A ADAPTABILITY OF OUR SEAL

We contend that the components of our SEAL framework are highly adaptable to user preferences. For instance, users have the option to choose a different reference line to visualize distributed performance, reorganize the SE test sets into new groups, and utilize any IQA metrics for evaluation.

A.1 EXTENSION ON NEW IQA METRIC

To illustrate the adaptability of our SEAL framework, we have opted for SSIM as the IQA metric to perform a comprehensive evaluation of real-SR methods. As depicted in Tab. 6, RealESRNet surpasses other methods in terms of AR, an outcome that can be credited to the use of sharpened Ground-Truth images. It is significant that RealESRNet and SwinIR exhibit remarkable stability, as evidenced by their RPR_I values. Furthermore, our findings indicate that SwinIR attains the highest RPR_A value, implying that transformer-based networks favor acceptance degradation cases. As evidenced by these observations, our proposed evaluation framework displays considerable adaptability. It accommodates various IQA metrics to systematically evaluate real-SR methods from diverse angles, such as reconstruction capability (PSNR) and structural similarity (SSIM).

Table 6: Results and ranking of different methods on SSIM by our SEAL framework. The subscript denotes the rank order. $_\times$ represents a failed SR model in a large degradation space.

Set14-SE	$AR\uparrow$	$RPR_I\downarrow$	$RPR_A\uparrow$	$RPR_U\uparrow$	Rank
SRResNet	$0.00_{(\times)}$	0.04	0.00	0.04	$\times$
DASR	$0.00_{(\times)}$	0.03	0.00	0.04	$\times$
BSRNet	$0.76_{(3)}$	$0.27_{(5)}$	$0.70_{(2)}$	$0.36_{(4)}$	3
RealESRNet	$0.91_{(1)}$	$0.16_{(1)}$	$0.67_{(3)}$	$0.43_{(1)}$	1
RDSR	$0.32_{(5)}$	$0.22_{(3)}$	$0.59_{(5)}$	$0.33_{(5)}$	5
RealESRNet-GD	$0.69_{(4)}$	$0.26_{(4)}$	$0.67_{(3)}$	$0.39_{(2)}$	4
SwinIR	$0.84_{(2)}$	$0.17_{(2)}$	$0.72_{(1)}$	$0.38_{(3)}$	2

A.2 USER-CUSTOMIZED SE TEST SETS

In order to accommodate varying user preferences, such as the analysis of the quantitative performance of IQA metrics, the SE test sets are organized in ascending order based on the PSNR values of the FSRCNN-mz output. These sets are then partitioned into five groups of equal size. Group 1 encompasses the most challenging cases, while Group 5 includes the least challenging ones. As shown in Table 7, the average RPR value of BSRNet closely matches that of RealESRNet-GD. However, there is a variation in their performance across different groups. RealESRNet-GD outperforms in groups $\{3, 4, 5\}$, whereas BSRNet takes the lead in groups $\{1, 2\}$.

Table 7: RPR value of different methods on the Set14-SE with PSNR. Blue: better than FSRCNN-mz.

Model	SRResNet	DASR	BSRNet	RealESRNet	RDSR	RealESRNet-GD	SwinIR
Group 1	0.03	0.03	**0.64**	0.37	0.26	0.34	0.48
Group 2	0.02	0.02	**0.60**	0.37	0.21	0.45	0.43
Group 3	0.07	0.07	0.51	0.42	0.31	**0.57**	0.40
Group 4	0.02	0.01	0.37	0.40	0.29	**0.68**	0.27
Group 5	0.03	0.03	0.44	0.36	0.17	**0.55**	0.36
Average	0.03	0.03	0.51	0.38	0.25	**0.52**	0.39

A.3 EXTENSION ON ACCEPTANCE LINE AND EXCELLENCE LINE.

For the acceptance line, we hope it can represent an acceptable lower bound of performance with good discrimination for different models. Concretely, the acceptance line cannot be so high that AR of most methods cannot exceed 0, nor can it be so low that AR can easily reach 1.0. FSRCNN is a small network (0.4M Params.) while it can distinguish the performance difference well, as shown in Tab. 1. Therefore, we choose FSRCNN-mz as the acceptance line.

