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

Haoyu Chen[1], Wenbo Li[2], Jinjin Gu[3], Jingjing Ren[1], Haoze Sun[4],
Xueyi Zou[2], Zhensong Zhang[2], Youliang Yan[2], Lei Zhu[1,5*]

[1]The Hong Kong University of Science and Technology (Guangzhou) [2]Huawei Noah's Ark Lab [3]The University of Sydney
[4]Tsinghua University [5]The Hong Kong University of Science and Technology

Project page: https://haoyuchen.com/LWay

Abstract

For image super-resolution (SR), bridging the gap between the performance on synthetic datasets and real-world degradation scenarios remains a challenge. This work introduces a novel "Low-Res Leads the Way" (LWay) training framework, merging Supervised Pre-training with Self-supervised Learning to enhance the adaptability of SR models to real-world images. Our approach utilizes a low-resolution (LR) reconstruction network to extract degradation embeddings from LR images, merging them with super-resolved outputs for LR reconstruction. Leveraging unseen LR images for self-supervised learning guides the model to adapt its modeling space to the target domain, facilitating fine-tuning of SR models without requiring paired high-resolution (HR) images. The integration of Discrete Wavelet Transform (DWT) further refines the focus on high-frequency details. Extensive evaluations show that our method significantly improves the generalization and detail restoration capabilities of SR models on unseen real-world datasets, outperforming existing methods. Our training regime is universally compatible, requiring no network architecture modifications, making it a practical solution for real-world SR applications.

1. Introduction

Image super-resolution (SR) aims to restore high-resolution (HR) images from their low-resolution (LR) or degraded counterparts. The inception of the deep-learning-based SR model can be traced back to SRCNN [14]. Recently, advancements in deep learning models have substantially enhanced SR performance [1, 6, 8–10, 12, 25–27, 39, 51, 52, 54, 56], particularly in addressing specific degradation types like bicubic downsampling. Nevertheless, the efficacy of SR models is generally restricted by the degradation strategies employed during the training phase, posing great challenges in complex real-world applications.

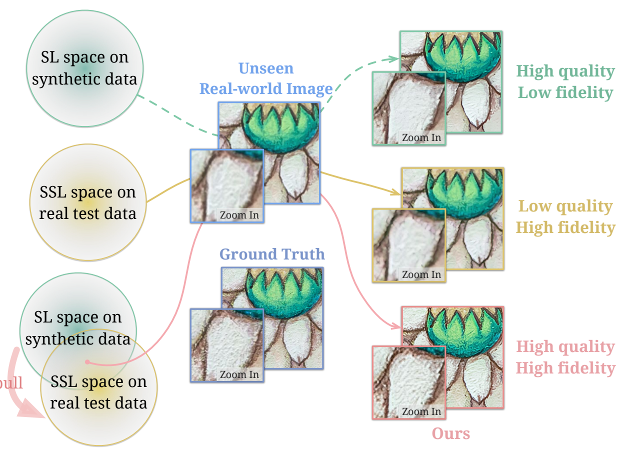

Figure 1. Our proposed training method combine the benefits of supervised learning (SL) on synthetic data and self-supervised learning (SSL) on the unseen test images, achieve high quality and high fidelity SR results.

In the realm of real-world SR, as shown in Figure 2, training approaches can primarily be categorized into three main paradigms. **(a) Unsupervised Learning with Unpaired Data**: Methods within this paradigm [2, 3, 15, 38, 40, 45, 46, 55] commonly utilize Generative Adversarial Networks (GAN) architecture to learn target distributions without paired data. Using one or multiple discriminators, they distinguish between generated images and actual samples, guiding the generator to model accurately. However, as this approach heavily relies on external data, it encounters significant challenges when facing scarce target domain data, particularly in real-world scenarios. The GAN framework for unsupervised learning also has some drawbacks. Firstly, it inherently struggles with stability during training, leading to noticeable artifacts in SR outputs. Secondly, it is difficult for a single 0/1 plane modelled by a discriminator to accurately separate the target domain [31]. This can result in imprecise distribution learning. **(b) Supervised Learning with Paired Synthetic Data**: BSRGAN [47] and Real-ESRGAN [42] have largely enhanced the SR model's

*Lei Zhu (leizhu@ust.hk) is the corresponding author.

generalization ability by simulating more realistic degradation. However, synthetic data, despite mimicking certain real-world conditions, inadequately captures the complex and variable nature of real scenarios, the gap between synthetic and real degradation persists. Consequently, the limited degradation patterns in synthetic data may lead to an over-smoothness issue, sacrificing crucial details and textures. Adapting effectively to complex, variable, or unknown degradations thus remains a formidable challenge. **(c) Self-supervised Learning with a Single Image**: Techniques falling within this category [11, 34, 37] leverage the intrinsic statistical characteristics of natural images, eliminating the necessity for external datasets. Generally, these methods enable self-supervised learning directly from the input LR image. Despite its inherent flexibility, this approach may exhibit reduced efficacy when handling images lacking repetitive patterns. As a result, in real-world scenarios, where necessary recurring structure are absent, these techniques tends to underperform compared to supervised learning methods that employ paired synthetic data.

It's notable that real LR/HR image pairs in the target domain are often prohibitively expensive or unavailable. Furthermore, a significant gap persists between synthesized data and real-world data. Given the intrinsic limitations of current methodologies, a critical question arises: Is there an approach that combines the strengths of these diverse strategies? In addressing this, we propose the novel "Low-Res Leads the Way" (LWay) training framework, which merges supervised learning (SL) pre-training with self-supervised learning (SSL) (see Figure 2 (d)). This approach aims to narrow the disparity between synthetic training data and real test images, as depicted in Figure 1. By integrating supervised learning's predictive capabilities with the ability to swiftly adapt to unique characteristics present in test LR images, this framework effectively produces high-quality results for unseen real-world images.

