Highlights

Post-Processing Ensemble Framework for Balancing Fidelity and Perception in Super-Resolution

Dong-Yun Kim, Jae-Ho Nah

- A novel, lightweight post-processing ensemble framework for Super-Resolution.
- Balances the trade-off between image fidelity (PSNR/SSIM) and perceptual quality.
- Effectively mitigates color, structure, and facial artifacts from perception-oriented models.
- Requires no model retraining, making it a versatile plug-and-play solution.

Post-Processing Ensemble Framework for Balancing Fidelity and Perception in Super-Resolution⋆

Dong-Yun Kim[a], Jae-Ho Nah[a,*]

[a]Dept. of Computer Science, Sangmyung University, 20, Hongjimun 2-gil, Jongno-gu, Seoul, 03016, Republic of Korea

ARTICLE INFO

Keywords:
Super-resolution
Ensemble Methods
Image Enhancement
Artifact Correction

ABSTRACT

Single Image Super-Resolution (SISR) methods are often specialized, excelling either in fidelity (high PSNR) or perceptual quality (realistic textures), but rarely both. Fidelity-oriented models tend to produce blurry results, while perception-oriented models often introduce undesirable artifacts. To address this trade-off, we propose a novel post-processing ensemble framework that synergistically combines the outputs of two complementary SR models without requiring any retraining. Our method leverages the structural integrity of a fidelity-focused model (e.g., SwinIR classical) and the textural detail of a perception-focused model (e.g., SwinIR real-world). The proposed pipeline involves three stages: (1) color stabilization via histogram matching to correct chromatic distortions from the perceptual model; (2) structure-aware blending using an edge mask to merge sharp textures into structurally coherent regions; and (3) facial fidelity enhancement using a face mask to mitigate identity-distorting artifacts. The experiments demonstrate that our method achieves a superior balance between fidelity and perception. It notably enhances perceptual quality compared to both baseline models, while remaining competitive in terms of fidelity, as evaluated on a comprehensive set of full-reference and no-reference image quality assessment metrics.

1. Introduction

Single Image Super-Resolution (SISR) aims to reconstruct a high-resolution (HR) image from a single low-resolution (LR) counterpart. It is a fundamental task in computer vision with wide-ranging applications in medical imaging, satellite surveillance, digital photography, and texture mapping in 3D graphics [1]. The advent of deep learning has led to remarkable progress, with models generally categorized into fidelity-oriented and perception-oriented approaches.

Fidelity-based methods, often optimized with pixel-wise losses like L_1 or MSE, excel at achieving high Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [2] scores [3, 4]. Recent Transformer-based architectures like SwinIR [5] have pushed the boundaries of these fidelity metrics even further. However, this pixel-level accuracy often comes at the cost of perceptual quality, as the averaging nature of these loss functions tends to produce overly smooth and blurry textures, failing to restore fine, realistic details [6].

Conversely, perception-based methods, primarily driven by Generative Adversarial Networks (GANs) and diffusion models, prioritize visual realism [6, 7]. By employing adversarial and perceptual losses, or iterative denoising processes, these models generate images with sharp details and convincing textures. Advanced GAN-based techniques like BSRGAN [8] and Real-ESRGAN [9], as well as recent diffusion-based models like TSD-SR [10], have pushed the boundaries of perceptual quality. However, this realism is often accompanied by a significant drawback: the generation of undesirable artifacts [11]. As shown in Figure 1, perception-oriented models can introduce structural distortions, color shifts, and unnatural patterns, particularly when processing real-world images with complex degradations. Similarly, diffusion models frequently exhibit structural distortions, particularly when reconstructing repetitive patterns such as grid-like structures.

This dichotomy presents a critical challenge: no single model offers an optimal solution for all scenarios. While recent trends favor developing complex, learnable neural fusion modules to improve outputs, these often introduce high computational costs and require extensive retraining. Instead, we propose a novel, lightweight post-processing ensemble framework built upon classical computer vision techniques. By relying on established, parameter-efficient methods rather than a heavyweight AI-driven fusion module, our approach ensures low computational overhead, interpretability, and true plug-and-play capability without any retraining. Our method synergistically combines the outputs of a fidelity-oriented model (e.g., SwinIR classical) and a perception-oriented model (e.g., SwinIR real-world).

Our contributions are:

- A multi-stage ensemble pipeline that operates purely in the post-processing domain, relying on classical computer vision efficiency and requiring no model retraining.
- A combination of histogram matching, edge-aware blending, and face-aware correction to effectively reduce color, structure, and facial artifacts.

⋆This document is a revised version of a user-provided manuscript, enhanced for clarity, structure, and academic rigor.

*Corresponding author
edw0728@gmail.com (D. Kim); jaeho.nah@smu.ac.kr (J. Nah)
https://tama0728.github.io/ (D. Kim); https://nahjaeho.github.io/ (J. Nah)
ORCID(s): 0009-0002-8496-0756 (D. Kim); 0000-0001-7805-5333 (J. Nah)

- A comprehensive evaluation demonstrating that our method provides a superior balance of fidelity and perceptual quality, confirmed by both full-reference and no-reference metrics.

2. Related Works

2.1. CNN-based SR

Convolutional Neural Networks (CNNs) laid the foundation for deep learning in SR. SRCNN [3] pioneered this direction with a simple three-layer network, introducing an end-to-end mapping from interpolated LR images to HR images. Subsequent works like VDSR [4] and DRCN [12] explored deeper architectures to improve performance. While effective at extracting local features, CNN-based methods often struggle to model long-range dependencies, leading to a loss of textural detail, especially at higher scaling factors.

