Bridging the Reference Gap: Dual-Reference SR-IQA with Generated Image Guidance as Perceptual Anchors

Anonymous Author(s)

Abstract

Existing image quality assessment (IQA) methods fall short for super-resolution (SR): Full-Reference metrics are impractical without ground-truth references, while No-Reference metrics reflect image content rather than reconstruction quality. To address this challenge, we propose **DualRefIQA**, leveraging large generative models for dual-reference SR-IQA. The method uses two references: the original low-resolution image to preserve content fidelity, and a generated high-quality image (from large generative models) to guide perceptual quality assessment. To support this paradigm, we construct a new **D**ual Reference **SR** **Q**uality Assessment (**DSRQ**) dataset, a benchmark of over 170,000 triplets with human-annotated Mean Opinion Scores (MOS). Each triplet consists of an LR/HR pair and a generated reference that, by being augmented directly from the LR by large models, empirically approximates the HR distribution to facilitate content-quality disentanglement. Furthermore, we design a dual-branch architecture to fully exploit the information provided by both references: a Fidelity Branch that captures local structural fidelity against the aligned LR input, and a Perceptual Branch that leverages cross-attention to assess global texture quality using the generated reference. The two branches focus on disentangled aspects of content and quality, respectively, enabling more accurate evaluation. Extensive experiments show DualRefIQA achieves state-of-the-art correlation with human subjective scores, outperforming existing Full/No-Reference metrics.

CCS Concepts

• **Computing methodologies** → **Image manipulation**; **Neural networks**; *Transfer learning*.

Keywords

Image Quality Assessment, Image Super-Resolution, Generative Models, Dual-Reference Paradigm, Perceptual Naturalness

ACM Reference Format:

Anonymous Author(s). 2026. Bridging the Reference Gap: Dual-Reference SR-IQA with Generated Image Guidance as Perceptual Anchors. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 9 pages.

1 Introduction

Image Super-Resolution (SR) evolves rapidly, from handcrafted methods to CNN-based ones, and more recently, to generative models such as Generative Adversarial Networks (GANs) [10, 25] and Diffusion Models [17, 18]. These models can synthesize highly realistic textures for severely degraded low-resolution (LR) images.

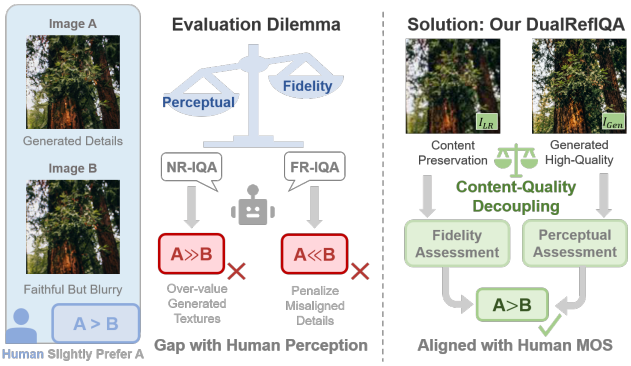


Figure 1: SR Evaluation Dilemma and Our Solution. *Left:* NR-IQA favors hallucinated textures while FR-IQA over-penalizes misalignment, yielding contradictory rankings against humans. *Right: DualRefIQA* decouples fidelity and quality via generative proxies and LR reference for MOS-consistent evaluation.

Existing Image Quality Assessment (IQA) methods for SR evaluation fall into three paradigms, each with critical limitations. Full-Reference (FR) metrics [5, 26, 37] measure pixel-level or feature-level distances to a ground truth (GT) reference, but in real-world SR, clean GT images are not available. When surrogates are used, FR metrics penalize perceptually plausible generative details that deviate structurally from reference, precisely the outputs that modern SR models are designed to produce. No-Reference (NR) metrics [15, 28, 35], on the other hand, evaluate images purely from a perceptual perspective. Without any fidelity anchor, they are blind to semantic distortions and actively reward hallucinated content, directly compromising the evaluation of SR fidelity. A third paradigm, Reduced-Reference (RR) methods [7, 16], attempts a middle ground with LR input as another reference. However, the LR image lies in a substantially different manifold from its HR one, making it an insufficient anchor for disentangling content from perception.