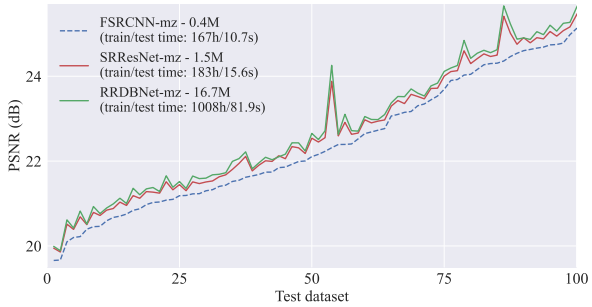


Figure 10: **Comparison of network structures** for the acceptance and excellence lines.

Table 8: **Basic strategies for model comparison.** We use three basic strategies to compare the overall performance of real-SR models. The real-SR models are trained on: (1) different network structures, (2) different training datasets, and (3) the RealESRGAN degradation model with different gate probability as proposed in Zhang et al. (2022).

		$AR\uparrow$	$RPR_I\downarrow$	$RPR_A\uparrow$	$RPR_U\uparrow$	Rank
Network (Parameter [M])	SRResNet (1.5)	$0.12_{(\times)}$	0.20	0.63	0.26	$\times$
	RCAN (15.6)	$0.37_{(2)}$	$\mathbf{0.15}_{(1)}$	$0.62_{(2)}$	$0.39_{(2)}$	2
	RRDBNet (16.7)	$0.37_{(2)}$	$0.33_{(3)}$	$0.68_{(1)}$	$0.32_{(3)}$	3
	SwinIR (11.9)	$\mathbf{0.67}_{(1)}$	$\mathbf{0.15}_{(1)}$	$0.62_{(2)}$	$0.41_{(1)}$	1
Training dataset	DIV2K	$0.32_{(3)}$	$0.25_{(3)}$	$0.64_{(2)}$	$0.33_{(3)}$	3
	DF2K	$0.43_{(2)}$	$0.24_{(2)}$	$0.67_{(1)}$	$0.39_{(2)}$	2
	ImageNet	$\mathbf{0.63}_{(1)}$	$0.22_{(1)}$	$0.67_{(1)}$	$0.41_{(1)}$	1
Gate probability	1.00	$0.37_{(4)}$	$0.33_{(1)}$	$0.68_{(3)}$	$0.32_{(2)}$	3
	0.75	$\mathbf{0.44}_{(1)}$	$0.35_{(2)}$	$0.69_{(2)}$	$0.34_{(1)}$	1
	0.50	$0.43_{(2)}$	$0.35_{(2)}$	$0.70_{(1)}$	$0.31_{(3)}$	1
	0.25	$0.40_{(3)}$	$0.43_{(4)}$	$0.66_{(4)}$	$0.21_{(4)}$	4
BSRNet (SOTA)		$0.59_{(2)}$	$0.42_{(2)}$	$0.72_{(1)}$	$0.27_{(2)}$	2
SwinIR-GD-I (Ours)		$\mathbf{0.85}_{(1)}$	$\mathbf{0.25}_{(1)}$	$\mathbf{0.72}_{(1)}$	$\mathbf{0.40}_{(1)}$	1

For the excellence line, we compare the networks of SRResNet (1.5M Params.) and RRDBNet (16.7M Params.). In Fig. 10, we observe that SRResNet-mz and FSRCNN-mz can already distinguish the performance difference. Although RRDBNet-mz exhibits a slight performance improvement, it comes at the expense of increased training and testing time, far surpassing those of other models. Considering the trade-off between performance and time costs, we choose SRResNet-mz as the excellence line. Nonetheless, we emphasize that our rationale for choosing these two lines is that they can well differentiate the methods for comparison. Note that the two lines can be changed flexibly to meet specific requirements of other scenarios.

B Developing New Strong Real-SR Models

According to the evaluation results by our framework, as shown in Tab. 8, we can improve the real-SR performance in **three** aspects to develop a stronger real-SR model: **1)** A powerful backbone is vital for overall performance. We can observe that SwinIR obtains the highest AR and the lowest RPR_I. **2)** Using a large-scale dataset (i.e., ImageNet Deng et al. (2009)) can also greatly improve the real-SR performance. **3)** A degradation model with the appropriate distribution (i.e., gate probability: 0.75 Zhang et al. (2022)) also has a non-negligible impact on the real-SR performance. Based on these observations, we use SwinIR as the backbone to train a new strong real-SR model on ImageNet with a high-order gate degradation (GD) model (gate probability: 0.75), denoted as SwinIR-GD-I. The evaluation results in Tab. 8 show that SwinIR-GD-I obtain a significant improvement over the SOTA performance of BSRNet. Fig. 11 shows that the visual results of SwinIR-GD-I are obviously better than BSRNet and SwinIR. We believe our framework would inspire more powerful real-SR methods in the future.