The initial step involves training an LR reconstruction network specifically designed to extract a degradation embedding from the LR image. This degradation embedding is then applied to the HR image, facilitating the regeneration of LR content. Upon encountering a test image, we derive its super-resolved result from an off-the-shelf SR model pre-trained on synthetic data. This output is fed into the fixed LR reconstruction network to produce the corresponding degraded counterpart. Subsequently, a self-supervised loss is computed by comparing this degraded counterpart to the original LR image, thereby updating specific parameters within the SR model. Given our observation that pre-trained SR models adeptly handle low-frequency domains but falter in high-frequency areas, we incorporate Discrete Wavelet Transform (DWT) to isolate high-frequency elements from the LR image. This component effectively shifts the model's focus to the recuperation

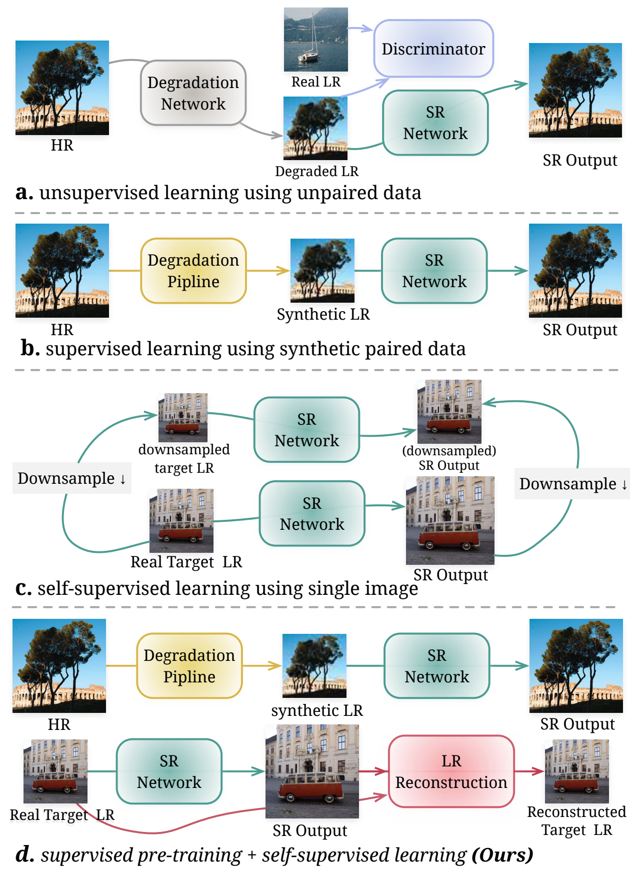

Figure 2. Comparison of different learning approaches for real-world image SR.

of high-frequency nuances, and avoids negative impacts on low-frequency areas.

With this innovative framework, our approach eliminates the need for paired LR/HR target domain images, significantly enhancing the performance of SL pre-trained models on unseen real-world data. Our method not only retains the essential content of LR images but also adds high-definition characteristics, ensuring a balance between fidelity and quality. Moreover, this training regime requires no modifications to the network architecture, offering broad compatibility across all SR models. Through extensive evaluations on real-world datasets, we have demonstrated our method's substantial improvements in generalization performance.

2. Related Work

2.1. Supervised Learning for Real-World SR

While recent years have witnessed significant advancements in the field of super-resolution (SR), conventional SR models such as SRCNN [14], VDSR [19], EDSR [29], RCAN [50], among others [1, 6, 9, 10, 12, 20, 23–27, 32, 51, 52, 54], have predominantly relied upon predefined degradation processes, such as bicubic downsampling. This simplification, while contributing to the theoretical understanding of SR, often falls short in capturing the intricate and diverse degradations inherent in real-world imaging scenarios, limiting practical adaptability across applications. Consequently, there is a pressing need to explore more sophisticated and realistic degradation models.

To this end, recent efforts have been directed toward methods capturing paired low-resolution (LR) and high-resolution (HR) images from real-world environments, as demonstrated by datasets like RealSR [4] and DRealSR [44]. However, these methods face challenges, including precise image alignment, complex hardware setups, and specific degradation characteristics (*e.g.*, Canon 5D3 and Nikon D810 cameras in RealSR), posing obstacles to practicality and scalability. Recent techniques, including Real-ESRGAN [42] and BSRGAN [47], have attempted to address these shortcomings by synthesizing LR images with more realistic degradation. Despite these advancements, a notable disparity persists between synthesized and authentic degradation. This often results in over-smoothed images that sacrifice fine textural details, as illustrated by [49]. Certain studies [7] have endeavored to enhance the generalizability using limited degradation data; however, the practical application scenarios remain restricted.

As a result, there is a growing demand for innovative approaches that are capable of adapting to the intricate and mixed degradation patterns that typify real-world applications. The SR results should not only exhibit high resolution but also encompass rich detail, ensuring fidelity.

2.2. Unsupervised Learning for Real-world SR

Unsupervised super-resolution [2, 3, 15, 38, 40, 45, 46, 55] serves as a technique to mitigate generation bias inherent in synthetic datasets. These approaches deviate from the conventional reliance on extensive paired data by harnessing the data-generating capabilities inherent in convolutional neural networks (CNNs). Ulyanov et al. [40] posited CNNs as implicit priors for capturing natural image statistics, a concept further explored by the Zero-Shot Super-Resolution (ZSSR) [37] model, which uniquely tailors SR algorithms to the repeating patterns within the input image itself. Generative Adversarial Networks (GANs) have significantly propelled the field forward. KernelGAN [2], for instance, aligns the statistical distribution of downscaled images with their original versions, enhancing the refinement of SR methods' outputs. CinCGAN [46] marks an early exploration into utilizing unpaired data for implicit degradation modeling. It employs a strategy that transforms LR images into noise-free 'clean' states through bicubic downsampling. This approach, backed by a dual CycleGAN architecture [55], fosters a cycle-consistent adaptation that eliminates the need for paired datasets. The unsupervised approach utilizing GANs also encompasses methods such as Degradation GAN [3], FSSR [15], DASR [45] and pseudo-supervision [33], which all employ discriminators to learn the distributions of HR or LR images, or even clean LR images. These methods are instrumental in constraining the network to transform the generated images to align with the corresponding distributions.

Despite considerable advancements in unsupervised methods, they still exhibit certain limitations. For instance, ZSSR and similar methods typically rely on the prerequisite assumption that images possess repetitive patterns. GAN-based approaches, in particular, require substantial data to fit certain specific degradation types effectively. They also face stability challenges during training, which often results in artifacts in SR outputs. Furthermore, the challenge for a discriminator to accurately distinguish the target domain using a binary (0/1) plane model can lead to imprecise learning of distributions. These constraints pose challenges to the practical utility of these methods in real-world scenarios. Exploring more generalized and flexible approaches becomes imperative.

3. Method

In the pursuit of practical applications for image SR, we introduce an unprecedented training methodology. This novel strategy marks a departure from established paradigms, fusing the precision of supervised pre-training with the innovation of self-supervised learning to address the complexities of real-world image degradation. Our proposed framework is detailed in Figure 3.