The rapid rise of the SR method based on generative models [13, 24, 29, 30, 33] has deepened these challenges. While such models achieve impressive perceptual realism, they often hallucinate structures inconsistent with the original scene, and their tendency to over-generate trades fidelity for appearance, making existing metrics even less reliable at jointly capturing both dimensions. **At a deeper level, the root cause lies in the representation space itself: content, aesthetics, and quality are difficult to cleanly separate in existing image embeddings.** Thus, any attempt to assess one dimension is inevitably influenced by the others, a fundamental entanglement that no purely metric-based approach has yet resolved. Generative models offer an opportunity: their ability to traverse and probe the high-dimensional image manifold provides a principled means to characterize the neighborhood structure around any given image. Inspired by this, we use generative priors to address the challenge of quality assessment for SR.

To address the semantic-quality entanglement described above, we propose **DualRefIQA**, a dual-reference IQA framework that tries to decouple content and perceptual quality. We use the LR input as a Content Anchor to assess structural fidelity, and a generative model to synthesize an SR image as a Perceptual Anchor to measure texture realism. Since the generated reference is not pixel-aligned with the SR output, we design a Spatial De-Align (SDA) module to inject texture statistics into the pipeline without spatial alignment constraints. An asymmetric dual-branch architecture then processes the two references according to their respective properties: a CNN branch handles the LR Content Anchor to capture local structural deviations, while a Cross-Attention branch handles the generated Perceptual Anchor to match texture distributions globally, remaining robust to spatial shifts.

Our main contributions are summarized as follows:

- We propose **DualRefIQA**, a novel dual-reference paradigm that takes the LR input as a Content Anchor for fidelity assessment and a generative model synthesized HR image as a Perceptual Anchor for texture quality, explicitly decoupling the two evaluation dimensions.
- We construct **DSRQ** dataset, a triplet benchmark of over 170,000 LR/SR/generated-reference samples with human-annotated MOS, where the generated references approximate the HR manifold to serve as Perceptual Anchors for content-quality disentanglement.
- We design an asymmetric dual-branch architecture with a Spatial De-Align (SDA) module, where a CNN branch assesses local structural fidelity against the LR Content Anchor, and a Cross-Attention branch evaluates perceptual quality against the generated Perceptual Anchor in a spatially invariant manner.
- DualRefIQA achieves state-of-the-art correlation with human perception on standard benchmarks, and serves as an effective training objective: fine-tuning SR networks via our metric improves the output's perceptual quality.

2 Related Work

1) General Image Quality Assessment. Traditional IQA is primarily divided into FR and NR metrics. Early FR-IQA metrics, such as PSNR and SSIM [26], evaluate image quality based on signal fidelity and structural similarity. With the advent of deep learning, feature-based FR metrics like LPIPS [37] and DISTS [5] have been proposed to better align with human perception by computing distance in deep feature spaces. However, FR metrics fundamentally rely on perfectly aligned high-resolution ground truths, making them overly sensitive to spatial misalignment and unjustly penalizing generated realistic details. In contrast, NR-IQA evaluates image quality without any reference. From traditional statistical models like BRISQUE [14] to deep architectures such as HyperIQA [20] and TOPIQ [3], NR metrics focus solely on perceptual aesthetics. Recently, Vision-Language Models (VLMs) have also been introduced to zero-shot NR-IQA, with methods like CLIPIQA [21], LIQE [38], and Q-Align [28] demonstrating strong generalization. Furthermore, hybrid adaptive metrics such as A-FINE attempt to dynamically merge naturalness and fidelity evaluation. Despite their flexibility, the lack of a reliable Content Anchor makes NR metrics blind to structural distortions, often rewarding severe semantic hallucinations. Unlike FR and NR metrics, DualRefIQA decouples structural content and perceptual quality via an LR content reference and a generated reference, requiring no pristine HR image.