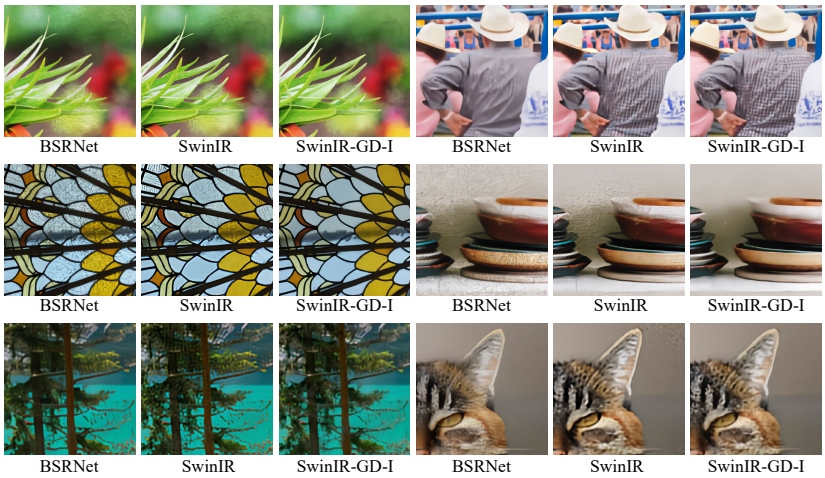


Figure 11: Visual results of the proposed baseline SwinIR-GD-I with real-SR methods.

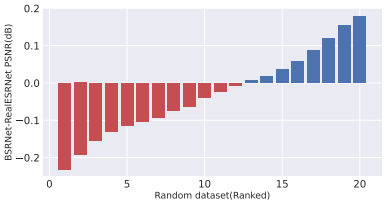


Figure 12: Results of conventional evaluation on 20 random test sets.

Figure 13: Similar visual effects of different degradation combinations.

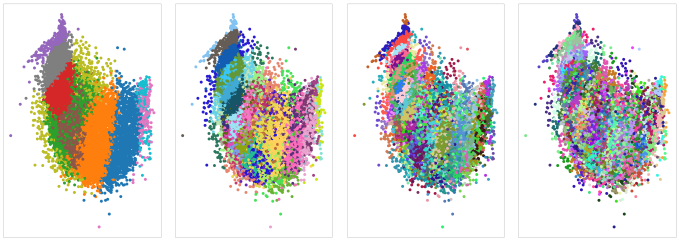
C Details of Conventional Evalution

To better understand the biased comparison of conventional evaluation on real-SR, we use common degradation types to illustrate the number of degradation combinations on a simple one-order degradation model. In Tab. 9, the number of degradation combinations is obtained using uniform sampling for each degradation type. It can be seen that the number of degradation combinations can reach up to 1.08×10^5. Furthermore, RealESRGAN Wang et al. (2021b) adopts a high-order degradation model with complex degradation combinations, such as an-isotropic blur, Poisson noise, gray noise, and *sinc* filter. The combination cases of the high-order degradation models will reach astronomical numbers. As described in the main paper, such a huge space cannot be sufficiently sampled by a few samples due to different combinations of degradation may produce similar visual effects.

Table 9: The number of combination degradation cases on a simple one-order degradation model.

	Gaussian Blur	Gaussian noise	Resize	Comression
Type	iso	color	nearest, bilinear, bicubic	JPEG
Range	Sigma: [0, 2.8]	Sigma: [2, 25]	scale: [0.125, 2]	range: [30, 95]
Sampling interval	0.2	2	0.1	5
Number	14	11	54	13
Total number	$14 \times 11 \times 54 \times 13 = 1.08 \times 10^5$			

Inspired by our proposed evaluation framework, which utilizes hundreds of representative test sets to evaluate real-SR models, we create 20 random test sets to further analyze existing evaluation methods. Using a large degradation model, we randomly apply degradations to the images from the DIV2K_val dataset. As shown in Fig. 12, each bar represents the PSNR difference between BSRNet and RealESRNet in a single test set. It reveals that the comparison conclusions are often inconsistent (e.g., -0.22 dB in test set 1 and 0.18 dB in test set 20) or indistinguishable (e.g., -0.01 dB in test set 12) using a single test set. When we average the PSNR results across all test sets, we find an average



(a) $k5$ $s0.12$ (b) $k40$ $s0.04$ (c) $k100$ $s0.03$ (d) $k200$ $s0.01$

Figure 14: **Ablation of the number of clusters** k. s denotes silhouette score.

difference of only 0.02 dB, which is statistically insignificant. These findings indicate that existing evaluation methods may not be sufficient to evaluate the real-SR methods.

