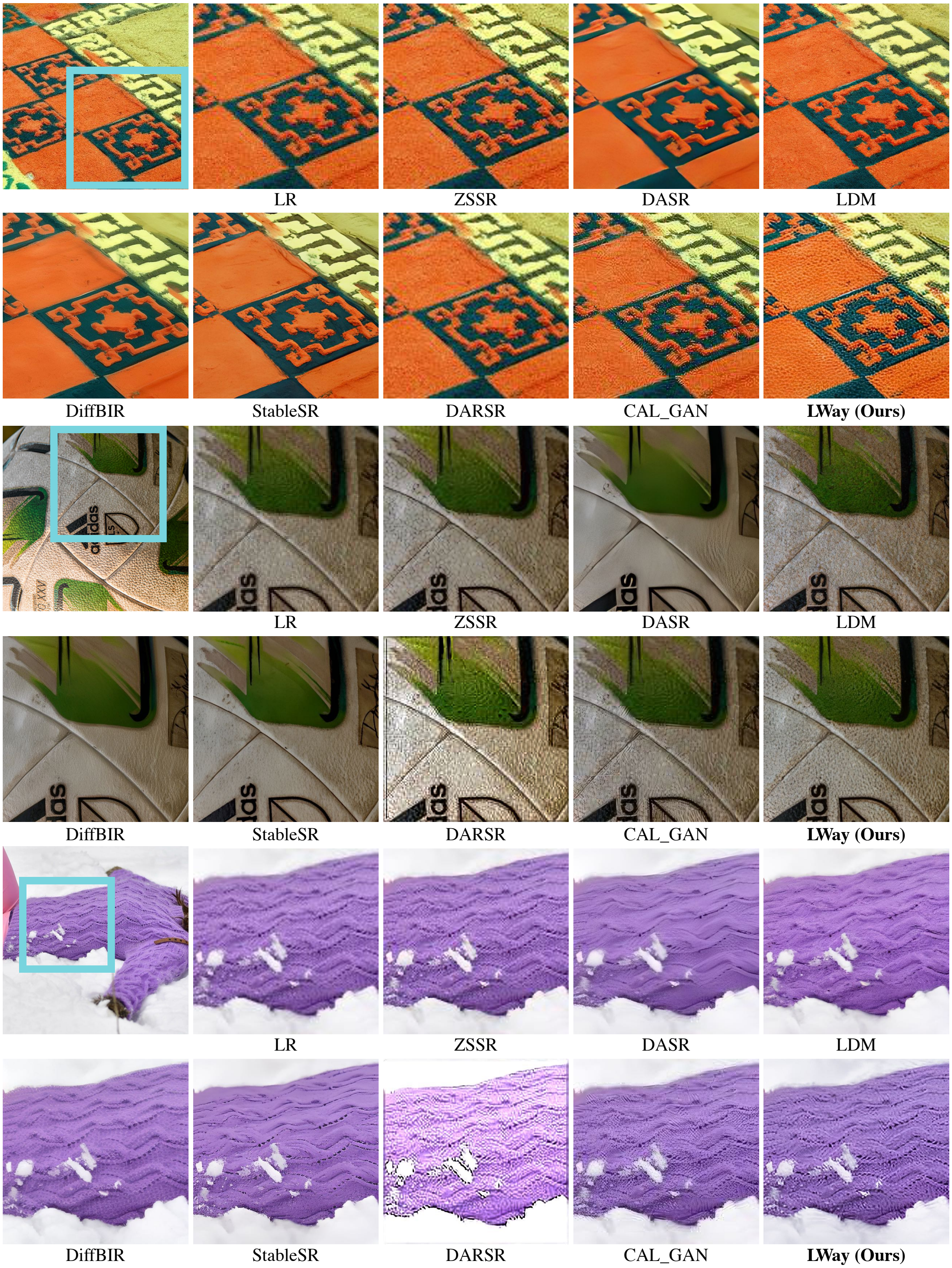

Figure 15. More visual comparisons.

