DEFERREF: SELECTIVE FALLBACK FOR RETRIEVAL-BASED MULTI-REFERENCE IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-image super-resolution (SISR) restores a high-resolution (HR) image from only a low-resolution (LR) input, making it robust but forcing the model to infer missing high-frequency details from learned priors. Reference-based super-resolution (RefSR) can reduce this ambiguity by transferring texture from external HR reference images. However, this advantage depends on the references being locally compatible with the query. In practical retrieval-based systems, references are often obtained automatically from an image gallery or the web; they may be semantically related but differ in viewpoint, geometry, texture scale, or degradation, and forced reference transfer can then degrade the final result. We formulate this as per-image output selection between a multi-reference super-resolved (SR) output produced from retrieved references and a no-reference SISR fallback, without access to the HR ground truth. We propose *DeferRef*, a lightweight selective fallback mechanism. DeferRef does not synthesize a new image; it selects between two already generated SR candidates by combining two cues: a pre-SR retrieval compatibility score computed from DINOv2 similarities between the query and the top-K retrieved references, and a post-SR no-reference quality difference computed from BRISQUE scores of the two candidate outputs. Thresholds are calibrated by five-fold cross-validation over image IDs, avoiding tuning on evaluation images. On the 80-image WR-SR benchmark with a decontaminated 6,597-image retrieval gallery, DeferRef selects the multi-reference output for 20.0% of images and falls back otherwise. It improves PSNR from 27.107 dB for always using MRefSR and 27.083 dB for always using SwinIR to 27.324 dB, with simultaneous gains in SSIM, LPIPS, and DISTS over both fixed systems. These results show that retrieved references can help but remain risky per image, and that calibrated fallback is an effective way to reduce reference-induced failures in retrieval-based RefSR.

1 INTRODUCTION

Image super-resolution (SR) reconstructs a high-resolution (HR) image from a low-resolution (LR) observation. Deep single-image super-resolution (SISR) has advanced substantially through convolutional, residual, attention-based, and Transformer architectures (Dong et al., 2016; Lim et al., 2017; Zhang et al., 2018b; Liang et al., 2021). However, SISR is constrained by the information in the LR input: when high-frequency content is lost, the model must infer it from learned priors rather than recover it from the observation.

Reference-based super-resolution (RefSR) reduces this ambiguity by transferring texture from one or more HR reference images (Zheng et al., 2018; Zhang et al., 2019; Yang et al., 2020; Jiang et al., 2021; Lu et al., 2021; Cao et al., 2022). Multi-reference SR further allows several candidate references, reflecting practical settings in which related images are available from photo collections, video frames, or web search. The LMR dataset and its MRefSR baseline show that multiple references can improve reconstruction when the reference set contains compatible content (Zhang et al., 2023).

The practical bottleneck is therefore not only how to fuse references, but also how to obtain reliable ones. Manual pairing is costly, while gallery or web retrieval offers a more scalable alternative; WR-SR provides a representative benchmark for this setting (Jiang et al., 2021). Yet semantic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

similarity does not guarantee local compatibility: retrieved images may differ in viewpoint, geometry, texture scale, or degradation. Indiscriminate transfer from such references can introduce incorrect structure and make RefSR underperform a no-reference SISR model.

This paper studies selective fallback in retrieval-based multi-reference SR. Given an LR input, a top-K retrieved reference set, a multi-reference SR output, and a no-reference SISR output, the system must decide which output to return. We separate reference reliability into two signals: pre-SR compatibility between the query and retrieved references, and post-SR quality evidence from the generated candidates. A retrieval compatibility score measures whether the retrieved references look compatible with the query, but it cannot see artifacts in the final SR image. A no-reference image quality assessment (NR-IQA) score can compare the produced SR candidates without HR ground truth at test time, but by itself it may prefer smoother outputs and miss useful reference texture. DeferRef therefore combines retrieval compatibility and NR-IQA evidence in a calibrated threshold rule.

Concretely, we retrieve top-5 references from a decontaminated 6,597-image gallery built from WR-SR references and standard SR image collections. The gallery includes LMR-2000, CUFED5, DIV2K, Flickr2K, BSD100, Urban100, Set14, and Set5 images (Zhang et al., 2023; 2019; Agustsson & Timofte, 2017; Lim et al., 2017; Martin et al., 2001; Huang et al., 2015; Zeyde et al., 2012; Bevilacqua et al., 2012). We compute retrieval compatibility by extracting DINOv2 ViT-S/14 descriptors with register tokens for the LR query and each gallery image, L2-normalizing the descriptors, and ranking references by cosine similarity (Oquab et al., 2024; Darcet et al., 2024). We remove gallery images that are near-identical target duplicates of the WR-SR target HR image, while retaining normal high-similarity and same-sample web references. The resulting retrieval scores define the compatibility function used by DeferRef. We then use BRISQUE to compare the two candidate SR outputs (Mittal et al., 2012). The compatibility and BRISQUE thresholds are chosen by five-fold cross-validation over image IDs: each fold fits thresholds on 64 images and evaluates them on the remaining 16 images.