3.1. LR Reconstruction Pre-training

We introduce an LR reconstruction branch that plays a pivotal role in finetuning our SR model $\mathcal{S}$ on test images derived from real-world environments. Central to this process is the Degradation Encoder $\mathcal{E}$, engineered to distill the degradation signatures from LR images I_{LR} into a concise degradation embedding $\mathbf{e}$. The dimension is 512, formulated as $\mathbf{e} = \mathcal{E}(I_{\text{LR}})$. Subsequently, the Reconstructor $\mathcal{R}$ employs $\mathbf{e}$ and a high-resolution image I_{HR} to synthesize an estimated LR image $\hat{I}_{\text{LR}}$, aiming to fulfill $\hat{I}_{\text{LR}} = \mathcal{R}(I_{\text{HR}}, \mathbf{e})$. To ensure the integrity of $\mathbf{e}$, we incorporate a dual-component loss function $\mathcal{L}$, integrating both an L1 norm and the Learned Perceptual Image Patch Similarity (LPIPS) metric. The combined loss function is thus articulated as $\mathcal{L}(I_{\text{LR}}, \hat{I}_{\text{LR}}) = \mathcal{L}_1 + \mathcal{L}_{\text{LPIPS}}$, meticulously tuning the reconstruction fidelity. Notably, LR reconstruction branch has great robustness, requiring only minimal data for training, is precisely why we advocate for the inclusion of an LR reconstruction branch. This ensures that even when faced with new forms of degradation, its support in the finetuning of the SR model remains uncompromised. The efficiency and robustness of this approach, pivotal in our methodology, will be detailed and validated in the following sections.

3.2. Self-supervised Learning on Test Images

Our approach innovatively fine-tunes a subset of parameters in a SR network, specifically tailored for processing previously unseen real-world images. This method refines

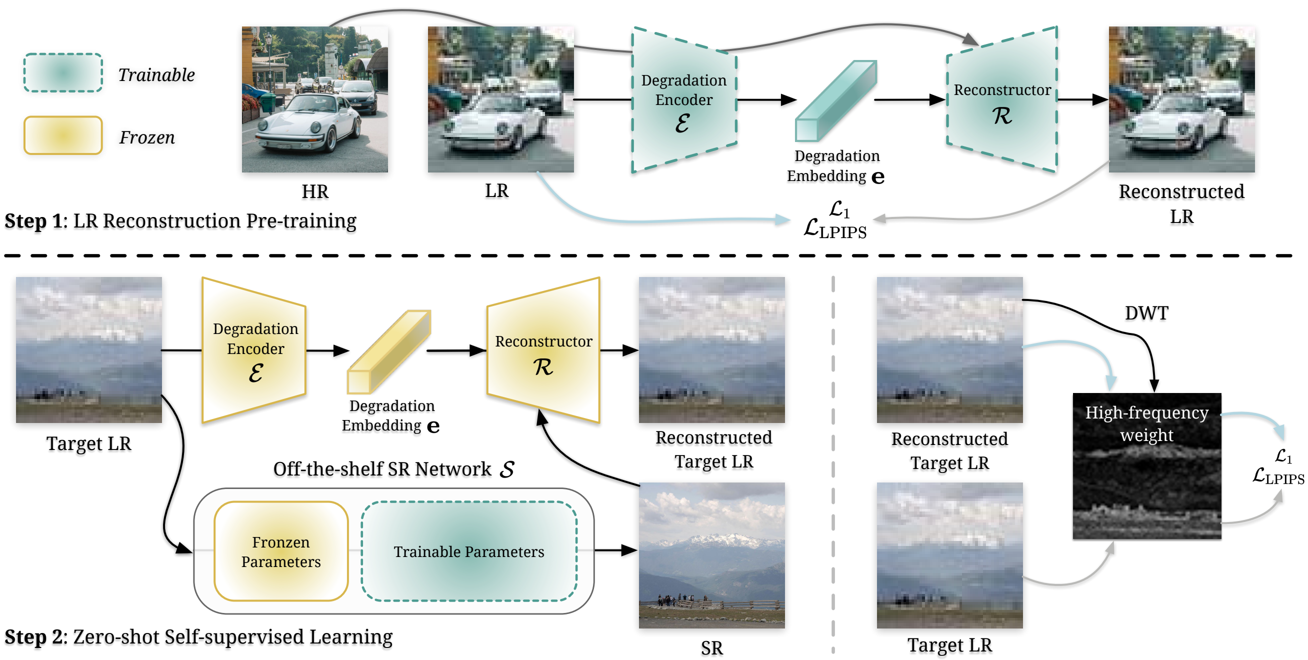

Figure 3. The proposed training pipeline (LWay) consists of two steps. In Step 1, we pre-train a LR reconstruction network to capture degradation embedding from LR images. This embedding is then applied to HR images, regenerating LR content. Moving to Step 2, for test images, a pre-trained SR model generates SR outputs, which are then degraded by the fixed LR reconstruction network. We iteratively update the SR model using a self-supervised learning loss applied to LR images, with a focus on high-frequency details through weighted loss. This refinement process enhances the SR model's generalization performance on previously unseen images.

the SR network to adeptly handle the complexities of actual degradation patterns. For an real-world LR test image $I_{\text{LR}}^{\text{test}}$, the SR network $\mathcal{S}$ initially produces a super-resolved image $I_{\text{SR}}^{\text{init}}$. The pre-trained LR reconstruction branch, with its parameters frozen, extracts a degradation embedding $\mathbf{e}^{\text{test}}$ from $I_{\text{LR}}^{\text{test}}$, expressed as $\mathbf{e}^{\text{test}} = \mathcal{E}(I_{\text{LR}}^{\text{test}})$. The self-supervised fine-tuning then commences, leveraging $I_{\text{SR}}^{\text{init}}$ and $\mathbf{e}^{\text{test}}$ to adjust a specific subset of the SR network's parameters θ_{ft}. This fine-tuning is formulated as an optimization problem:

$$\theta_{\text{ft}}^* = \arg\min_{\theta_{\text{ft}}} \mathcal{L}(\mathcal{R}(\mathcal{S}_\theta(I_{\text{LR}}^{\text{test}}), \mathbf{e}^{\text{test}}), I_{\text{LR}}^{\text{test}})\,,$$

where θ_{ft}^* is the optimized parameters from full model θ.

This strategic adjustment enhances the SR network's capability to reconstruct images with high fidelity to the LR inputs, enhances the SR network's ability to generalize to real-world degradation without the need for paired data.

Focused enhancement of high-frequency details. Conventional SR methods tend to proficiently reconstruct low-frequency regions but often neglect or inadequately restore high-frequency details. In addition, the low-frequency regions do not require LR reconstruction due to the absence of detailed texture. Therefore, our approach aims to concentrate the LR reconstruction process specifically on high-frequency areas, thereby preventing the introduction of artifacts into the low-frequency areas. Specifically, we apply Discrete Wavelet Transform (DWT) to obtain the high-frequency component, and then normalize it to yield a weight map $\mathbf{W} \in [0, 1]$. This weight map is then utilized to calculate a weighted loss, ensuring the fidelity to high-frequency details:

$$\mathcal{L} = \mathcal{L}_1(\mathbf{W} \odot \hat{I}_{\text{LR}}^{\text{test}}, \mathbf{W} \odot I_{\text{LR}}^{\text{test}}) + \mathcal{L}_{\text{LPIPS}}(\mathbf{W} \odot \hat{I}_{\text{LR}}^{\text{test}}, \mathbf{W} \odot I_{\text{LR}}^{\text{test}})\,,$$

where $\odot$ denotes element-wise multiplication. The combined loss effectively guides the network to restore high-frequency details with greater precision, improving the perceptual quality of the super-resolved image without compromising low-frequency content.