D Details of Degradation Clustering

D.1 Spectral Clustering

We use the shuffled degradation model of BSRGAN Zhang et al. (2021) and the high-order degradation model of RealESRGAN Wang et al. (2021b) to construct a large degradation space. The degraded images are generated by the shuffled Zhang et al. (2021) and high-order Wang et al. (2021b) degradation models with probabilities of $\{0.5, 0.5\}$. The degradation types mainly consist of 1) various types of Gaussian blur; 2) commonly-used noise: Gaussian, Poisson, and Speckle noise with gray and color scale; 3) multiple resize strategy: area, bilinear and bicubic; 4) JPEG noise.

Algorithm 1 Image degradation clustering

Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct.
1: Compute Adjacency Matrix W and Degree Matrix D.
2: Compute Laplacian Matrix $L = D - W$.
3: Compute the first K eigenvectors $u_1, ..., u_k$ of L.
4: Let $U \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $u_1, ..., u_k$ as columns.
5: For $i = 1, ..., n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of U.
6: Cluster the points $(y_i)_{i=1,...,n}$ in $\mathbb{R}^k$ with the k-means algorithm into clusters $C_1, ..., C_k$
Output: Cluster centers $c_1, ..., c_K$ with $c_i \in C_i$.

We use spectral clustering to group the degraded images (x) due to its effectiveness and flexibility in finding arbitrarily shaped clusters. First, we use the histogram (h) Tang et al. (2011); Ye & Doermann (2012) with 256 values (bins) as the image feature to calculate the similarity $s_{ij} = L_1(h(x_i), h(x_j))$. The similarity matrix is defined as a symmetric matrix S, where s_{ij} represents a measure of the similarity between data points x with indices i and j for n data points. We execute Algo. 1 step by step to obtain the degradation parameter of cluster centers $\mathcal{D} = \{c_1, c_2, ..., c_K\}$. Then, we use the degradation parameter of cluster centers as the representative degradations to construct the systematic set.

D.2 Similarity Metrics

In this section, we provide more experimental details for the *Sec. similarity metrics* in the main paper. To select an appropriate similarity metric, we create two dataset with simple degradation types – Gaussian blur with a range of [0.1, 4.0] and Gaussian noise [1, 40]. We use the image *lenna* in Set14 Zeyde et al. (2010) as our Ground-Truth image. Firstly, we generate 100 low-quality images named Blur100 by applying Gaussian blur within a range of [0.1, 4.0]. Each cluster is assigned a label based on the degradation intensity. We label the low-quality images with $\{$[0.1, 1.0], [1.0, 2.0], [2.0, 3.0], [3.0, 4.0]$\}$ as $\{1, 2, 3, 4\}$ respectively. Similarly, we generate 100 low-quality images named Noise100 by applying Gaussian noise within the range [1, 40], labeled as $\{1, 2, 3, 4\}$ based on noise intensity.