The results show that retrieved references are valuable but risky. On WR-SR, always using MRefSR outperforms always using SwinIR on all four metrics, including a 0.024 dB advantage in PSNR. DeferRef selects MRefSR for 20.0% of images and falls back for the rest, improving PSNR to 27.324 dB while also improving SSIM, LPIPS, and DISTS over both fixed choices. A same-family control with MRefSR-SelfRef further shows that the gain is not merely a consequence of replacing MRefSR with a different SISR architecture: even a weaker MRefSR-family fallback can recover a small but measurable PSNR improvement when selected image by image.

Our contributions are:

1. We formulate reference risk in retrieval-based multi-reference SR as a per-image output selection problem between RefSR and no-reference SISR candidates.
2. We introduce DeferRef, a calibrated selector that combines pre-SR retrieval compatibility and post-SR no-reference quality evidence.
3. We evaluate DeferRef on WR-SR with a decontaminated 6,597-image retrieval gallery and show PSNR gains of 0.217 dB over always using MRefSR and 0.241 dB over always using SwinIR, together with consistent improvements in SSIM, LPIPS, and DISTS.

2 Related Work

Single-image super-resolution. SISR estimates an HR image from only one LR input. SRCNN showed that a compact convolutional neural network can learn the LR-to-HR mapping end to end (Dong et al., 2016). VDSR increased depth and predicted residual details (Kim et al., 2016). EDSR simplified residual blocks and enlarged model capacity (Lim et al., 2017). RCAN introduced residual-in-residual connections and channel attention to strengthen high-frequency modeling (Zhang et al., 2018b). SwinIR brought shifted-window Transformers into image restoration and remains a strong classical SR baseline (Liang et al., 2021). These models are robust because they do not depend on external images, but they can only infer missing detail from the LR image and learned priors.

Reference-based super-resolution. RefSR uses an additional HR image to provide texture missing from the LR query. CrossNet aligns references by cross-scale warping (Zheng et al., 2018). SRNTT

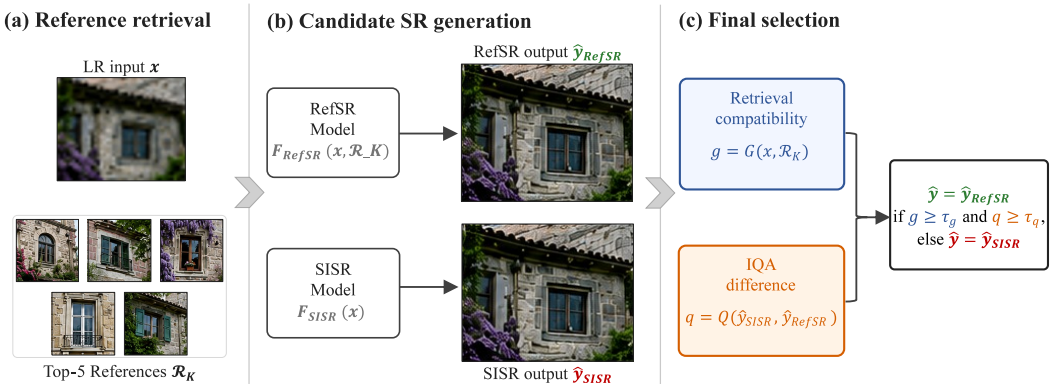

Figure 1: Overview of DeferRef. Given an LR input, the system retrieves top-K references and generates RefSR and SISR candidates. It returns the RefSR candidate only when both retrieval compatibility g and NR-IQA difference q exceed their calibrated thresholds; otherwise, it falls back to SISR.

transfers neural texture features from matched references (Zhang et al., 2019). TTSR searches and transfers reference texture with a texture Transformer (Yang et al., 2020). SADC and MASA-SR improve matching robustness and spatial adaptation (Shim et al., 2020; Lu et al., 2021). C2-Matching learns correspondences across transformation and resolution gaps (Jiang et al., 2021), and DATSR aggregates reference features with deformable attention (Cao et al., 2022). These methods differ in matching, alignment, and fusion design, but they generally assume that the provided reference contains useful information for the target image.

Multi-reference and retrieved-reference SR. Multi-reference SR is closer to how references are obtained in practical systems. LMR provides a large-scale multi-reference dataset where each sample group contains five reference images, and MRefSR aggregates an arbitrary number of references through multi-reference attention (Zhang et al., 2023). WR-SR addresses a related setting in which the reference image is collected through web search and then manually processed for evaluation (Jiang et al., 2021). Recent retrieval-augmented RefSR systems automatically retrieve HR references from external databases and condition SR generators on them (Lee et al., 2025; Yan et al., 2026). These settings make reference quality an explicit variable. A retrieved set can include useful texture, but it can also include images that are only semantically similar and locally mismatched.