3.3. Discussion

By combining supervised learning (SL) on synthetic data with self-supervised learning (SSL) on test images with unknown degradation, we dynamically adjust the modeling space based on the intrinsic features of test images, steering the SL space towards a more precise SSL space. Figure 4 shows the effectiveness of our method during the fine-tuning process. Our method achieves high-quality and high-fidelity SR while maintaining general compatibility across all models. The primary advantages of our approach compared to other methods are included in the following:

General Degradation Modeling. The transformation from LR to HR images is recognized as a challenging task, while the reverse HR to LR transformation is comparatively simpler and more robust. Our method capitalizes on this observation, avoiding excessive reliance on extensive paired datasets. Instead, we opt to pre-train a universal degradation embedding extraction and LR reconstruction model. This characteristic ensures that our approach is not bound by as-

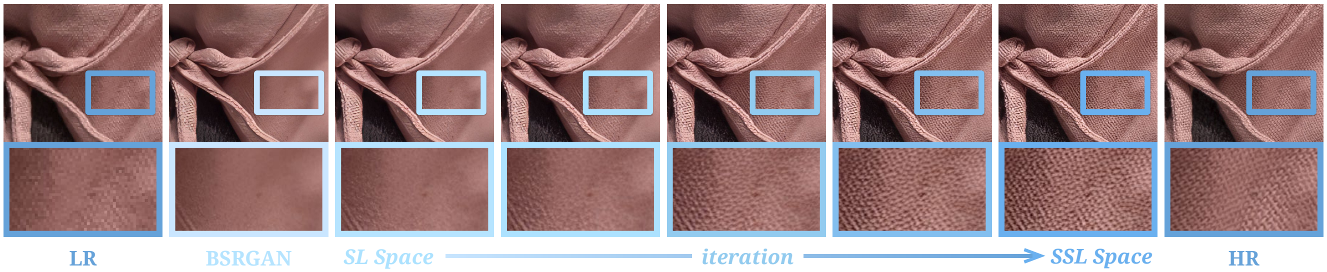

Figure 4. The SR model advances through the proposed fine-tuning iterations, moving from the supervised learning (SL) space of synthetic degradation to the self-supervised learning (SSL) space learned from test images. This results in enhanced SR quality and fidelity.

sumptions of uniform degradation across image datasets. During the training of the SR model, these parameters remain fixed, allowing the SR model for flexible adaption to unknown distributions in real-world scenarios. On the contrary, CycleGAN-based methods simultaneously learn the mappings from LR to HR and HR to LR. This process relies heavily on a substantial amount of data. Furthermore, because CycleGAN implicitly learns the HR to LR mapping without an explicit degradation extraction process, its underlying assumption is that the degradation across the entire dataset is consistent. Consequently, it can only fit certain degradation patterns, largely impacting its performance in real-world scenarios with limited data availability.

Dense pixelwise self-supervision. Through self-supervised learning, our method operates independently of external labels, leveraging dense LR pixel-level signals for supervision. This allows the model to learn richer texture features from the intrinsic image structure. This stands in contrast to traditional supervised approaches that rely on discriminators, which may learn inaccurate features due to the sparsity of supervision signals, leading to suboptimal results.

Robust regularization. Our approach can be viewed as a form of regularization constraint. By integrating degradation embedding extraction and decoupling it from the LR image reconstruction, our method maintains effectiveness in guiding the reconstruction process even when faced with imperfect degradation prediction. This substantially boosts the robustness of our approach, enabling it to learn rich and accurate texture information from the test images.

4. Experiments

4.1. Experimental Settings

Testing methods. Our proposed method serves as a universally applicable self-supervised learning strategy for various cutting-edge blind SR models, eliminating the necessity for architectural modifications. We conduct evaluations on a diverse range of advanced SR methods, including BSRGAN [47] and Real-ESRGAN+ [42] employing conventional CNN frameworks, SwinIR-GAN [27] integrating Transformer structures, FeMaSR [5] utilizing VQGAN, and StableSR [41] based on pre-trained diffusion. We use officially released SR models as baselines and conduct self-supervised fine-tuning on targeted test datasets. While fine-tuning a single image can lead to superior performance, for improved training efficiency, we opt to fine-tune the entire test dataset collectively. All experiments are conducted under this configuration unless otherwise specified.

Implementation details. We adopt the Adam [21] optimizer. For StableSR, we set the learning rate to 5e-5 and the batch size to 1. For the remaining models, a learning rate of 2e-6 and a batch size of 6 are used. Each model undergoes rapid fine-tuning on a single V100 GPU. The duration of training varies among models and images, typically spanning 150 to 500 iterations. *More details are provided in the supplementary materials.*

Training datasets. Our self-supervised fine-tuning approach is directly applied to the test set, without the need for a separate training set. The only prerequisite training is allocated for the LR reconstruction network, which is trained using 6,000 real paired images collected in-house. It is critical to note that these data were invisible to the SR network.

Testing datasets. Our method is evaluated on real-world paired datasets, including RealSR [4] and DRealSR [44]. These datasets are meticulously curated from diverse device sensors to reflect various degradation characteristics. To ensure a fair comparison with other methods, we follow the standard setting of cropping each image into multiple patches for a 4$\times$ SR. The LR image patch size is 128 $\times$ 128, while the corresponding HR size is 512 $\times$ 512.

Evaluation metrics. We employ LPIPS [48], DISTIS [13], and NLPD [17] metrics that closely align with human perception [16, 18]. Additionally, traditional metrics such as PSNR, SSIM [43], and MAD [22] are included for a comprehensive assessment. Six different metrics provide a comprehensive evaluation.