2) Image Quality Assessment for Super-Resolution. Similarly, IQA methods for SR evaluation also fall into three paradigms. Full-Reference (FR) metrics [5, 26, 37] measure pixel- or feature-level distances to a GT reference, but clean GT images are rarely available in real-world SR. Even with surrogate references, FR metrics penalize perceptually plausible generative details that structurally deviate from the reference—precisely the outputs modern SR models aim to produce. No-Reference (NR) metrics [15, 28, 35] assess quality from a purely perceptual standpoint without any fidelity anchor, rendering them blind to semantic distortions and prone to rewarding hallucinated content. Reduced-Reference (RR) methods [7, 16] attempt a middle ground by leveraging the LR input as a partial reference, yet the LR image lies in a substantially different manifold from its HR counterpart, making it insufficient for disentangling fidelity from perception. These limitations motivate a new paradigm that jointly considers both aspects in a reference-aware manner. In contrast, we depart from this single-reference paradigm by introducing dual complementary references that ground fidelity assessment in the LR input while anchoring perceptual evaluation to a generative HR reference, offering a more complete and diagnostically meaningful evaluation of SR quality.

3) Generative Priors and Feature-Level Alignment. To address the lack of reliable high-resolution references, researchers have increasingly leveraged the robust representations embedded within pre-trained generative models. Works such as GFP-GAN [23] and CodeFormer [40] successfully utilize generative facial priors and codebook representations to guide the restoration of severely degraded images. These methods demonstrate that generative models can serve as highly effective texture surrogates, providing rich texture statistics when ground truths are absent. Furthermore, evaluating or utilizing unaligned reference features requires robust non-local matching strategies. Instead of relying on strict spatial convolutions, mechanisms such as Non-local Neural Networks [22] and Texture Transformers (*e.g.*, TTSR [31]) have proven highly effective in reference-based tasks. By employing cross-attention, these architectures can capture long-range dependencies and establish structural correspondence across images despite significant coordinate shifts. Inspired by these advancements, our DualRefIQA introduces a generative texture proxy and employs non-local attention to evaluate texture statistics without demanding rigid spatial alignment. Unlike prior works that apply generative priors and non-local matching to image restoration, DualRefIQA adapts these techniques to IQA, assessing texture naturalness via cross-attention without requiring spatial alignment.

3 Dual-Reference Image Quality Assessment

3.1 Motivation: Decoupling Content and Quality Assessment via Dual References

Generative model-based SR methods typically exhibit a fidelity-perception trade-off, where enhancing perceptual quality often leads to synthesized high-frequency details, sometimes at the cost of structural consistency. Traditional FR metrics tend to penalize