Image quality assessment. Full-reference image quality assessment (FR-IQA) compares an output image with an HR ground truth image. PSNR and SSIM are standard SR metrics, while LPIPS and DISTS use learned or perceptual features to better reflect visual similarity in some cases (Wang et al., 2004; Zhang et al., 2018a; Ding et al., 2022). No-reference image quality assessment (NR-IQA) estimates image quality without an undistorted reference. BRISQUE is a classical NR-IQA method based on natural scene statistics in the spatial domain (Mittal et al., 2012). In SR benchmarks, FR-IQA is more direct for final evaluation because HR ground truth is available. At deployment time, however, HR ground truth is absent; NR-IQA can therefore serve as a practical decision signal for comparing candidate SR outputs produced from the same LR input.

3 Method

3.1 Overall Framework

DeferRef selects between two already generated SR candidates: a retrieved-reference MRefSR output and a no-reference SISR fallback. For each LR input x, a retrieval system returns a top-K reference set $\mathcal{R}_K = \{r_1, \ldots, r_K\}$. A multi-reference SR model produces

$$\hat{y}_{\text{RefSR}} = F_{\text{RefSR}}(x, \mathcal{R}_K), \tag{1}$$

and a no-reference SISR model produces

$$\hat{y}_{\text{SISR}} = F_{\text{SISR}}(x). \tag{2}$$

As shown in Figure 1, DeferRef does not synthesize a third image. It decides which of these two already generated candidates should be trusted.

The decision uses two signals. The retrieval compatibility function G maps the LR query and retrieved set to a scalar

$$g = G(x, \mathcal{R}_K), \tag{3}$$

which estimates whether the retrieved references are likely to be useful before SR is performed. The image-quality comparison function Q maps the two SR candidates to

$$q = Q(\hat{y}_{\mathrm{SISR}}, \hat{y}_{\mathrm{RefSR}}), \tag{4}$$

which estimates whether the RefSR output looks better than the SISR fallback after SR is performed. The final rule is

$$\hat{y} = \begin{cases} \hat{y}_{\mathrm{RefSR}}, & \text{if } g \geq \tau_g \text{ and } q \geq \tau_q, \\ \hat{y}_{\mathrm{SISR}}, & \text{otherwise.} \end{cases} \tag{5}$$

The thresholds τ_g and τ_q are fitted on calibration folds and then kept fixed on the corresponding evaluation folds.

3.2 Top-K Reference Retrieval

We construct a retrieval gallery from real HR images and search it for each WR-SR query. We encode the LR query and all gallery images with DINOv2 ViT-S/14 with register tokens (Oquab et al., 2024; Darcet et al., 2024). Specifically, we extract the class-token descriptor $f_{\mathrm{cls}}(z)$ for each image z and L2-normalize it:

$$\phi(z) = \frac{f_{\mathrm{cls}}(z)}{\|f_{\mathrm{cls}}(z)\|_2}. \tag{6}$$

The retrieval score between query x and a gallery image r is the cosine similarity between their normalized descriptors:

$$s(x, r) = \phi(x)^\top \phi(r). \tag{7}$$

The top-K references are the K gallery images with the largest scores. In our experiments, $K = 5$, and the sorted scores are denoted $s_1 \geq s_2 \geq \cdots \geq s_5$.

Because MRefSR consumes a reference set, we consider five summaries of the sorted top-5 similarities $s_1 \geq \cdots \geq s_5$:

$$G_1 = s_1, \quad G_2 = \frac{1}{5}\sum_{k=1}^{5} s_k, \quad G_3 = s_5, \quad G_4 = s_1 - s_2, \quad G_5 = s_1 - s_5. \tag{8}$$

These quantities represent, respectively, the strongest match, mean set compatibility, weakest retained match, top-two separation, and top-to-bottom spread. During calibration, DeferRef selects one summary and its orientation so that larger values favor accepting the RefSR output. The resulting scalar is denoted $g = G(x, \mathcal{R}_K)$ and compared with τ_g.

3.3 No-Reference Quality Difference

Retrieval compatibility is computed before SR and cannot observe artifacts in the final images. DeferRef therefore also compares the two SR candidates with BRISQUE, a no-reference image quality assessment model based on natural scene statistics (Mittal et al., 2012). BRISQUE assigns lower scores to images that better match natural scene statistics. We define

$$q = Q(\hat{y}_{\mathrm{SISR}}, \hat{y}_{\mathrm{RefSR}}) = \mathrm{BRISQUE}(\hat{y}_{\mathrm{SISR}}) - \mathrm{BRISQUE}(\hat{y}_{\mathrm{RefSR}}). \tag{9}$$

Thus $q > 0$ means BRISQUE prefers the RefSR output, while $q < 0$ means it prefers the SISR fallback. We use the score difference rather than an absolute BRISQUE value because the decision is relative to the same LR input and the same pair of candidate outputs.