2.2. Transformer-based SR

To capture global context, recent works have adapted the Transformer architecture, originally from NLP, for vision tasks. The Vision Transformer (ViT) [13] demonstrated the potential of self-attention for image recognition. In image restoration, IPT [14] and SwinIR [5] models (classical and lightweight) have set new state-of-the-art benchmarks. SwinIR, based on the Swin Transformer [15], uses shifted windows for efficient and effective modeling of long-range dependencies. Other notable models like DAT [16] and HAT [17] have further improved performance. DRCT [18] addresses the information bottleneck observed in deep networks like SwinIR and HAT—where the intensity distribution of feature maps decreases sharply as the network deepens by using inter-layer Dense-residual Connections to alleviate spatial information loss. Despite their strength in capturing global context, Transformer models optimized with pixel-wise losses can still produce results that lack perceptual quality.

2.3. GAN-based SR

GANs address texture issues by learning realistic image distributions via adversarial training. SRGAN [6] first applied GANs to SR with perceptual loss, and ESRGAN [7] further improved performance with enhanced architecture and training strategies. To handle real-world degradations, BSRGAN [8], Real-ESRGAN [9], and SwinIR [5] (for real-world SR) adopt complex degradation modeling. Beyond general SR, GAN-based methods such as GFPGAN [19] focus on face restoration using generative facial priors. However, these methods often produce artifacts, including structural distortions and unnatural textures. While DeSRA [11] mitigates this via a detect-and-delete mechanism, its effectiveness is limited to known artifact types, leaving artifacts a key challenge for practical deployment.

2.4. Diffusion-based SR

Diffusion models have emerged as a compelling alternative to GANs for SR, iteratively denoising a signal to synthesize realistic outputs via pre-trained generative priors. Recent works have focused on reducing the inference cost of this paradigm. TSD-SR [10] proposes a one-step diffusion model using target score distillation, attaining strong perceptual quality without iterative sampling, while PiSA-SR [20] introduces pixel-level and semantic-level adaptation to adaptively balance fidelity and perceptual quality.

2.5. Ensemble-based SR

Ensemble learning, which combines multiple models to achieve better performance, has also been applied to SR. The effectiveness of an ensemble relies on its models being both accurate and diverse, meaning they should possess different strengths and weaknesses [21]. Early work involved ensembling Sparse Coding Networks [22]. Jiang et al. [23] proposed an ensemble framework based on Maximum A Posteriori (MAP) estimation, weighting individual SR models by their PSNR and SSIM scores. While effective, these methods often increase computational complexity and model size, and some require retraining or fine-tuning, limiting their flexibility. Our work differs by proposing a post-processing ensemble that is lightweight and model-agnostic.

3. Motivation and Preliminaries

3.1. Complementary Nature of SR Models

Modern SR models exhibit a clear performance trade-off, which forms the motivation for our work.

Fidelity-Oriented Models (e.g., SwinIR classical, PFT-SR [24]) These models are trained with pixel-wise losses (L_1 or L_2) to maximize metrics like PSNR and SSIM. As a result, they excel at preserving the global structure, contours, and overall layout of the original image. However, their tendency to average possible solutions leads to overly smooth textures and a lack of high-frequency detail, as illustrated in Figure 1.

Perception-Oriented Models (e.g., SwinIR real-world, TSD-SR [10]) These models are trained with a combination of adversarial, perceptual, and pixel-wise losses. This encourages the generator to produce statistically realistic textures, resulting in visually sharper and more detailed images. However, this process is not perfectly constrained, often leading to undesirable artifacts [25]:

- **Color Distortion:** The model may generate colors that are inconsistent with the original image.
- **Structural Distortion:** Straight lines can become warped, and geometric patterns may be unfaithfully reconstructed (see Fig. 1).
- **Facial Artifacts:** Faces are a particularly challenging area in which perception-oriented models can produce distorted or unnatural features, severely impacting perceived quality.

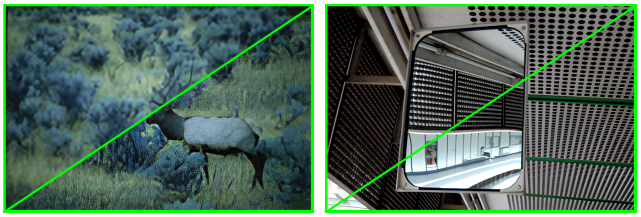

Figure 1: Comparison of SwinIR classical (top-left) and real-world (bottom-right) models. The classical model preserves structures but lacks texture, resulting in a loss of detail (Left: BSD100 38082), while the real-world model maintains texture but can introduce distortions, such as circular holes reconstructed as squares (Right: Urban100 img004).

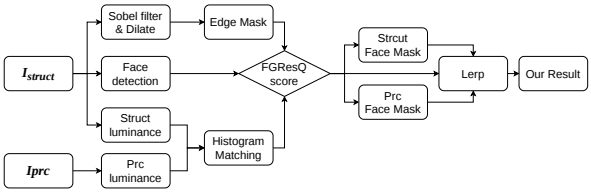

Figure 2: The proposed ensemble pipeline. The process starts with two inputs, I_{struct} and I_{prc}. First, color stabilization is applied to I_{prc} via histogram matching. Second, an edge mask derived from I_{struct} is used to guide the structure-aware blending of the two images. Finally, a face mask is utilized to adaptively correct facial artifacts, producing the final output.