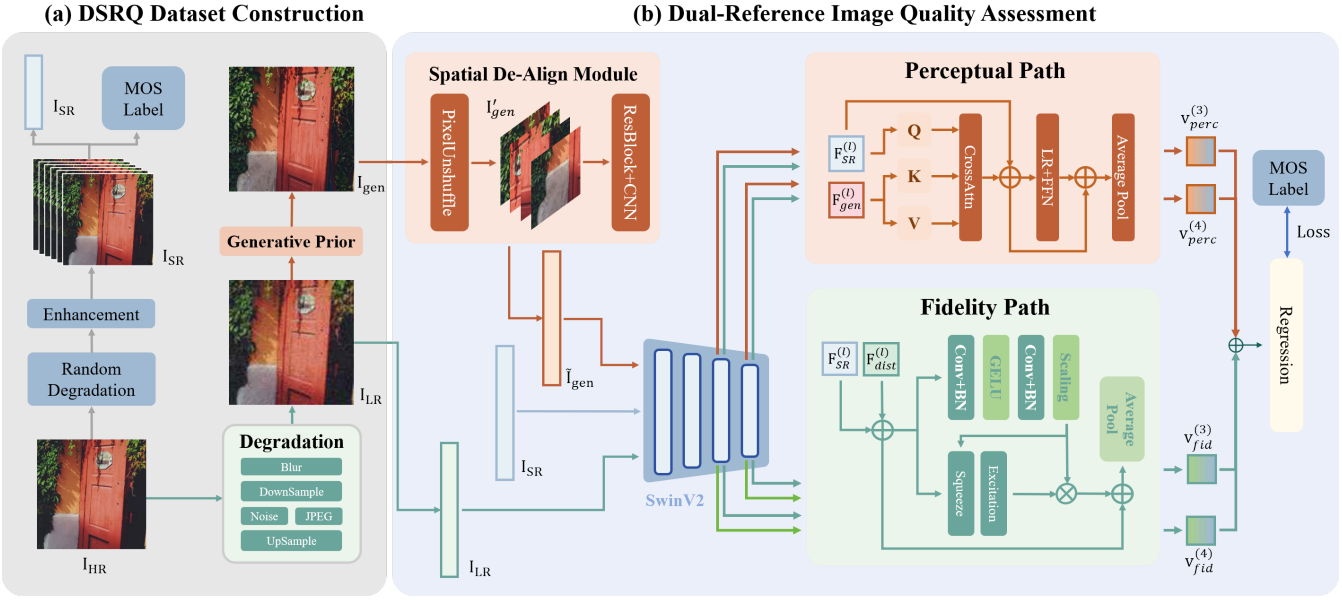


Figure 2: Overall architecture of DualRefIQA. Given an SR output, the framework extracts hierarchical features from an input triplet and evaluates quality via two branches: the Fidelity Path (*bottom*) measures structural fidelity against the LR reference, while the Perceptual Path (*top*) assesses texture quality against the generated HR reference. The fused representations are passed to a regression head to predict the MOS.

high-frequency details due to spatial misalignments, and thus fail to encourage the generation of perceptually realistic details, whereas NR metrics primarily focus on perceptual aesthetics and fail to trade-off between alignment with LR content and SR image quality.

To address these issues, we decouple the quality assessment into two dimensions: *Content* and *Quality*. To be specific, we leverage two complementary references to provide content guidance while promoting superior quality, such as richer texture that goes beyond ground-truth and better matches human perceptual preference:

- **Content Reference ($\mathbf{I}_{LR}$)**: The original low-resolution input. It preserves the low-frequency macro-structure of the scene, serving as a reliable anchor for content assessment.
- **Generated Reference ($\mathbf{I}_{gen}$)**: A high-quality image synthesized by advanced generative models, which denotes the visual quality reference. While we note that $\mathbf{I}_{gen}$ is not an absolute ground truth due to potential structural hallucinations, it serves as a high-quality perceptual proxy of the human-preferred texture quality. **Our goal is to evaluate how closely an SR image approaches the perceptual quality proxied by a generated SR reference, while respecting the structural constraints imposed by the content reference.**

Given a distorted SR image $\mathbf{I}_{SR}$ alongside two reference images, we extract hierarchical features using a pre-trained visual backbone Φ (*e.g.*, Swin Transformer). Crucially, instead of utilizing features from all hierarchical layers as seen in conventional perceptual metrics, we strictly extract features only from the deeper layers. The rationale behind this design is twofold: First, shallow layers primarily encode pixel-level variations and edges, making them hypersensitive to minor generative misalignments and unjustly penalizing perceptually reconstructions. Second, the deeper layers possess significantly larger receptive fields, effectively capturing robust high-level semantics, macro-structural contexts, and deep texture statistics. This deliberate layer selection aligns with our objective of evaluating structural fidelity and perceptual naturalness rather than pixel-wise differences. In practice, we select features from layers in the 3^{rd} and 4^{th} blocks in the Swin Transformer. For the l-th block ($l \in \mathcal{L} = \{3, 4\}$), we extract feature maps $\mathbf{F}^{(l)}_{SR}, \mathbf{F}^{(l)}_{LR}, \mathbf{F}^{(l)}_{gen} \in \mathbb{R}^{h_l \times w_l \times c_l}$ from $\mathbf{I}_{SR}$, $\mathbf{I}_{LR}$ and $\mathbf{I}_{gen}$, respectively.