Algorithm 1 Calibration of the DeferRef policy

Require: Calibration set with LR inputs, top-K references and retrieval scores, and HR targets; models F_{RefSR} and F_{SISR}
Ensure: Calibrated policy $\pi^\star = (G^\star, \tau_g^\star, \tau_q^\star)$
1: **for all** calibration images $i \in \mathcal{C}$ **do**
2: Generate $\hat{y}_{\mathrm{RefSR}}^{(i)} = F_{\mathrm{RefSR}}(x^{(i)}, \mathcal{R}_K^{(i)})$ and $\hat{y}_{\mathrm{SISR}}^{(i)} = F_{\mathrm{SISR}}(x^{(i)})$
3: Compute $G_1^{(i)}, \ldots, G_5^{(i)}$ and $q^{(i)} = Q(\hat{y}_{\mathrm{SISR}}^{(i)}, \hat{y}_{\mathrm{RefSR}}^{(i)})$
4: Compute the full-reference metrics of both candidates against $y^{(i)}$
5: **end for**
6: Construct threshold grids from all distinct calibration values, including the two endpoint policies
7: Construct the policy set $\Pi = \{(G, \tau_g, \tau_q) : G \in \mathcal{G}, \tau_g \in \mathcal{T}_g(G), \tau_q \in \mathcal{T}_q\}$.
8: **for all** candidate policies $\pi \in \Pi$ **do**
9: Apply Eq. (11) to obtain $\{d_i^\pi\}_{i \in \mathcal{C}}$
10: Compute selected-output PSNR $U(\pi)$ using Eq. (12)
11: Compute severe-error rate $E(\pi)$ and MRefSR selection rate $A(\pi)$
12: **end for**
13: $\pi^\star \leftarrow \mathrm{lex\,argmax}_{\pi \in \Pi}(U(\pi), -E(\pi), -A(\pi))$
14: **return** $\pi^\star$

3.4 Selection Rule and Threshold Fitting

Let

$$\mathcal{G} = \{G_1, -G_1, \ldots, G_5, -G_5\} \tag{10}$$

denote the candidate retrieval summaries and their two orientations. The orientation absorbs the comparison direction, so a larger value of the selected function always provides stronger evidence for accepting the RefSR output. A DeferRef policy is a tuple $\pi = (G, \tau_g, \tau_q)$, where $G \in \mathcal{G}$, and makes the decision

$$\begin{aligned} d^\pi &= \mathrm{SelectRefSR}(x, \mathcal{R}_K, \hat{y}_{\mathrm{RefSR}}, \hat{y}_{\mathrm{SISR}}) \\ &= \mathbf{1}[G(x, \mathcal{R}_K) \geq \tau_g \ \wedge \ Q(\hat{y}_{\mathrm{SISR}}, \hat{y}_{\mathrm{RefSR}}) \geq \tau_q]\,. \end{aligned} \tag{11}$$

If $d^\pi = 1$, DeferRef returns $\hat{y}_{\mathrm{RefSR}}$; otherwise it returns $\hat{y}_{\mathrm{SISR}}$.

For each $G \in \mathcal{G}$, let $\mathcal{T}_g(G)$ contain all distinct values of G observed on the calibration fold, augmented with endpoints that realize the always-accept and always-defer policies. Let $\mathcal{T}_q$ be defined analogously for q, with 0 additionally included. For a policy π, let $d_i^\pi \in \{0, 1\}$ denote its decision on calibration image i. We evaluate the mean PSNR of the selected outputs as

$$U(\pi) = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \left[d_i^\pi P_i^{\mathrm{RefSR}} + (1 - d_i^\pi) P_i^{\mathrm{SISR}} \right], \tag{12}$$

where P_i^{RefSR} and P_i^{SISR} are computed against the HR ground truth. We select the policy with the largest $U(\pi)$. Exact ties are broken by lower severe-selection-error rate and then lower MRefSR selection rate, where a severe error means that the selected output is worse on at least three of PSNR, SSIM, LPIPS, and DISTS.

We evaluate calibration by five-fold cross-validation over the 80 WR-SR image IDs. In each run, 64 images are used to select the policy and the remaining 16 are evaluated with that policy fixed. Decisions from all five evaluation folds are pooled for reporting. HR images are used only for calibration-fold policy fitting and final metric computation; inference uses only retrieval and BRISQUE scores. Algorithm 1 summarizes this calibration procedure; inference applies the resulting fixed policy through Eq. (11).

3.5 Fallback Model Choices

The main fallback model is SwinIR-M $\times 4$ classical SR with public DF2K weights (Liang et al., 2021). SwinIR serves as the no-reference fallback because it provides a strong SR candidate independent of the retrieved set. The multi-reference model is the official MRefSR implementation under the LMR MSE setting (Zhang et al., 2023).

We additionally train MRefSR-SelfRef as a same-family fallback control. It retains the MRefSR architecture and training pipeline, but replaces each external HR reference with the bicubic-upsampled LR input, repeated across the five reference slots. This produces a fallback candidate that does not depend on retrieved external images while remaining in the MRefSR model family.