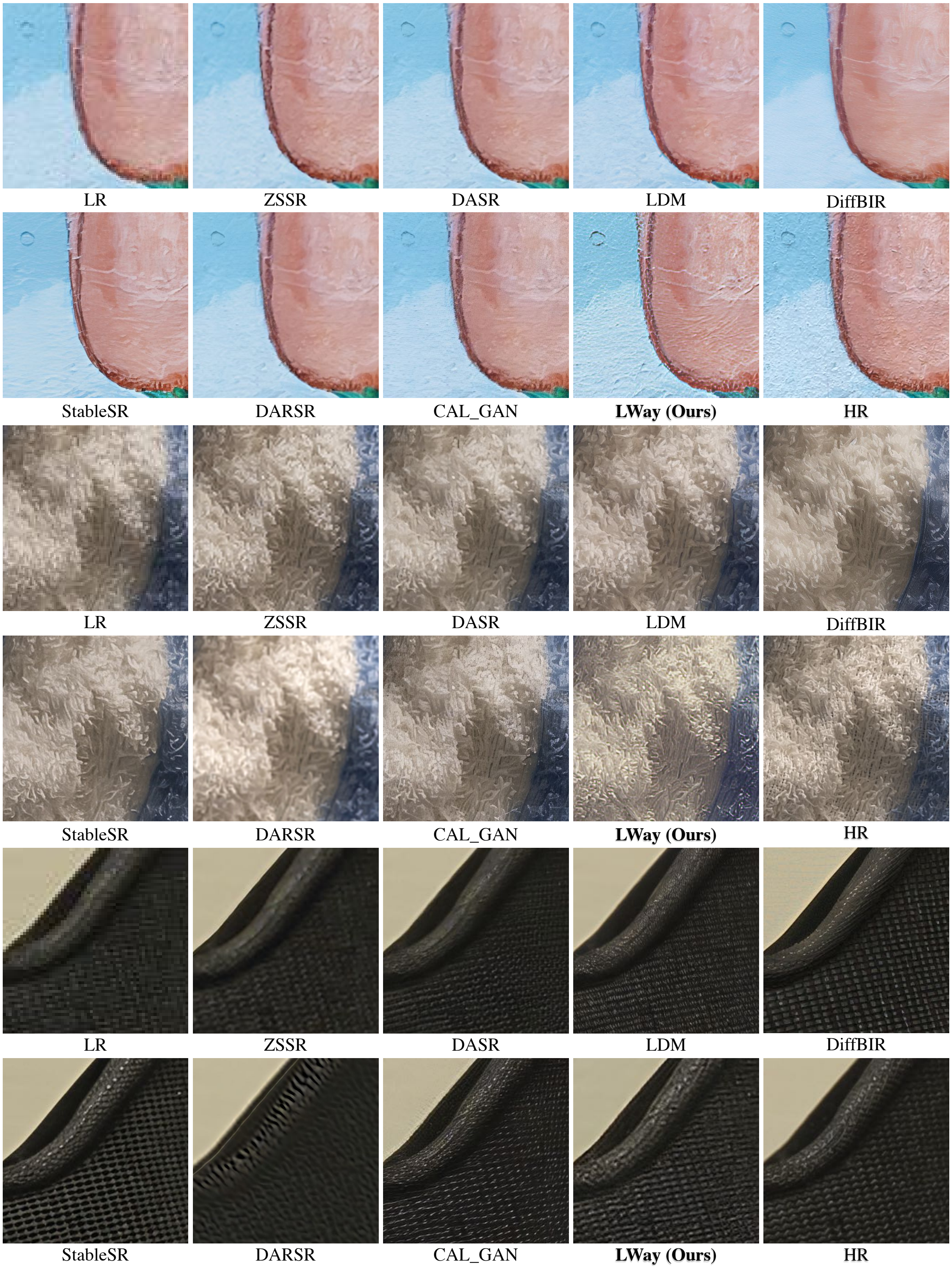

Figure 16. More visual comparisons.