3.2. Problem Formulation

Given a low-resolution input I_{LR}, we use two pre-trained SR models to generate two intermediate HR images:

- I_{struct}: A structurally faithful image from a fidelity-oriented model.
- I_{prc}: A perceptually rich image from a perception-oriented model.

In our primary experiments, I_{struct} and I_{prc} are generated by SwinIR classical and SwinIR real-world. To demonstrate the versatility of our framework, we also evaluate alternative pairings using PFT-SR as the structural baseline combined with TSD-SR (a diffusion-based model).

The goal of our framework is to synthesize a final output image $\hat{I}$ that combines the strengths of both inputs:

$$\hat{I} = \mathcal{F}(I_{struct}, I_{prc})$$

where $\mathcal{F}$ is our proposed multi-stage ensemble function. This function aims to maximize both fidelity and perceptual quality by preserving the structure from I_{struct} while selectively incorporating the valid textures from I_{prc} and correcting its artifacts.

4. Proposed Method

Our proposed framework is a three-stage pipeline designed to intelligently merge the outputs of fidelity- and perception-oriented SR models. The overall process is illustrated in Figure 2.

4.1. Stage 1: Color Stabilization via Histogram Matching

To address the color distortion often present in perception-oriented models' outputs (I_{prc}), we align its luminance distribution to that of the more faithful I_{struct}. Rather than applying a full per-channel CDF-based histogram mapping, we adopt a *Reinhard Color Transfer* [26] approach operating in the LAB color space, which corrects tonal shifts without introducing quantization noise or banding artifacts.

The process is as follows:

1. Both I_{struct} and I_{prc} are converted from BGR to the CIE LAB color space and cast to floating-point.
2. The mean μ and standard deviation σ of the L (luminance) channel are computed for each image.
3. The L channel of I_{prc} is linearly rescaled so that its distribution matches that of I_{struct}:

$$L' = \left(L_{prc} - \mu_{prc}\right) \cdot \frac{\sigma_{struct}}{\sigma_{prc}} + \mu_{struct}$$

4. The rescaled L' is clipped to $[0, 255]$, and the image is converted back to BGR to produce the color-corrected image $\tilde{I}_{prc}$.

By operating solely on the luminance channel, this approach corrects brightness and contrast discrepancies while leaving the a^* and b^* chroma channels of I_{prc} intact, preserving its perceptual texture without introducing color noise. This step ensures that the subsequent blending stages operate on images with consistent tonal characteristics.

4.2. Stage 2: Structure Preservation via Edge-Aware Blending

The core of our method is to selectively blend the detailed textures of $\tilde{I}_{prc}$ into the structurally sound I_{struct}. We hypothesize that the structural integrity is most critical along edges and contours, while flat regions are more tolerant to texture distortion.

To achieve this, we create a soft edge mask M_e from the structurally reliable image, I_{struct}:

1. We apply a 3×3 Sobel filter to the grayscale version of I_{struct} to compute the gradient magnitude, highlighting the edges.
2. The resulting edge map is dilated using a 7×7 kernel. Dilation thickens the edges, creating a smoother transition zone for blending and ensuring that the areas immediately surrounding contours are also preserved from I_{struct}.
3. The mask is normalized to the range [0, 1].

Using this mask, we perform a linear interpolation between the two images (to get I_S, a structure-preserved image):

$$I_S = M_e \odot I_{struct} + (1 - M_e) \odot \tilde{I}_{prc}$$

where $\odot$ denotes element-wise multiplication. This operation preserves the pixels of I_{struct} in the edge regions ($M_e \approx 1$) while seamlessly introducing the textures from $\tilde{I}_{prc}$ in the non-edge regions ($M_e \approx 0$).

4.3. Stage 3: Face-Aware Correction

Perception-oriented models often distort facial features due to insufficient structural constraints. To address this, we use the MediaPipe Face Detector [27] to selectively perform *per-face quality comparison* using the FGResQ metric [28], adaptively blending the optimal facial features.

For each detected face, we extract crops $(I^f_{struct}, I^f_{prc}, I^f_S)$ using a bounding box expanded by 50 % to encompass the entire head. We then calculate their respective FGResQ scores $(q_{struct}, q_{prc}, q_s)$. If the intermediate edge-blended result I_S yields the highest score (q_s), no correction is applied.

Otherwise, we generate a 2D Gaussian kernel ($\sigma = \frac{1}{3}$ of the crop dimension) and accumulate it into one of two global masks, depending on which source model provides superior quality:

- M_{struct}: updated if $q_{struct} > q_{prc}$, steering the correction toward the fidelity-oriented I_{struct}.
- M_{prc}: updated if $q_{prc} \geq q_{struct}$, steering the correction toward the perception-oriented I_{prc}.

The final image $\hat{I}$ is obtained by sequentially blending the intermediate result I_S using the generated masks:

$$I_{temp} = M_{struct} \odot I_{struct} + (1 - M_{struct}) \odot I_S$$
$$\hat{I} = M_{prc} \odot I_{prc} + (1 - M_{prc}) \odot I_{temp}$$

This adaptive per-face strategy ensures that the final image recovers realistic facial details without compromising structural accuracy.