This control addresses the possibility that the main gain comes only from using SwinIR, whose architecture differs from MRefSR. If selective fallback still improves over always using MRefSR when the fallback is a weaker model from the same family, the result supports the reference-risk explanation: some automatically retrieved reference sets reduce SR quality, and per-image selection can avoid part of that degradation.

4 Experiments

4.1 Experimental Setup

Dataset and retrieved references. We evaluate on WR-SR, an 80-image web-retrieval RefSR benchmark (Jiang et al., 2021). Each WR-SR sample provides an HR target and a web-searched reference that was manually processed for RefSR evaluation. We generate the LR query by mod-cropping the HR target and bicubic-downsampling it by a factor of four. WR-SR provides a controlled retrieval-based setting: references originate from web search and are processed for RefSR evaluation, but they may still differ from the query in geometry, viewpoint, texture scale, or degradation.

For every query, we start from a 6,605-image gallery built from WR-SR references, LMR-2000, CUFED5, DIV2K train/validation, Flickr2K, BSD100, Urban100, Set14, and Set5. To prevent target leakage, we remove eight gallery images identified as near-identical to WR-SR HR targets by average hash, difference hash, and thumbnail MSE, leaving 6,597 images for retrieval. No reference is otherwise forced into or excluded from the selected set. All remaining images are ranked by DINOv2 cosine similarity, and the five highest-scoring images are passed to MRefSR. Across the 400 selected reference slots, 139 images come from WR-SR, 110 from Flickr2K, 71 from DIV2K, 26 from BSD100, 27 from CUFED5, 26 from LMR-2000, and 1 from Set14. The setup keeps WR-SR's web references available while also allowing the retrieval system to draw from a broader SR image pool.

SR models. The multi-reference model is the official MRefSR implementation under the LMR MSE setting (Zhang et al., 2023). The main fallback model is SwinIR-M $\times 4$ classical SR with public DF2K weights (Liang et al., 2021). For the same-family control, we use MRefSR-SelfRef, trained in the MRefSR codebase under the same training setup. It replaces the five external HR references with repeated bicubic-upsampled LR self-references.

Metrics and threshold selection. We report PSNR and SSIM, where higher is better, and LPIPS and DISTS, where lower is better (Wang et al., 2004; Zhang et al., 2018a; Ding et al., 2022). We also report the MRefSR selection rate. For diagnosis, we report a severe selection error rate: the fraction of images for which the selected output is worse than the unselected candidate on at least three of the four metrics. This measures pairwise disagreement with an oracle over the two available outputs. Severe selection error is pair-specific. Table 1 compares MRefSR with SwinIR, whereas Table 3 compares MRefSR with MRefSR-SelfRef; these rates should therefore be interpreted within each table rather than compared across tables. The oracle used in the tables is a diagnostic upper bound over the two already generated outputs under the same metric-voting rule. All threshold-based results use five image-ID folds. Each fold fits thresholds on 64 calibration images and evaluates the fixed thresholds on 16 evaluation images.

Implementation details. Both candidate pipelines follow the WR-SR $\times 4$ bicubic protocol and derive their LR observations from the same HR target. MRefSR uses its official test-time preprocessing, while the SwinIR query is generated once by mod-cropping the target and bicubic-downsampling it by four. PSNR, SSIM, LPIPS, and DISTS are evaluated on RGB images in the $[0, 1]$ range after cropping four pixels from each border; LPIPS uses the AlexNet backbone. For retrieval, DINOv2 directly encodes the LR image without pre-upsampling. Before MRefSR inference, the selected references are resized or cropped to dimensions divisible by four, with the longer side capped at 720 pixels. BRISQUE is computed on the RGB SR outputs. Image-ID folds use seed 20260620,

Table 1: Main results on WR-SR using a 6,597-image gallery obtained by removing eight near-identical target duplicates from the original 6,605 images. †Random selection uses DeferRef's 0.200 selection probability and is averaged over five seeds; its realized mean selection rate is 0.160. The two-candidate oracle selects MRefSR when it outperforms SwinIR on at least three of the four metrics; otherwise, it selects SwinIR. It is a diagnostic bound over the two generated candidates, and DeferRef is only 0.025 dB below it.

Method	MRefSR selection rate	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	Severe sel. err.↓
Always MRefSR	1.000	27.107	0.7902	0.2892	0.1645	0.738
Always SwinIR	0.000	27.083	0.7889	0.2945	0.1676	0.200
Random, same expected selection rate	0.200†	27.088	0.7897	0.2927	0.1669	0.285
DeferRef	**0.200**	**27.324**	**0.7966**	**0.2823**	**0.1600**	**0.088**
Oracle over two candidates	0.200	27.349	0.7972	0.2812	0.1590	0.000

Table 2: Ablation of DeferRef's two decision signals. Retrieval compatibility assesses the reference set before SR, whereas the BRISQUE difference compares the generated MRefSR and SwinIR outputs. Their combination gives the highest PSNR and the lowest severe selection error, showing that retrieval compatibility provides a modest complementary signal beyond output-quality comparison alone.