4.2. Improvements on Existing Methods

The results outlined in Table 1 compellingly demonstrate our method's effectiveness in significantly advancing SR quality. Notably, improvements are consistently observed across all models, datasets, and metrics, underscoring the universal applicability of our approach. For CNN-based models like Real-ESRGAN+, our method achieves a notable enhancement on the Nikon dataset, delivering a 1.77dB improvement in PSNR and a 0.0388 increase in SSIM. These improvements contribute to more precise re-

Dataset	Sensors	Metrics	CNN-based						Transformer-based			VQ-based			Diffusion-based		
			Real-ESRGAN+	+ LWay	Gain	BSRGAN	+ LWay	Gain	SwinIR-GAN	+ LWay	Gain	FeMaSR	+ LWay	Gain	StableSR	+ LWay	Gain
RealSR	Canon	PSNR ↑	27.51	29.18	+1.67	28.81	28.85	+0.04	28.12	28.96	+0.84	25.72	28.16	+2.44	25.50	27.22	+1.72
		SSIM ↑	0.8348	0.8688	+0.034	0.8473	0.8496	+0.0023	0.8486	0.8579	+0.0093	0.7811	0.8383	+0.0572	0.7684	0.8043	+0.0359
		LPIPS ↓	0.1947	0.1479	-0.0468	0.1988	0.1572	-0.0416	0.1850	0.1469	-0.0381	0.2543	0.1747	-0.0796	0.2636	0.2019	-0.0617
		MAD ↓	133.96	111.91	-22.05	119.08	116.77	-2.31	125.17	111.71	-13.46	143.38	117.48	-25.90	145.36	124.15	-21.21
		NLPD ↓	0.2807	0.2437	-0.037	0.2594	0.2569	-0.0025	0.2670	0.2541	-0.0129	0.3239	0.2778	-0.0461	0.3426	0.3074	-0.0352
		DISTIS ↓	0.1621	0.1444	-0.0177	0.1794	0.1558	-0.0236	0.1557	0.1352	-0.0205	0.2116	0.1808	-0.0308	0.1897	0.1596	-0.0301
	Nikon	PSNR ↑	26.81	28.58	+1.77	28.13	28.65	+0.52	27.54	28.55	+1.01	25.41	27.87	+2.46	25.54	26.92	+1.38
		SSIM ↑	0.7861	0.8249	+0.0388	0.8012	0.8057	+0.0045	0.8043	0.813	+0.0087	0.7314	0.7936	+0.0622	0.7370	0.7686	+0.0316
		LPIPS ↓	0.2300	0.1769	-0.0531	0.2302	0.1750	-0.0552	0.2154	0.176	-0.0394	0.2738	0.2028	-0.071	0.2711	0.2156	-0.0555
		MAD ↓	131.62	108.18	-23.44	118.48	105.64	-12.84	122.65	106.73	-15.92	137.54	110.79	-26.75	139.26	119.29	-19.97
		NLPD ↓	0.3061	0.2667	-0.0394	0.2805	0.2758	-0.0047	0.2844	0.272	-0.0124	0.3419	0.297	-0.0449	0.3513	0.3215	-0.0298
		DISTIS ↓	0.1950	0.1714	-0.0236	0.2102	0.1791	-0.0311	0.1842	0.1639	-0.0203	0.2340	0.2042	-0.0298	0.2131	0.1837	-0.0294
DRealSR	sony	PSNR ↑	30.16	31.4	+1.24	30.47	31.23	+0.76	29.92	30.77	+0.85	27.51	29.75	+2.24	28.63	29.28	+0.65
		SSIM ↑	0.8326	0.8597	+0.0271	0.8260	0.8442	+0.0182	0.8213	0.8398	+0.0185	0.7725	0.8096	+0.0371	0.7648	0.7785	+0.0137
		LPIPS ↓	0.2488	0.2341	-0.0147	0.2685	0.2469	-0.0216	0.2565	0.2383	-0.0182	0.3228	0.2931	-0.0297	0.3331	0.3017	-0.0314
		MAD ↓	125.20	112.1	-13.10	123.22	115.14	-8.08	124.85	114.09	-10.76	140.50	125.52	-14.98	141.13	130.01	-11.12
		NLPD ↓	0.3032	0.2751	-0.0281	0.3034	0.2857	-0.0177	0.3105	0.2895	-0.021	0.3502	0.3152	-0.035	0.3503	0.3402	-0.0101
		DISTIS ↓	0.1859	0.1765	-0.0094	0.2115	0.1934	-0.0181	0.1883	0.1783	-0.01	0.2314	0.2168	-0.0146	0.2296	0.2176	-0.012
	olympus	PSNR ↑	29.53	29.88	+0.35	29.16	29.4	+0.24	28.94	29.57	+0.63	26.42	28.26	+1.84	28.69	29.05	+0.36
		SSIM ↑	0.8050	0.8206	+0.0156	0.7931	0.7944	+0.0013	0.8002	0.8071	+0.0069	0.6976	0.7557	+0.0581	0.7460	0.7487	+0.0027
		LPIPS ↓	0.3107	0.308	-0.0027	0.3275	0.2926	-0.0349	0.3184	0.3093	-0.0091	0.4129	0.3762	-0.0367	0.3853	0.3800	-0.0053
		MAD ↓	127.91	125.04	-2.87	130.94	126.87	-4.07	131.73	126.04	-5.69	151.35	138.85	-12.50	137.60	132.71	-4.89
		NLPD ↓	0.3016	0.2899	-0.0117	0.3157	0.3129	-0.0028	0.3093	0.3005	-0.0088	0.3897	0.3425	-0.0472	0.3410	0.3353	-0.0057
		DISTIS ↓	0.2130	0.2118	-0.0012	0.2276	0.2145	-0.0131	0.2181	0.2109	-0.0072	0.2552	0.2406	-0.0146	0.2412	0.2371	-0.0041
	panasonic	PSNR ↑	29.81	30.83	+1.02	29.98	31.05	+1.07	29.11	30.94	+1.83	27.83	29.44	+1.61	29.13	29.88	+0.75
		SSIM ↑	0.8094	0.8283	+0.0189	0.7987	0.8236	+0.0249	0.7918	0.8193	+0.0275	0.7413	0.7798	+0.0385	0.7428	0.7554	+0.0126
		LPIPS ↓	0.2592	0.2581	-0.0011	0.2738	0.2624	-0.0114	0.2688	0.2517	-0.0171	0.3144	0.2973	-0.0171	0.3143	0.3021	-0.0122
		MAD ↓	124.51	116.18	-8.33	124.38	114.04	-10.34	126.61	112.79	-13.82	137.50	124.81	-12.69	132.36	122.85	-9.51
		NLPD ↓	0.304	0.2825	-0.0215	0.3109	0.2852	-0.0257	0.3184	0.2869	-0.0315	0.3604	0.3215	-0.0389	0.3444	0.3312	-0.0132
		DISTIS ↓	0.2000	0.1974	-0.0026	0.2130	0.2021	-0.0109	0.2046	0.1948	-0.0098	0.2243	0.2121	-0.0122	0.2255	0.2196	-0.0059

Table 1. The performance improvements across various model types utilizing our proposed training methodology.

# of Fine-tuning Images Per Model	Description	LPIPS ↓	DISTIS ↓	MAD ↓
0	baseline, without fine-tuning	0.3136	0.2353	117.71
1	fine-tuning on every single images	**0.2351**	**0.1919**	111.46
10	fine-tuning on the entire testset	0.2536	0.2044	111.63
50	fine-tuning with 40 additional images from the same sensors	0.2571	0.2037	**108.62**

Table 2. The impact of the number of images used for a single fine-tuning training. Our method can be fine-tuned either on individual images or on the entire test set, which greatly reduces cost.

construction of high-quality images. Furthermore, the validation of enhanced perceptual quality is evident through an LPIPS reduction of 0.0532. Additionally, when applied to Transformer models such as SwinIR-GAN, our method showcases considerable improvements. On the Olympus dataset, we observe a 0.63 dB increase in PSNR and a significant decrease in MAD by 5.69, highlighting the framework's capacity to enhance fidelity and sharpness.