4.4. Adaptive Facial Restoration Framework

Our face enhancement algorithm evaluates and selects the optimal result among I_{struct}, I_{prc}, and the intermediate output up to Stage 2 by employing the FGResQ metric. However, these outputs are generated by models that are not primarily focused on face restoration. Therefore, as proposed in [29], incorporating a dedicated face restoration model (e.g., GFPGAN [19]) could provide superior facial quality. We do not integrate this into our main pipeline because GFPGAN employs the RetinaFace_ResNet50 model for face detection, which incurs a longer processing time than the MediaPipe detector we selected. Furthermore, the face restoration process itself requires additional inference time during image processing. Such computational overhead is unsuitable for the lightweight image post-processing system pursued in our study. Nevertheless, we conducted supplementary experiments to demonstrate the extensibility and general applicability of our framework.

5. Experiments

5.1. Experimental Setup

Datasets We evaluate our method on widely used benchmark datasets, including Set5 [30], Set14 [31], BSD100 [32], Urban100 [33], and Manga109 [34]. Additionally, we use two real-world datasets, RealSR [35] and DRealSR [36], to assess performance on authentic degradations. All experiments are conducted for a 4× upscaling factor.

Implementation Details Our framework is governed by a small set of parameters, prioritizing simplicity and robust generalization. For edge detection, we utilize a 3×3 Sobel filter followed by a 7×7 dilation kernel. These specific kernel sizes were determined by evaluating 49 candidate combinations—ranging from 3×3 to 15×15 with odd-numbered increments to identify the optimal balance between edge sensitivity and structural connectivity. For facial masking, we employ the MediaPipe "Full-range" model with a min detection confidence of 0.5. All parameters were set heuristically and kept identical across all datasets to demonstrate the plug-and-play capability of our pipeline. While dataset-specific fine-tuning could yield marginal performance gains, our fixed configuration underscores the universal applicability of the proposed solution.

Models We use pre-trained SwinIR models. The classical SR model (SwinIR-C) was trained on DIV2K and Flickr2K. The real-world SR model (SwinIR-RW) was trained on DIV2K, Flickr2K, and OST. We also utilized PFT-SR and TSD-SR [10] to evaluate the universal applicability of our framework across a diverse range of modern SR architectures.

Evaluation Metrics We use a comprehensive set of metrics to evaluate both fidelity and perceptual quality. For all metrics except FGResQ, we utilize the PyIQA library [37] for standardized evaluation.

- **Fidelity Metrics:** PSNR and SSIM (higher is better), which measure fidelity against a ground-truth image.
- **Perceptual Metrics:** NIQE [38] (lower is better) and FGResQ [28] (higher is better), which assess perceptual quality without a ground-truth, and LPIPS [39] (lower is better) to evaluate perceptual quality with a ground-truth.

Methods for Comparison

- **SwinIR-C, PFT-SR**: The baseline fidelity-oriented models.
- **SwinIR-RW, TSD-SR**: The baseline perception-oriented models.
- **Our1 (Edge Blend)**: The intermediate output after edge blending only.
- **Our2 (Final)**: The final output from our full pipeline, which includes all refinement stages (Stage 1: histogram matching, Stage 2: edge blending, Stage 3: face correction).

Table 1
Quantitative results utilizing SwinIR outputs. For each metric, the best and second-best scores are highlighted.

Metric	Benchmark	SwinIR-C	SwinIR-RW	Our1 (Edge)	Our2 (Final)
PSNR↑	Set5	31.4824	26.1221	28.6486	28.6419
	Set14	27.4848	24.0935	25.7713	26.0191
	B100	26.6148	23.8285	25.0752	25.2667
	Urban100	26.0991	21.4291	23.8290	24.1037
	Manga109	30.7627	24.3835	28.0251	28.1471
	RealSR	26.2299	24.5537	25.8860	26.3594
	DRealSR	29.2974	26.8728	28.4385	29.1680
SSIM↑	Set5	0.8755	0.7111	0.7632	0.7902
	Set14	0.7596	0.6378	0.6781	0.6883
	B100	0.7265	0.6147	0.6504	0.6525
	Urban100	0.8069	0.6562	0.7204	0.7137
	Manga109	0.9013	0.7709	0.8299	0.8293
	RealSR	0.7383	0.7308	0.7471	0.7438
	DRealSR	0.7897	0.7510	0.7758	0.7821
FGResQ↑	Set5	0.5351	0.5641	0.5627	0.5816
	Set14	0.6502	0.7210	0.7075	0.7128
	B100	0.5989	0.7027	0.6688	0.6732
	Urban100	0.8490	0.8673	0.8648	0.8637
	Manga109	0.8270	0.8406	0.8401	0.8416
	RealSR	0.3446	0.6381	0.5424	0.5342
	DRealSR	0.2326	0.5134	0.4131	0.4055
LPIPS↓	Set5	0.1663	0.1660	0.1397	0.1467
	Set14	0.2664	0.2267	0.2147	0.2165
	B100	0.3536	0.2597	0.2579	0.2592
	Urban100	0.1840	0.2014	0.1821	0.1820
	Manga109	0.0920	0.1454	0.1103	0.1105
	RealSR	0.3963	0.2614	0.2528	0.2544
	DRealSR	0.4390	0.2830	0.2764	0.2746
NIQE↓	Set5	7.1313	7.2162	6.2083	6.6400
	Set14	6.2128	4.7497	4.8210	4.9061
	B100	6.1092	4.1204	4.2027	4.2733
	Urban100	5.4528	4.3865	4.4106	4.5018
	Manga109	5.3244	4.4098	4.3348	4.4502
	RealSR	8.3344	5.7362	6.0240	6.0937
	DRealSR	10.4529	6.5676	6.7367	6.8129

Table 2
Quantitative results utilizing PFT and TSD-SR outputs. For each metric, the best and second-best scores are highlighted.