Selector	MRefSR selection rate	PSNR↑	Gain vs. MRefSR	LPIPS↓	DISTS↓	Severe sel. err.↓
Always MRefSR	1.000	27.107	0.000	0.2892	0.1645	0.738
Always SwinIR	0.000	27.083	-0.024	0.2945	0.1676	0.200
Retrieval compatibility only	0.363	27.204	+0.096	0.2892	0.1637	0.300
BRISQUE difference only	0.225	27.317	+0.210	0.2824	0.1600	0.100
Retrieval compatibility + BRISQUE	**0.200**	**27.324**	**+0.217**	**0.2823**	**0.1600**	**0.088**

and the matched-random baseline uses seeds 0–4. Each deterministic candidate-generation and evaluation pipeline was run once with fixed checkpoints; MRefSR-SelfRef was trained once, the selector was evaluated over five folds, and the matched-random baseline was averaged over five seeds. All experiments were implemented in PyTorch and conducted using four NVIDIA GeForce RTX 4090 GPUs.

4.2 Main Results

Table 1 reports the main result. MRefSR is the stronger fixed system, outperforming SwinIR on all four metrics and by 0.024 dB in PSNR. DeferRef selects MRefSR for 20.0% of images and falls back to SwinIR for the rest, improving PSNR from the better fixed value of 27.107 dB to 27.324 dB. The same decisions also improve SSIM, LPIPS, and DISTS. The severe selection error rate drops from 73.8% for always using MRefSR and 20.0% for always using SwinIR to 8.8%, showing that many severe per-image failures are avoided. The two-candidate oracle reaches 27.349 dB PSNR, leaving only a 0.025 dB gap under this diagnostic oracle. Across PSNR, SSIM, LPIPS, and DISTS, paired two-sided Wilcoxon tests remain significant against both fixed systems after Holm correction (all adjusted $p \leq 0.0082$).

4.3 Signal Ablation

Table 2 separates the two signals used by DeferRef. Retrieval compatibility alone improves PSNR over always using MRefSR, but it accepts too many MRefSR outputs and leaves many severe selection errors. BRISQUE alone reaches 27.317 dB with a 10.0% severe selection error rate. Combining retrieval compatibility with BRISQUE further improves PSNR to 27.324 dB and reduces the severe selection error rate to 8.8%, indicating that the pre-SR retrieval signal provides modest complementary evidence beyond output quality alone.

Figure 2 visualizes why a selective rule is useful. Many images have negative ΔPSNR, meaning that MRefSR is worse than SwinIR for that image despite remaining better on several aggregate

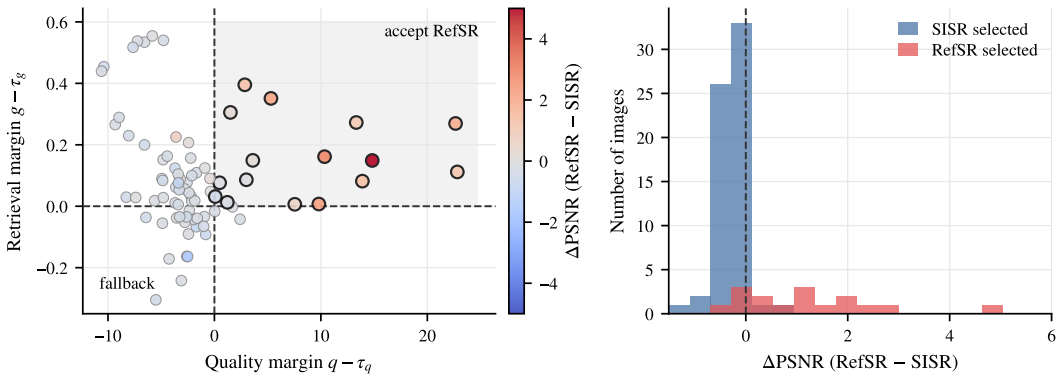

Figure 2: Decision diagnostics on WR-SR under the same five-fold policies used in Table 1. Left: each point represents one evaluation image, plotted by the BRISQUE margin $q - \tau_q$ and retrieval-compatibility margin $g - \tau_g$. The dashed lines mark the fold-specific thresholds, so the upper-right quadrant selects MRefSR. Color denotes $\Delta\text{PSNR} = \text{PSNR}(\text{MRefSR}) - \text{PSNR}(\text{SwinIR})$; blue points indicate images where MRefSR underperforms the SISR fallback. Right: the same ΔPSNR values grouped by DeferRef's selected output. Most of the negative tail is assigned to SwinIR.