As depicted in Figure 5, in the first example, all SR models fail to preserve the original textures present in the input images, resulting in excessively smoothed fabric patterns. However, upon applying our self-supervised fine-tuning method, significant improvements are observed across all approaches, successfully reconstructing clear fabric textures. A similar improvement is evident in the second example of oil paintings. The existing SR models struggle to capture the intricate details of the paintings. Conversely, our method effectively restores the artistic effects, particularly showcasing notable enhancement for StableSR. For other examples, the results are similar as well, our method significantly improving high-frequency detail recovery, yielding results that were both sharp and rich in detail.

4.3. Application on Real-world Scenes

Old films often exhibit issues like graininess, color fading, and lower resolution, making them an ideal testbed for evaluating the practical capabilities of SR models. To conduct a comprehensive comparison, we curate a selection of state-of-the-art real-world SR models. These encompass various methodologies: ZSSR [37], a self-supervised learning model; DASR [28], a degradation-adaptive approach; large diffusion models such as LDM [36], DiffBIR [30], and StableSR [41]; DARSR [53], which leverages unsupervised techniques for enhanced model performance; and CAL_GAN [35], a photo-realistic SR model. We employ StableSR as the base model and implement the proposed self-supervised learning strategy. The first case in Figure 6 involves a 480p low-resolution film, namely "My Fair Lady". Among the assessed models, ZSSR, DASR, and DARSR exhibit minimal improvements, while DiffBIR introduces unpleasing artifacts. Other models achieve slightly smoother results. Notably, our model not only accurately reproduces the hat with clear fabric textures but also effectively restores facial features, including wrinkles and contours. In contrast to some methods that may introduce unnatural effects or overly smooth distortions, our model adeptly balances the restoration of fine textures with preserving overall image clarity.

User study. We conducted a user study with the participation of 24 experienced researchers. Each participant was tasked with assigning a visual perceptual quality score rang-

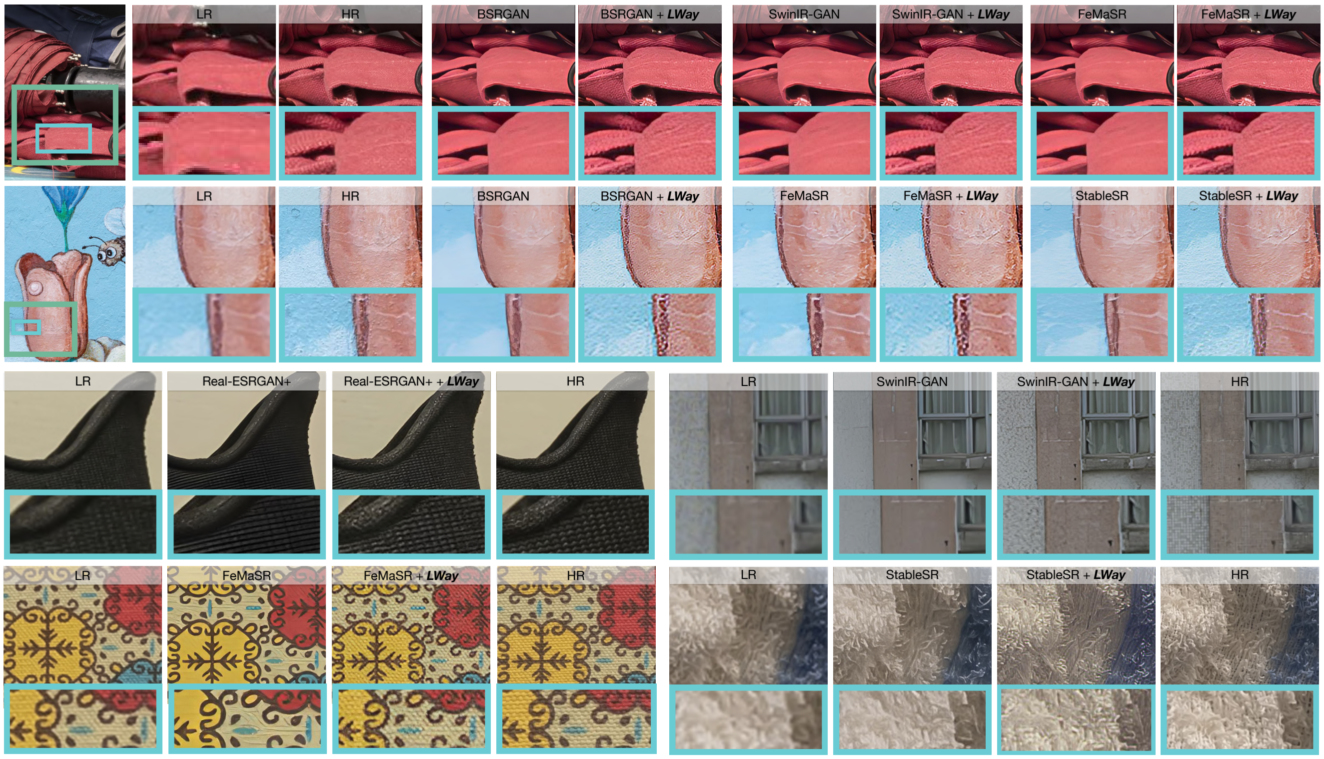

Figure 5. Qualitative comparisons on real-world datasets. The content within the blue box represents a zoomed-in image.

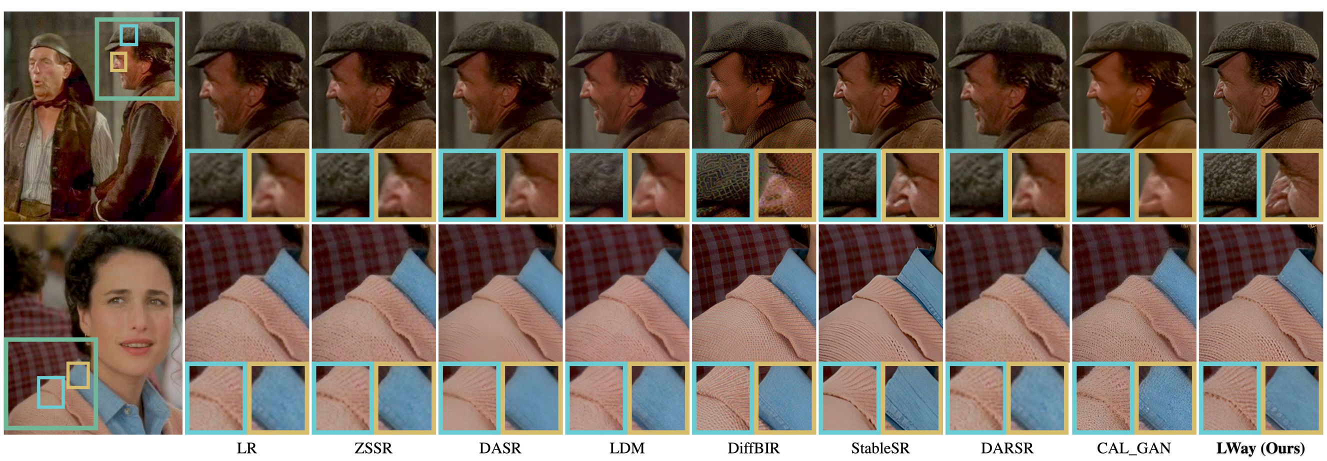

Figure 6. Qualitative comparisons on two old films.