Metric	Benchmark	PFT-SR	TSD-SR	Our1 (Edge)	Our2 (Final)
PSNR↑	Set5	31.6767	24.3916	28.0182	27.2450
	Set14	27.6472	22.3697	24.7628	24.7531
	B100	26.6958	22.3693	24.2418	24.4179
	Urban100	26.8483	19.9651	23.2513	23.4226
	Manga109	31.3273	21.6307	26.3633	26.2804
	RealSR	26.2245	22.0823	24.4746	25.1717
	DRealSR	29.2615	24.8758	27.3529	28.1055
SSIM↑	Set5	0.8774	0.7162	0.7801	0.7532
	Set14	0.7632	0.5957	0.6491	0.6431
	B100	0.7298	0.5785	0.6257	0.6270
	Urban100	0.8232	0.6226	0.7027	0.6940
	Manga109	0.9051	0.7370	0.8101	0.8004
	RealSR	0.7376	0.6505	0.6915	0.6970
	DRealSR	0.7893	0.6660	0.7130	0.7297
FGResQ↑	Set5	0.5502	0.7521	0.7047	0.7200
	Set14	0.6535	0.8293	0.8040	0.7942
	B100	0.6058	0.8436	0.8113	0.7928
	Urban100	0.8468	0.8582	0.8613	0.8591
	Manga109	0.8255	0.8314	0.8421	0.8430
	RealSR	0.3471	0.7571	0.6750	0.6437
	DRealSR	0.2327	0.6713	0.5911	0.5682
LPIPS↓	Set5	0.1616	0.1439	0.1175	0.1249
	Set14	0.2604	0.1942	0.1729	0.1733
	B100	0.3476	0.2033	0.1939	0.1945
	Urban100	0.1677	0.1706	0.1482	0.1468
	Manga109	0.0872	0.1376	0.1025	0.1021
	RealSR	0.3972	0.2805	0.2527	0.2503
	DRealSR	0.4394	0.3115	0.2823	0.2717
NIQE↓	Set5	6.4653	5.1420	5.5910	5.3679
	Set14	6.0587	3.9963	3.9044	3.9721
	B100	5.9730	3.6788	3.6091	3.6349
	Urban100	5.2631	4.2533	4.1440	4.2573
	Manga109	4.9014	4.0514	3.9317	4.0644
	RealSR	8.2619	5.0851	5.1089	5.2986
	DRealSR	10.2892	5.7729	5.5731	5.6605

5.2. Quantitative Results

Tables 1 and 2 present the quantitative comparison across all datasets. The results show a clear trend. Fidelity-oriented models (SwinIR-C and PFT-SR) consistently achieve the highest PSNR/SSIM scores except for SwinIR-C on RealSR. In contrast, perception-oriented models (SwinIR-RW and TSD-SR) perform well on these two no-reference metrics: NIQE and FGResQ.

As shown in Tables 1 and 2, our proposed method, Our2 (Final), strikes an effective balance. It significantly improves PSNR and SSIM compared to the perception-oriented models, recovering much of the lost fidelity. For instance, on Urban100 in Table 1, our method improves PSNR from 21.43 (SwinIR-RW) to 24.10 (Our2). Furthermore, in RealSR, our method based on the two SwinIR models achieves the best PSNR score.

Rather than aggressively maximizing a single objective, our framework aims to maintain a robust balance between visual fidelity and perceptual realism. While it may not always outperform specialized baselines in their primary metrics, it effectively mitigates extreme structural distortions and textural blurring. This balanced performance is particularly evident. Our ensemble outputs (Our1 and Our2) consistently achieve the best or second-best scores across most datasets and metrics.

A significant practical advantage of our approach is its computational efficiency. Our framework is implemented as a post-processing step, entirely separate from model training, and requires no additional fine-tuning. As detailed in Table 3, the entire process takes only 14.137 seconds for all images (100 images), highlighting its practical feasibility.

Table 3
Total post-processing time (ms) for all images in the Urban100 dataset. For this measurement, we used images generated by SwinIR. The total size of the loaded images was about 250 MB, and the saved images were about 130 MB. This measurement was performed on a system equipped with an Intel Core i7-14700K CPU, 32GiB DDR5 RAM, and an NVIDIA GeForce RTX 4080 GPU.

Component	Time (ms)	Percentage
FGResQ Loading	2043	14.45%
MediaPipe Loading	2	0.01%
Image Loading	2117	14.97%
Histogram Matching	2928	20.71%
Edge Blending	4084	28.89%
Face Correction	892	6.32%
Image Saving	2071	14.65%
Total	14137	100%

5.3. Qualitative Analysis

Effect of Histogram Matching Figure 3 demonstrates the effectiveness of our color stabilization stage. In the top row using SwinIR, the image shows how the bright, unnatural glow of the clock generated by SwinIR-RW is successfully suppressed via histogram matching with SwinIR-C. Similarly, in the bottom image using PFT-SR and TSD-SR, our