Table 3: Same-family fallback control. MRefSR-SelfRef retains the MRefSR architecture and training procedure but replaces the external HR references with the bicubic-upsampled LR input repeated across all five reference slots. As a fixed system, it is 0.272 dB below MRefSR. After recalibration, DeferRef improves PSNR from 27.107 to 27.142 dB and numerically improves all four metrics; bootstrap gain intervals exclude zero for PSNR and SSIM, but not for LPIPS and DISTS. Severe selection error is defined within this candidate pair and is therefore not directly comparable with Table 1.

Method	MRefSR selection rate	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	Severe sel. err.↓
Always MRefSR	1.000	27.107	0.7902	0.2892	0.1645	0.538
Always MRefSR-SelfRef	0.000	26.835	0.7818	0.3027	0.1726	0.375
DeferRef with MRefSR-SelfRef fallback	**0.300**	**27.142**	**0.7913**	**0.2877**	**0.1637**	**0.163**

perceptual and structural metrics. DeferRef mostly keeps MRefSR only where both the retrieval margin and the BRISQUE margin support it, which explains the reduction in severe selection errors.

4.4 Same-Family Fallback Control

Table 3 examines whether the main improvement can be explained solely by the architectural difference between MRefSR and SwinIR. MRefSR-SelfRef is a deliberately weaker same-family control: when used for every image, it obtains 26.835 dB PSNR, 0.272 dB below MRefSR. After recalibrating DeferRef for this candidate pair, selective fallback reaches 27.142 dB, improving MRefSR by 0.035 dB and numerically improving SSIM, LPIPS, and DISTS; the bootstrap gain intervals exclude zero for PSNR and SSIM, but not for LPIPS and DISTS. The severe selection error rate decreases from 53.8% to 16.3% within this pair. Because the fallback retains the MRefSR architecture and training pipeline, this result indicates that the main effect is not solely caused by SwinIR's architecture: retrieved references reduce SR quality for a subset of images, and selective fallback can avoid some of these failures.

4.5 Qualitative Analysis

Figure 3 shows representative decisions. In fallback rows, the retrieved references do not translate into better local structure, and DeferRef selects SwinIR. In keep-reference rows, MRefSR can exploit the retrieved references to recover sharper object or facade details, and DeferRef keeps the multi-reference output.