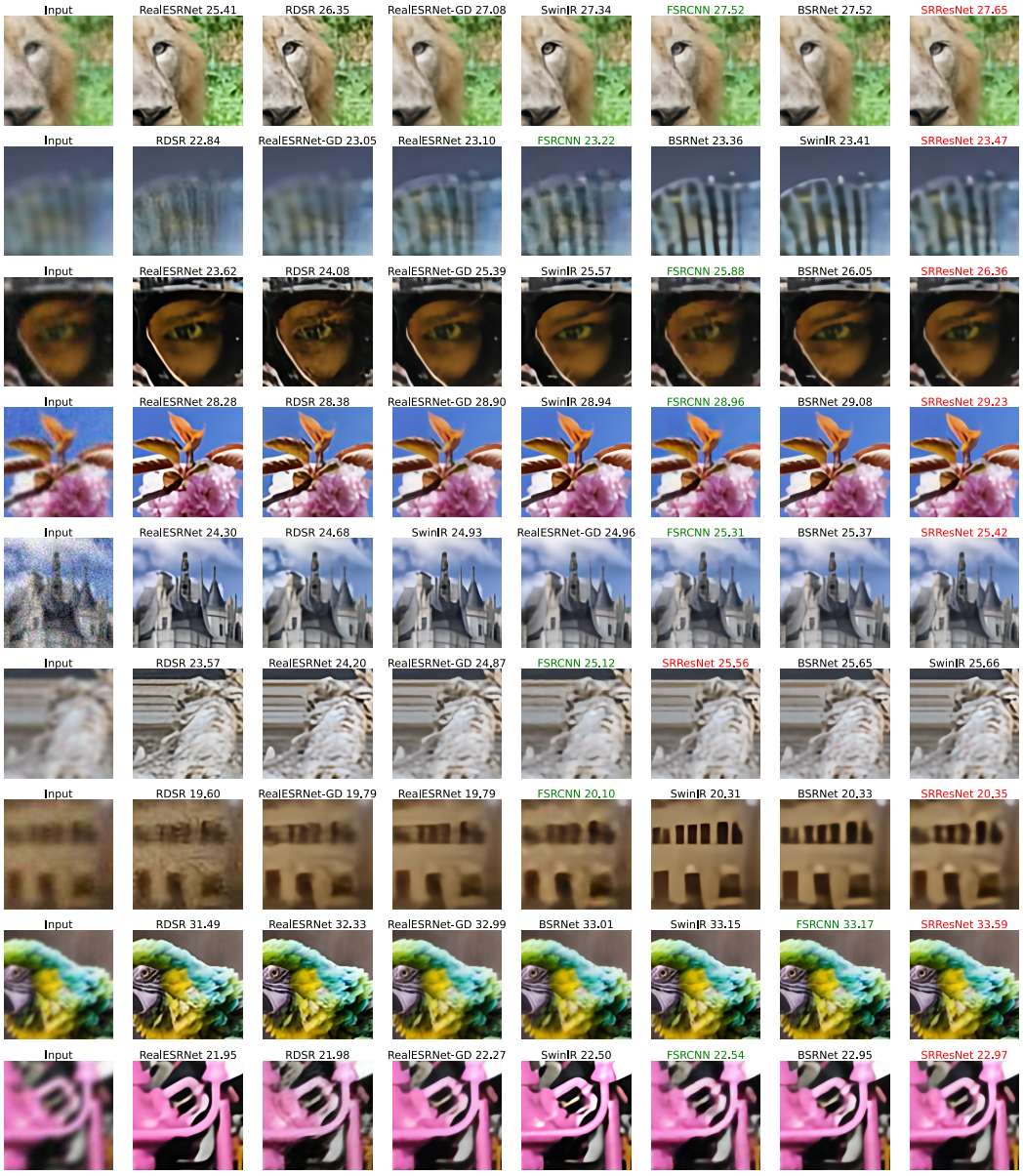


Figure 15: Visual results of MSE-based real-SR methods and the acceptance line FSRCNN and excellence line SRResNet with PSNR metric. It is best viewed in color.

To evaluate the effectiveness of the similarity metric, we combine Blur100 and Noise100 to produce BN100, which comprises 100 blurred images and 100 noised images. BN100 is labeled as {1, 2, 3, 4, 5, 6, 7, 8} using the same criteria as the previous datasets. We evaluate the clustering performance using purity accuracy, which divides the number of correctly matched class and cluster labels by the total number of data points.

D.3 THE NUMBER OF CLUSTERS

To determine the number of clusters, we use silhouette scores Rousseeuw (1987) to measure the quality of the clusters. A higher silhouette score represents a better cluster, while the clustering result is acceptable if the silhouette score is greater than 0. As demonstrated in Fig. 14, the silhouette scores of k=40 and k=100 are very close, thus we utilize 100 clusters to find the representative cases.