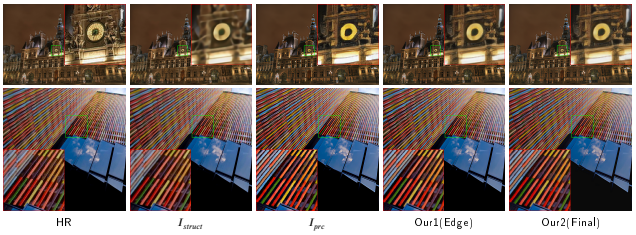

Figure 3: Effect of histogram matching. The I_{prc} outputs present color distortions and unrealistic color expressions. Top (SwinIR), Bottom (PFT-SR & TSD-SR): Our2 accurately suppresses the excessive artifacts introduced by SwinIR-RW, restoring the natural glow of the clock (Top) and returning overly dark, saturated colors to their original state (Bottom). (Top: Urban100 img022, Bottom: Urban100 img023).

approach pulls back overly saturated colors introduced by the perceptual model, successfully restoring the original natural colors of the scene.

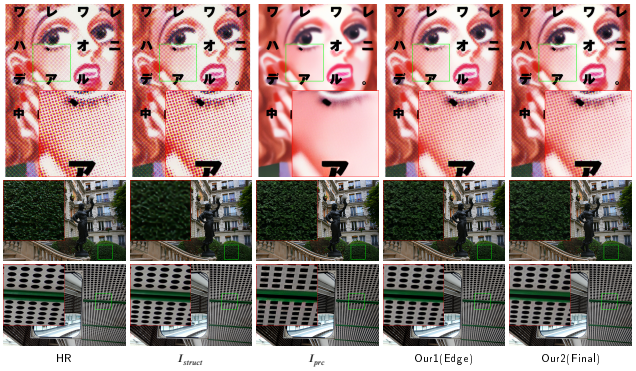

Figure 4: Effect of edge-aware blending. By blending with I_{struct} at structural boundaries, our method preserves fine shapes and textures without being completely dominated by perceptual hallucinations. Top (SwinIR): Halftone dots erased from the original manga by SwinIR-RW are restored. Middle (PFT-SR & TSD-SR): Grass texture generated well by I_{prc} is preserved and seamlessly blended. Bottom (SwinIR): Structural circular holes are distinctly preserved using the geometry of I_{struct}. (Top: Manga109 WarewareHaOniDearu, Middle: Urban100 img003, Bottom: Urban100 img004).

Effect of Edge-Aware Blending Figure 4 shows the benefit of our edge-aware blending strategy, which leverages the structural reliability of I_{struct} to anchor intricate shapes. The top image shows that SwinIR-RW entirely erases the specific halftone dots present in the manga image; however, our framework flawlessly restores these repeating patterns. In the middle image, utilizing PFT-SR and TSD-SR, the complex organic textures of the grass well-synthesized by I_{prc} are carefully preserved without introducing geometric irregularities. The bottom row highlights structural accuracy, as our blending strategy prioritizes the clear circular holes robustly restored by SwinIR-C over the distorted shapes presented by the real-world model.

Effect of Face-Aware Correction Figure 5 highlights the critical role of our dedicated face correction stage within the main pipeline. By evaluating facial crops through the FGResQ metric, the framework autonomously selects the most visually pleasing features. In the top row, the system accurately adopts the high-quality intermediate face naturally produced by Our1. Conversely, in the bottom row, where the perception-oriented TSD-SR produces excessive noise, the dynamic framework intelligently defaults to the slightly blurrier but structurally clean representation of I_{struct}.

Figure 5: Effect of face-aware correction within the main pipeline using PFT-SR and TSD-SR. Top: Our face-aware correction evaluates the outputs and appropriately selects the well-restored features from Our1. Bottom: When perceptual outputs introduce strong facial noise, the dynamic framework uses the slightly blurred, artifact-free I_{struct} as a fallback. Both cases demonstrate a reliable selection mechanism guided by FGResQ. (Top: B100 189080, Bottom: Manga109 DualJustice).

Figure 6: Effect of Adaptive Facial Restoration Framework using GFPGAN. When integrated with specialized facial restoration networks, the quality of detected faces improves. While TSD-SR manages to recover partial facial semantics compared to PFT-SR, our supplemental GFPGAN integration delivers superior realism and facial acuity. (Top: B100 85048, Middle: B100 216081, Bottom: Urban100 img009).

5.4. Adaptive Facial Restoration Framework (GFPGAN)

As introduced previously, combining our base framework with a supplementary specialized model highlights its significant extensibility. Figure 6 presents the results of utilizing the GFPGAN configuration on detected facial regions. Although TSD-SR produces an adequate foundational representation compared to PFT-SR, the intricate alignment of facial semantics remains incomplete. Additionally, it occasionally selects a blurrier face. The integration of GFPGAN corrects misaligned facial contours, restoring high-quality, realistic faces that out-perform both base implementations, proving that our framework can easily integrate with domain-specific models for targeted operations.

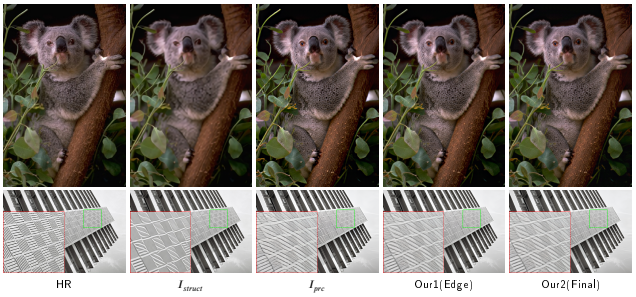

Figure 7: Failure cases. Top (PFT-SR & TSD-SR): The perceptually-oriented TSD-SR model hallucinates dotted artifacts on the koala's fur, which incorrectly bleed into the final result. Bottom (SwinIR): Structural misalignment artifact. The two models produce divergent restorations for the aliased diagonal line, resulting in a visible overlap in the blended output. (Top: BSD100 69015, Bottom: Urban100 img092).