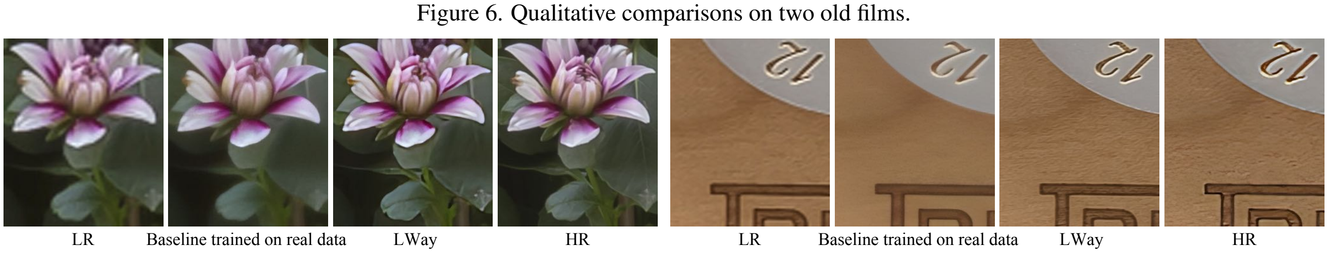

Figure 7. Supervised fine-tuning a baseline model on one real dataset doesn't perform well on another due to dataset gaps. Our proposed method self-supervised fine-tuned model for specific test images achieves superior performance.

ing from 0 to 10 to every image. The results, depicted in the Figure 8, reveal a significant lead of our proposed method over alternative approaches, surpassing the second-best method by more than 2 points. Notably, the scores for DASR, DiffBIR, and DARSR were even lower than those for LR images, indicating a limited effectiveness of these methods in handling real-world images.

Training Type	Number of Sensors	Number of Images	LPIPS ↓	DISTIS ↓
- (baseline)	-	-	0.2302	0.2102
Synthetic Data	-	2K	0.1836	0.1885
	-	6k	0.1816	0.1873
Real-world Data	1	0.6K	0.2003	0.1970
	2	2K	0.1785	0.1793
	2	4K	**0.1722**	**0.1772**
	3	6K	0.1800	0.1830

Table 3. Ablation on training data of LR reconstruction.

	baseline	128	256	512	1024	**2048**	4096
PSNR↑	28.13	28.92	28.54	28.85	29.10	**29.56**	29.20
LPIPS↓	0.2302	0.1804	0.1776	0.1722	0.1736	**0.1629**	0.1669
DISTS↓	0.2192	0.1792	0.1818	0.1772	0.1749	**0.1630**	0.1656

Table 4. Ablation on dimensions of degradation embedding.

Method	LPIPS ↓	DISTIS ↓
baseline	0.2302	0.2102
baseline + real data	0.2268	0.1989
LWay (ours)	**0.1722**	**0.1772**

Table 5. Our method versus supervised real data fine-tuning.

HF Loss	LPIPS ↓	DISTIS ↓
×	0.1858	0.1879
✓	0.1722	0.1772

Table 6. Ablation study on high-frequency (HF) loss.

4.4. Ablation Study

We conducted an ablation study on the RealSR Nikon test set using BSRGAN. We trained 65% of the model parameters to achieve the lowest LPIPS score on this test set.

Training data of LR reconstruction. In this section, we demonstrate the robustness of the LR reconstruction network trained with limited data, which forms the cornerstone of our design. As depicted in Table 3, we incorporated two types of training data. The first category includes synthetic data created using BSRGAN degradation, while the second involves real paired images collected for training. Both settings result in improved performance. Specifically, compared to synthetic data, which brings a 0.0486 improvement in LPIPS, the utilization of only 600 images brings a 0.0299 improvement, and 4000 images notably boosts LPIPS by 0.058. Adding more images beyond this threshold did not yield any further advancement. We attribute this to the inherent ease in mapping from HR to LR compared to the reverse LR to HR mapping, mitigating the necessity for extensive training data. This assertion finds further support in Figure 9, where t-SNE visualization distinctly separates distinct degradations, even for unseen degradation types.

Degradation embedding dimensions. Table 4 tests different embedding dimensions, all variants significantly enhance performance. While a dimension of 512 (default) is effective, higher one (2048) can further improve results.

Our method versus supervised fine-tuning. To comprehensively illustrate the efficacy of our method, we conduct additional supervised fine-tuning of the baseline model using the gathered real paired data. As depicted in Table 5, we note marginal improvements. This aligns with our contention that LR to HR mapping poses inherent difficulties. Training with data from one sensor type showed negligible benefits for another, suggesting a significant gap in degradation patterns. This was further corroborated by Figure 7,

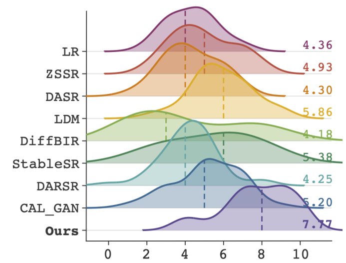

Figure 8. User study on the visual perceptual quality of results from different methods on real images.

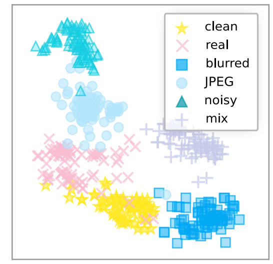

Figure 9. t-SNE visualization of embeddings from LR degradation encoder.

where it generates over-smoothed outputs. Conversely, our method showcases robustness and substantially enhances the final SR quality, and is more effective.