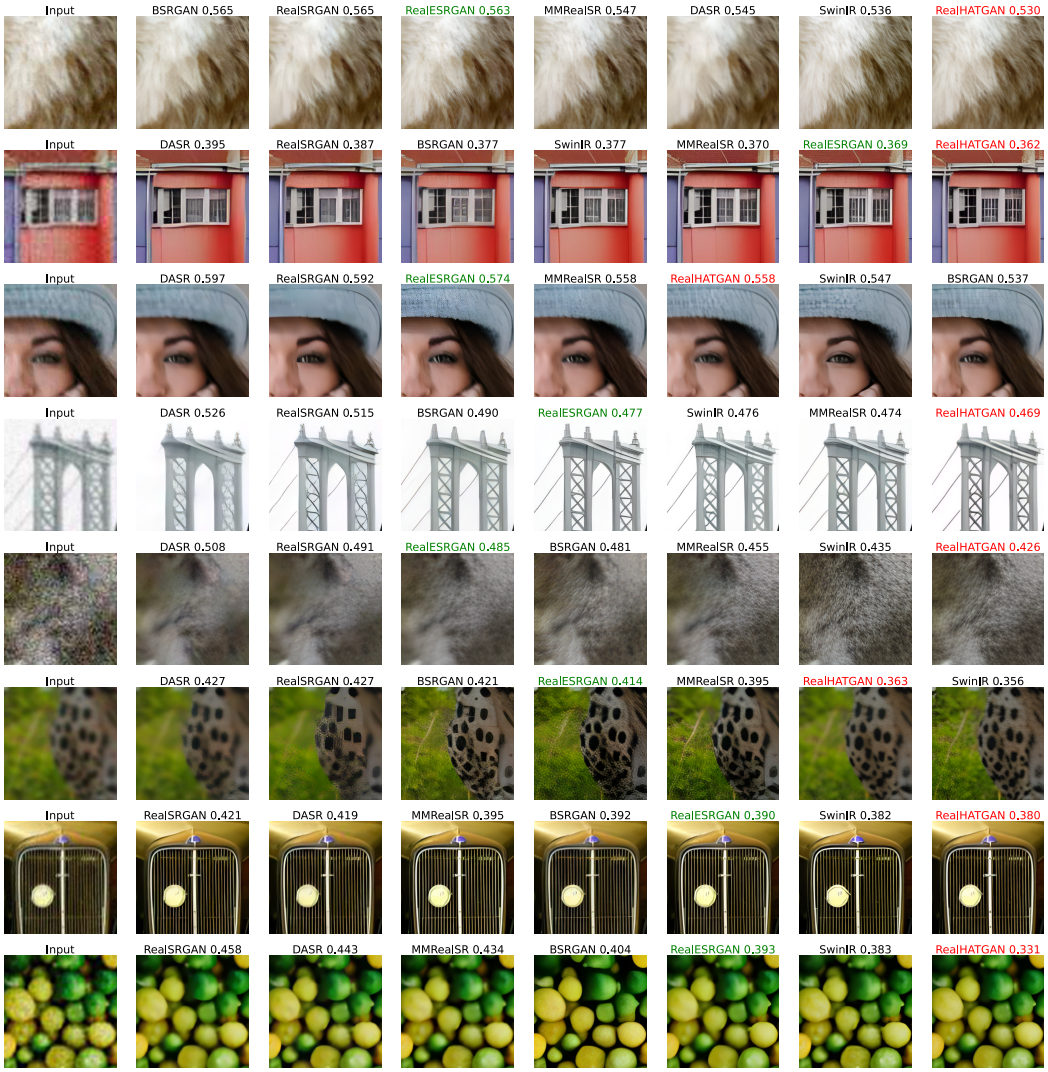


Figure 16: Visual results of GAN-based real-SR methods and the acceptance line RealESRGAN and excellence line RealHATGAN with LPIPS metric. It is best viewed in color.

E More Experimental Results

E.1 More Visual Results on real-SR methods

In this section, we further explore the effectiveness of our evaluation framework by providing additional qualitative results. We compare our proposed lines of acceptance and excellence with existing real-SR methods. The MSE-based methods that we consider include DASR Wang et al. (2021a), BSRNet Zhang et al. (2021), SwinIR Liang et al. (2021a), RealESRNet Wang et al. (2021b), RDSR Kong et al. (2022), and RealESRNet-GD Zhang et al. (2022). In Fig. 15, we use FSRCNN (green) to denote the acceptance line, and SRResNet (red) to represent the excellence line. Moving on to the GAN-based methods, we include DASR Liang et al. (2022), BSRGAN Zhang et al. (2021), MMRealSR Mou et al. (2022), SwinIR Liang et al. (2021a) and RealSRGAN Ledig et al. (2017). In Fig. 16, RealESRGAN (green) is used to denote the acceptance line, and RealHATGAN (red) is used to represent the excellence line. This comprehensive comparison provides a clear understanding of the performance of our proposed lines against the existing methods, thereby demonstrating the effectiveness of our evaluation framework.