5.5. Failure Cases

Despite its effectiveness, our method remains dependent on the baseline capabilities of the chosen SR models. Figure 7 illustrates typical failure cases. In the top image, the highly complex texture of koala fur leads TSD-SR to generate repetitive, dotted point artifacts. Because these artifacts exist as texture outside the structural boundary, our method unwittingly preserves them, leading to a degraded overall quality. The bottom image exhibits ghosting artifacts arising from structural misalignment, where the two base models produce divergent restorations for the aliased diagonal line. Consequently, attempting to blend these conflicting components exacerbates noise, rendering structural ghosts overlapping across the output.

These cases highlight avenues for future improvement. First, mitigating semantic-level artifacts like the unnatural dots generated by TSD-SR requires the development of dynamic, content-aware filtering beyond simple edge masking. Second, ghosting artifacts from structural inconsistencies heavily rely on the baseline robustness of the foundational models [40]. Finally, GFPGAN in our adaptive framework sometimes produces incomplete or unrealistic facial restoration results. Since it remains difficult to accurately determine which restoration is superior relying solely on current metrics, the development of more advanced no-reference metrics strongly aligned with human perception is essential.

6. Conclusion

In this work, we proposed a novel post-processing ensemble framework to address the persistent trade-off between fidelity and perception in SISR. By combining the outputs of fidelity- and perception-oriented SR models through a multi-stage process involving histogram matching, edge-aware blending, and face-aware correction, our method effectively mitigates common artifacts while producing images that are both structurally sound and texturally rich.

Our experiments confirm that the proposed method achieves a superior balance across both full-reference and no-reference quality metrics. The qualitative results further demonstrate its ability to correct specific color, structural, and facial distortions. A key advantage of our approach is its modularity and efficiency; as it requires no retraining, it can be readily applied to the outputs of various existing and future SR models, serving as a versatile tool for quality enhancement.

Future work could explore more sophisticated blending strategies, such as using semantic segmentation to create content-aware masks or developing a dynamic blending algorithm that adapts to the level of distortion. Addressing the limitations observed in our failure cases, particularly through more robust artifact detection, remains a promising direction for further research. Ultimately, our work represents a practical step toward generating high-quality super-resolved images that are suitable for real-world applications.

Appendix

The source code and online viewer are available at https://github.com/tama0728/ESR and https://tama0728.github.io/ESR_viewer, respectively.

Acknowledgements

We thank the reviewers for their constructive and insightful feedback. This research was funded by the 2024 and 2025 Research Grants from Sangmyung University (2024-A000-0231, 2025-A000-0247).

During the preparation of this work the authors used Google Gemini and ChatGPT in order to improve the readability and language of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] Nah, J.H., Kim, H., 2022. TexSR: Image super-resolution for high-quality texture mapping, in: SIGGRAPH Asia 2022 Posters. doi:10.1145/3550082.3564204.

[2] Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13, 600–612. doi:10.1109/TIP.2003.819861.

[3] Dong, C., Loy, C.C., He, K., Tang, X., 2016. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 295–307. doi:10.1109/TPAMI.2015.2439281.

[4] Kim, J., Lee, J.K., Lee, K.M., 2016. Accurate image super-resolution using very deep convolutional networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1646–1654. doi:10.1109/CVPR.2016.182.

[5] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., 2021. SwinIR: Image restoration using Swin Transformer, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 1833–1844. doi:10.1109/ICCVW54120.2021.00210.

[6] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 105–114. doi:10.1109/CVPR.2017.19.

[7] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C., 2019. ESRGAN: Enhanced super-resolution generative adversarial networks, in: Proceedings of the European Conference on Computer Vision Workshops (ECCVW), Springer. pp. 63–79. doi:10.1007/978-3-030-11021-5_5.

[8] Zhang, K., Liang, J., Van Gool, L., Timofte, R., 2021. Designing a practical degradation model for deep blind image super-resolution, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4771–4780. doi:10.1109/ICCV48922.2021.00475.

[9] Wang, X., Xie, L., Dong, C., Shan, Y., 2021. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 1905–1914. doi:10.1109/ICCVW54120.2021.00217.

[10] Dong, L., Fan, Q., Guo, Y., Wang, Z., Zhang, Q., Chen, J., Luo, Y., Zou, C., 2025. TSD-SR: One-step diffusion with target score distillation for real-world image super-resolution, in: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 23174–23184. doi:10.1109/CVPR52734.2025.02158.

[11] Xie, L., Wang, X., Chen, X., Li, G., Shan, Y., Zhou, J., Dong, C., 2023. DeSRA: detect and delete the artifacts of GAN-based real-world super-resolution models, in: Proceedings of the 40th International Conference on Machine Learning (ICML), JMLR.org. doi:10.5555/3618408.3619998.

[12] Kim, J., Lee, J.K., Lee, K.M., 2016. Deeply-recursive convolutional network for image super-resolution, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1637–1645. doi:10.1109/CVPR.2016.181.