Number of images used in fine-tuning. We employ self-supervised LR reconstruction fine-tuning on test images to optimize the SR model. This section investigates the impact of the number of fine-tuning images on the final performance. As indicated in Table 2, we establish a baseline without fine-tuning using ten real-world images. Conducting single-shot fine-tuning on individual images yields the most favorable results, allowing models to best adapt to the distribution of input images. Next, we conduct experiments involving collective fine-tuning of ten images. Results show significant improvements compared to the baseline but are not as effective as fine-tuning individual images separately. Furthermore, we extend our study by fine-tuning the model using an additional forty images to investigate whether acquiring more images from the same sensor would refine the model further. Our findings indicate that compared to training on ten images, there is a decline in LPIPS, while DISTIS and MAD exhibit slight improvements. This suggests a trade-off between fine-tuning performance and efficiency.

High-frequency loss. Table 6 illustrates the impact of the introduced high-frequency loss. The integration of the high-frequency loss results in a notable improvement. Importantly, it enhances high-frequency recovery and avoids the negative impact of our method on low-frequency areas.

5. Conclusion

In conclusion, our proposed super-resolution training strategy, termed “Low-Res Leads the Way”, represents an innovative approach that successfully bridges the disparity between synthetic data supervised training and real-world test image self-supervision. Demonstrating impressive performance and robustness across various SR frameworks and real-world benchmarks, our method marks a advancement toward achieving effective real-world applications.

Acknowledgment. This work is supported by Guangzhou Municipal Science and Technology Project (Grant No. 2023A03J0671), Guangzhou-HKUST(GZ) Joint Funding Program (No. 2024A03J0618), and the InnoHK funding launched by Innovation and Technology Commission, Hong Kong SAR.

References

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 252–268, 2018. 1, 2

[2] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3

[3] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *Proceedings of the European conference on computer vision (ECCV)*, pages 185–200, 2018. 1, 3

[4] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2019. 3, 5

[5] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1329–1338, 2022. 5

[6] Haoyu Chen, Jinjin Gu, and Zhi Zhang. Attention in attention network for image super-resolution. *arXiv preprint arXiv:2104.09497*, 2021. 1, 2

[7] Haoyu Chen, Jinjin Gu, Yihao Liu, Salma Abdel Magid, Chao Dong, Qiong Wang, Hanspeter Pfister, and Lei Zhu. Masked image training for generalizable deep image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1703, 2023. 3

[8] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, pages 22367–22377, 2023. 1

[9] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12312–12321, 2023. 2

[10] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, and Xiaokang Yang. Recursive generalization transformer for image super-resolution. In *International Conference on Learning Representations (ICLR)*, 2024. 1, 2

[11] Xi Cheng, Zhenyong Fu, and Jian Yang. Zero-shot image super-resolution with depth guided internal degradation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 265–280. Springer, 2020. 2

[12] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019. 1, 2

[13] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 5

[14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1, 2

[15] 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, 3

[16] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy Ren, and Chao Dong. Image quality assessment for perceptual image restoration: A new dataset, benchmark and metric. *arXiv preprint arXiv:2011.15002*, 2020. 5

[17] Alexander Hepburn, Valero Laparra, Ryan McConville, and Raul Santos-Rodriguez. Enforcing perceptual consistency on generative adversarial networks by using the normalised laplacian pyramid distance. *arXiv preprint arXiv:1908.04347*, 2019. 5

[18] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 633–651. Springer, 2020. 5

[19] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 2

[20] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016. 2

[21] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015. 5

[22] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006–011006, 2010. 5

[23] Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, and Jiaya Jia. Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. *NeurIPS*, 33:20343–20355, 2020. 2

[24] Wenbo Li, Kun Zhou, Lu Qi, Liying Lu, and Jiangbo Lu. Best-buddy gans for highly detailed image super-resolution. In *AAAI*, pages 1412–1420, 2022.

[25] Wenbo Li, Xin Lu, Shengju Qian, and Jiangbo Lu. On efficient transformer-based image pre-training for low-level vision. In *IJCAI-23*, pages 1089–1097. International Joint Conferences on Artificial Intelligence Organization, 2023. Main Track. 1

[26] Zheyuan Li, Yingqi Liu, Xiangyu Chen, Haoming Cai, Jinjin Gu, Yu Qiao, and Chao Dong. Blueprint separable residual network for efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 833–843, 2022.

[27] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1, 2, 5

[28] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *European Conference on Computer Vision*, pages 574–591. Springer, 2022. 6

[29] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2

[30] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 6

[31] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind image super-resolution: A survey and beyond. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5461–5480, 2022. 1

[32] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2359–2368, 2020. 2

[33] Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 291–300, 2020. 3

[34] Reyhaneh Neshatavar, Mohsen Yavartanoo, Sanghyun Son, and Kyoung Mu Lee. Icf-srsr: Invertible scale-conditional function for self-supervised real-world single image super-resolution. *arXiv preprint arXiv:2307.12751*, 2023. 2

[35] JoonKyu Park, Sanghyun Son, and Kyoung Mu Lee. Content-aware local gan for photo-realistic super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10585–10594, 2023. 6

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 6

[37] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3118–3126, 2018. 2, 3, 6

[38] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3516–3525, 2020. 1, 3

[39] Haoze Sun, Wenbo Li, Jianzhuang Liu, Haoyu Chen, Renjing Pei, Xueyi Zou, Youliang Yan, and Yujiu Yang. Coser: Bridging image and language for cognitive super-resolution. *arXiv preprint arXiv:2311.16512*, 2023. 1

[40] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018. 1, 3

[41] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. In *arXiv preprint arXiv:2305.07015*, 2023. 5, 6

[42] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 1, 3, 5

[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[44] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 101–117. Springer, 2020. 3, 5

[45] Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, and Hengjie Song. Unsupervised real-world image super resolution via domain-distance aware training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13385–13394, 2021. 1, 3

[46] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 701–710, 2018. 1, 3

[47] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 1, 3, 5

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[49] Ruofan Zhang, Jinjin Gu, Haoyu Chen, Chao Dong, Yulun Zhang, and Wenming Yang. Crafting training degradation distribution for the accuracy-generalization trade-off in real-world super-resolution. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 3

[50] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of*

the European conference on computer vision (ECCV), pages 286–301, 2018. 2

[51] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 1, 2

[52] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 1, 2

[53] Hongyang Zhou, Xiaobin Zhu, Jianqing Zhu, Zheng Han, Shi-Xue Zhang, Jingyan Qin, and Xu-Cheng Yin. Learning correction filter via degradation-adaptive regression for blind single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12365–12375, 2023. 6

[54] Lin Zhou, Haoming Cai, Jinjin Gu, Zheyuan Li, Yingqi Liu, Xiangyu Chen, Yu Qiao, and Chao Dong. Efficient image super-resolution using vast-receptive-field attention. In *European Conference on Computer Vision*, pages 256–272. Springer, 2022. 1, 2

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 3

[56] Wenbin Zou, Tian Ye, Weixin Zheng, Yunchen Zhang, Liang Chen, and Yi Wu. Self-calibrated efficient transformer for lightweight super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 930–939, 2022. 1