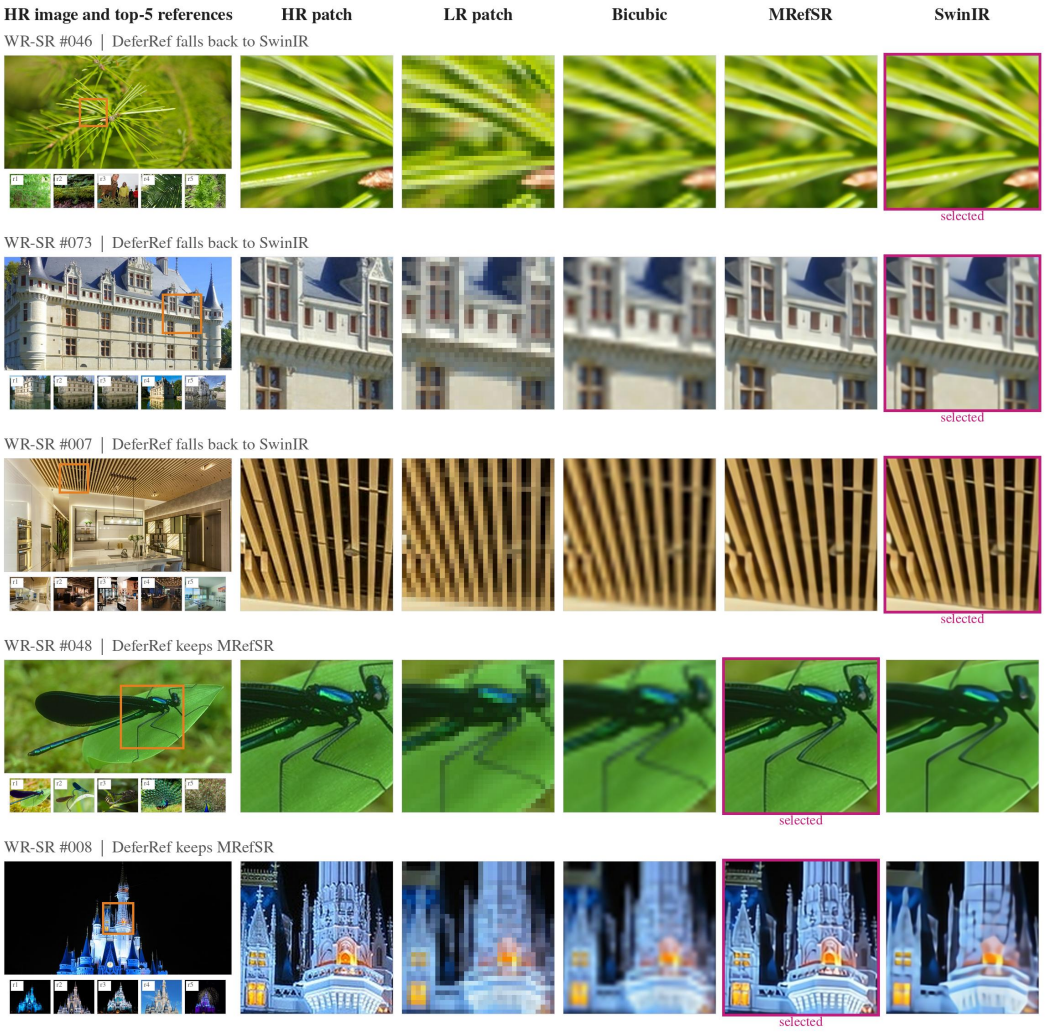

Figure 3: Representative DeferRef decisions on WR-SR. Each row shows the HR image and top-5 references, followed by HR/LR patches, bicubic, MRefSR, and SwinIR outputs; magenta marks the selected output. The first three rows show fallback cases, and the last two show cases in which retrieved references improve local detail.

5 Discussion and Conclusion

This paper studies selective reference use in retrieval-based multi-reference SR. Retrieved references help on average but can harm individual outputs. DeferRef combines retrieval compatibility and BRISQUE-based output comparison to choose between MRefSR and SISR. On WR-SR with a decontaminated 6,597-image gallery, it improves all four metrics over both fixed systems and approaches the two-candidate oracle, showing that reference reliability should be part of the SR decision rather than only retrieval preprocessing. The current selector is still post-hoc: because the BRISQUE difference is computed after SR, both the MRefSR and SISR candidates must be generated before DeferRef can make a decision. A useful future direction is to turn this post-hoc selector into a compute-aware policy: retrieval-only signals could first decide whether MRefSR should be run, while output-level quality signals could be reserved for ambiguous cases.

AI USE STATEMENT

We used AI-assisted coding tools to generate and refine experimental code through an iterative, prompt-based workflow. We also used generative AI tools to proofread the manuscript and improve its language, clarity, and readability. The authors take full responsibility for the final content of this work, including the code, experimental results, and written claims.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 specify the selector and its calibration procedure; Section 4.1 describes the data processing, evaluation protocol, and implementation settings. The anonymous supplementary code includes experiment scripts, environment snapshots, third-party revision identifiers, compatibility patches, a SelfRef training configuration, and numerical reference results with per-image decisions. Its README and dataset and checkpoint documentation explain the required external assets and identify artifacts not bundled with the code. CPU-only verification scripts support checks of the supplied numerical snapshots without rerunning SR inference; these checks are distinct from an end-to-end reproduction.

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 126–135, 2017.

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, pp. 135.1–135.10. BMVA Press, 2012. ISBN 1-901725-46-4. doi: http://dx.doi.org/10.5244/C.26.135.

Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *Computer Vision – ECCV 2022*, pp. 325–342. Springer, 2022.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024.

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, 2022. doi: 10.1109/TPAMI.2020.3045810.

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, 2016. doi: 10.1109/TPAMI.2015.2439281.

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5197–5206, 2015.

Yuming Jiang, Kelvin C. K. Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via C2-Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2103–2112, 2021.

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*, pp. 1646–1654, 2016.

Byeonghun Lee, Hyunmin Cho, Hong Gyu Choi, Soo Min Kang, Iljun Ahn, and Kyong Hwan Jin. Reference-based super-resolution via image-based retrieval-augmented generation diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10764–10774, 2025.

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 Workshops*, pp. 1833–1844, 2021.

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*, pp. 136–144, 2017.

Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. MASA-SR: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6368–6377, 2021.

David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 416–423, 2001.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Gyumin Shim, Jinsun Park, and In So Kweon. Robust reference-based super-resolution with similarity-aware deformable convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8425–8434, 2020.

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. doi: 10.1109/TIP.2003.819861.

Jiaqi Yan, Shuning Xu, Xiangyu Chen, Dell Zhang, Jiantao Zhou, Jie Tang, Gangshan Wu, and Jie Liu. RASR: Retrieval-augmented super resolution for practical reference-based image restoration. In *2026 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2228–2232, 2026. doi: 10.1109/ISCAS66217.2026.11562865.

Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5791–5800, 2020.

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, pp. 711–730. Springer, 2012. doi: 10.1007/978-3-642-27413-8_47.

Lin Zhang, Xin Li, Dongliang He, Fu Li, Errui Ding, and Zhaoxiang Zhang. LMR: A large-scale multi-reference dataset for reference-based super-resolution. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13072–13081, 2023. doi: 10.1109/ICCV51070.2023.01206.

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*, pp. 586–595, 2018a.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Computer Vision – ECCV 2018*, pp. 294–310. Springer, 2018b.

Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7982–7991, 2019.

Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. CrossNet: An end-to-end reference-based super resolution network using cross-scale warping. In *Computer Vision – ECCV 2018*, pp. 87–104. Springer, 2018.