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR). URL: https://openreview.net/forum?id=YicbFdNTTy.

[14] Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W., 2021. Pre-trained image processing Transformer, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12294–12305. doi:10.1109/CVPR46437.2021.01212.

[15] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022. doi:10.1109/ICCV48922.2021.00986.

[16] Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., Yu, F., 2023a. Dual aggregation Transformer for image super-resolution, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12278–12287. doi:10.1109/ICCV51070.2023.01131.

[17] Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C., 2023b. Activating more pixels in image super-resolution Transformer, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22367–22377. doi:10.1109/CVPR52729.2023.02142.

[18] Hsu, C.C., Lee, C.M., Chou, Y.S., 2024. DRCT: Saving image super-resolution away from information bottleneck, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 6133–6142. doi:10.1109/CVPRW63382.2024.00618.

[19] Long, W., Zhou, X., Zhang, L., Gu, S., 2025. Towards real-world blind face restoration with generative facial prior, in: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2279–2288. doi:10.1109/CVPR52734.2025.00218.

[20] Sun, L., Wu, R., Ma, Z., Liu, S., Yi, Q., Zhang, L., 2025. Pixel-level and semantic-level adjustable super-resolution: A dual-LoRA approach, in: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2333–2343. doi:10.1109/CVPR52734.2025.00223.

[21] Dietterich, T.G., 2000. Ensemble methods in machine learning, in: Multiple Classifier Systems, Springer. pp. 1–15. doi:10.1007/3-540-45014-9_1.

[22] Wang, L., Huang, Z., Gong, Y., Pan, C., 2017. Ensemble based deep networks for image super-resolution. Pattern Recognition 68, 191–198. doi:10.1016/j.patcog.2017.02.027.

[23] Jiang, J., Yu, Y., Wang, Z., Tang, S., Hu, R., Ma, J., 2020. Ensemble super-resolution with a reference dataset. IEEE Transactions on Cybernetics 50, 4694–4708. doi:10.1109/TCYB.2018.2890149.

[24] Long, W., Zhou, X., Zhang, L., Gu, S., 2025. Progressive focused Transformer for single image super-resolution, in: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2279–2288. doi:10.1109/CVPR52734.2025.00218.

[25] Liang, J., Zeng, H., Zhang, L., 2022. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5647–5656. doi:10.1109/CVPR52688.2022.00557.

[26] Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P., 2001. Color transfer between images. IEEE Computer Graphics and Applications 21, 34–41. doi:10.1109/38.946629.

[27] Google AI Edge, 2025. Face detection guide. URL: https://ai.google.dev/edge/mediapipe/solutions/vision/face_detector. mediaPipe Face Detector | Google AI Edge | Google AI for Developers.

[28] Sheng, X., Pan, X., Yang, Z., Chen, P., Li, L., 2026. Fine-grained image quality assessment for perceptual image restoration, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 8914–8922. doi:10.1609/aaai.v40i11.37846.

[29] Mun, J., Kim, J., 2020. Universal super-resolution for face and non-face regions via a facial feature network. Signal, Image and Video Processing 14, 1601–1608. doi:10.1007/s11760-020-01706-3.

[30] Bevilacqua, M., Roumy, A., Guillemot, C., line Alberi Morel, M., 2012. Low-complexity single-image super-resolution based on non-negative neighbor embedding, in: Proceedings of the British Machine Vision Conference (BMVC), BMVA Press. pp. 135.1–135.10. doi:10.5244/C.26.135.

[31] Zeyde, R., Elad, M., Protter, M., 2010. On single image scale-up using sparse-representations, in: International conference on curves and surfaces, Springer. pp. 711–730. doi:10.1007/978-3-642-27413-8_47.

[32] Martin, D., Fowlkes, C., Tal, D., Malik, J., 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Proceedings Eighth IEEE International Conference on Computer Vision (ICCV), pp. 416–423 vol.2. doi:10.1109/ICCV.2001.937655.

[33] Huang, J.B., Singh, A., Ahuja, N., 2015. Single image super-resolution from transformed self-exemplars, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5197–5206. doi:10.1109/CVPR.2015.7299156.

[34] Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K., 2017. Sketch-based manga retrieval using Manga109 dataset. Multimedia Tools and Applications 76, 21811–21838. doi:10.1007/s11042-016-4020-z.

[35] Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L., 2019. Toward real-world single image super-resolution: A new benchmark and a new model, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3086–3095. doi:10.1109/ICCV.2019.00318.

[36] Wei, P., Xie, Z., Lu, H., Zhan, Z., Ye, Q., Zuo, W., Lin, L., 2020. Component divide-and-conquer for real-world image super-resolution, in: European conference on computer vision (ECCV), Springer. pp. 101–117. doi:10.1007/978-3-030-58598-3_7.

[37] Chen, C., Mo, J., 2022. IQA-PyTorch: PyTorch toolbox for image quality assessment. [Online]. Available: https://github.com/chaofengc/IQA-PyTorch.

[38] Mittal, A., Soundararajan, R., Bovik, A.C., 2013. Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters 20, 209–212. doi:10.1109/LSP.2012.2227726.

[39] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 586–595. doi:10.1109/CVPR.2018.00068.

[40] Yang, H., Liu, R., Zhou, X., Zheng, Y., Zhao, R., 2025. Expert-scoring guided global information interaction network for lightweight image super-resolution. Image and Vision Computing 161, 105642. doi:10.1016/j.imavis.2025.105642.