E.2 DEGRADATION CLUSTERING RESULTS

In this section, we present visual results for the *lenna* image, processed with degradation parameters of cluster center [1, 100]. These results are sorted based on the PSNR values of the output from FSRCNN-mz. Fig. 17 showcases the most challenging cases encountered in our study. On the other hand, Fig. 21 highlights the cases that were relatively easier to handle. This comparative analysis provides a clear understanding of the performance range of our proposed evaluation framework.

Figure 17: The visual results with the degradation parameters of cluster center [1, 20]. Best viewed in color.

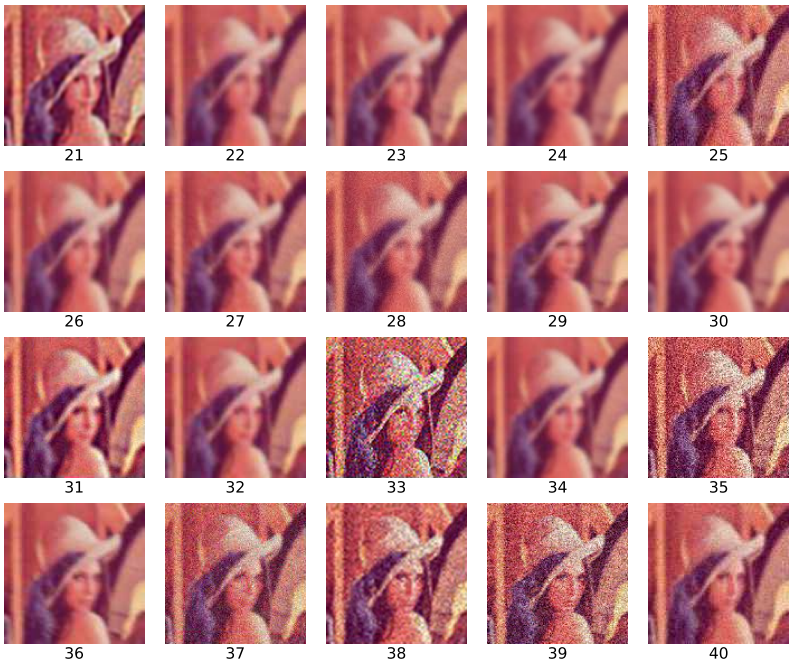


Figure 18: The visual results with the degradation parameters of cluster center [21, 40]. Best viewed in color.

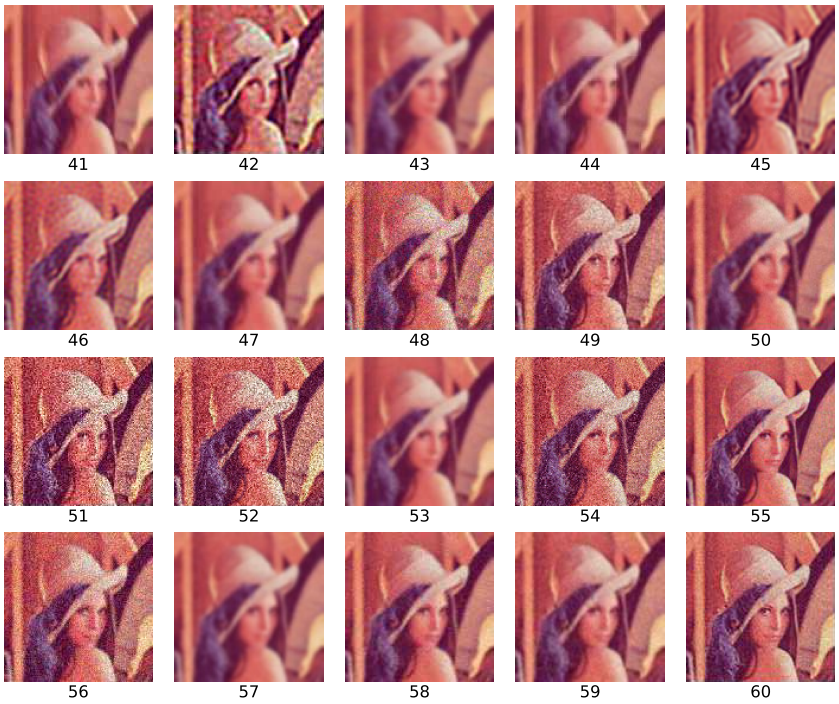


Figure 19: The visual results with the degradation parameters of cluster center [41, 60]. Best viewed in color.

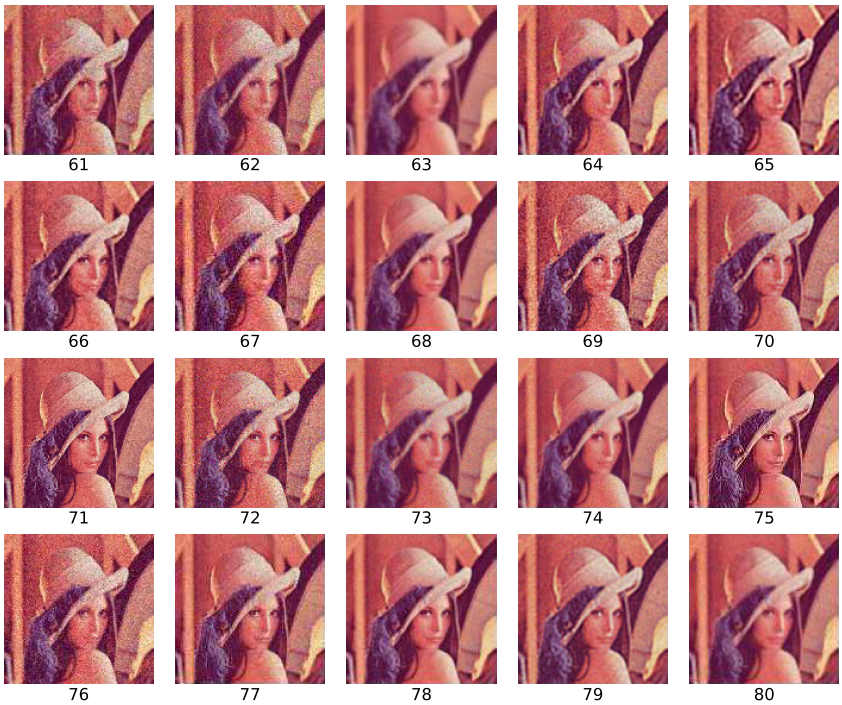


Figure 20: The visual results with the degradation parameters of cluster center [61, 80]. Best viewed in color.

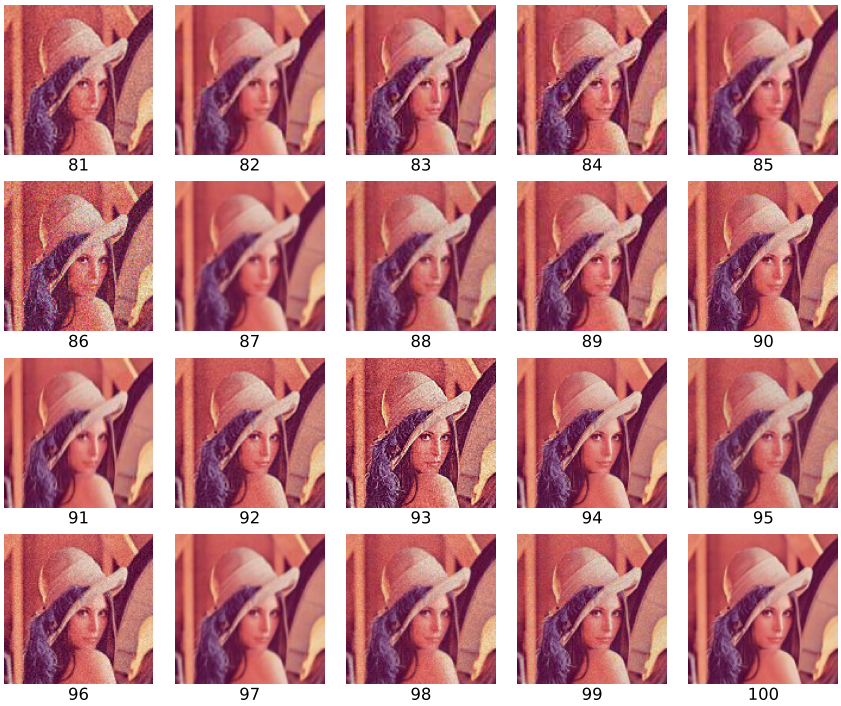


Figure 21: The visual results with the degradation parameters of cluster center [81, 100]. Best viewed in color.