3.2 Spatial De-Align (SDA) Module

We aim to encourage the generation of perceptually rich details that are more diverse and expressive than those available in the ground truth. To this end, We leverage large generative models to obtain a generated reference $\mathbf{I}_{gen}$ in a higher resolution (*e.g.*, $2H \times 2W$) with richer high-frequency details compared to the evaluation scale ($H \times W$). A naive bicubic downsampling would irreversibly blur and destroy these crucial texture statistics, compromising the purpose of using a high-quality generated reference.

To inject these high-frequency details losslessly into our evaluation pipeline, we design a dedicated Spatial De-Align (SDA) module $\mathcal{F}_{tex}$. Inspired by sub-pixel convolutional architectures, we employ a PixelUnshuffle operation with a downscale factor of 2. This operation spatially rearranges the high-resolution texture image $\mathbf{I}_{gen} \in \mathbb{R}^{2H \times 2W \times 3}$ into a lower-resolution but higher-dimensional tensor $\mathbf{I}'_{gen} \in \mathbb{R}^{H \times W \times 12}$ without discarding any pixel information:

$$\mathbf{I}'_{gen} = \text{PixelUnshuffle}_{2\times}(\mathbf{I}_{gen}). \tag{1}$$

Subsequently, we utilize a shallow fusion network $\mathcal{F}_{res}$ composed of residual convolutional blocks to compress the 12-channel tensor back to a 3-channel canonical representation while preserving the

essential high-frequency textures:

$$\tilde{\mathbf{I}}_{gen} = \mathcal{F}_{res}(\mathbf{I}'_{gen}) \in \mathbb{R}^{H\times W\times 3}. \tag{2}$$

This texture-preserved surrogate $\tilde{\mathbf{I}}_{gen}$ is then passed, alongside $\mathbf{I}_{SR}$ and $\mathbf{I}_{LR}$, into the shared frozen visual backbone Φ for multi-scale feature extraction.

3.3 Local Fidelity Assessment Based on Content Reference

The SR image $\mathbf{I}_{SR}$ and the content reference $\mathbf{I}_{LR}$ are approximately aligned in their low-frequency structures. Given that semantic distortion can be characterized by localized deviations, we leverage local convolutions to extract spatially aligned structural discrepancies for semantic fidelity assessment.

Rather than proposing a complex new structure, we adopt a standard residual convolution block equipped with a Squeeze-and-Excitation (SE) module to suit this local matching task. The features are concatenated and processed by the convolution module f_{conv}:

$$\mathbf{H}^{(l)} = f_{conv}\left(\mathbf{F}^{(l)}_{SR} \oplus \mathbf{F}^{(l)}_{LR}\right), \tag{3}$$

where $\oplus$ is element-wise addition, and H is the output feature. To adaptively re-weight feature channels based on structural distortion severity, the SE mechanism is applied:

$$\mathbf{w}^{(l)}_{se} = \sigma\left(\mathbf{W}_2\delta\left(\mathbf{W}_1\left(\frac{1}{h_l w_l}\sum_{i=1}^{h_l}\sum_{j=1}^{w_l}\mathbf{H}^{(l)}_{i,j}\right)\right)\right) \in \mathbb{R}^{c_l}, \tag{4}$$

where δ is the ReLU activation, σ is the Sigmoid function, and $\mathbf{W}_1, \mathbf{W}_2$ are linear transformations. The channel-wise re-weighted feature is globally pooled to form the fidelity descriptor:

$$\mathbf{v}^{(l)}_{fid} = \mathrm{GAP}\left(\mathbf{H}^{(l)} \odot \mathbf{w}^{(l)}_{se}\right) \in \mathbb{R}^{c_l}, \tag{5}$$

where GAP denotes Global Average Pooling and $\odot$ represents the Hadamard product.

3.4 Non-Local Perceptual Assessment via Generated Reference

While the content reference $\mathbf{I}_{LR}$ provides reliable structural and semantic information, the generated reference $\mathbf{I}_{gen}$ serves as a surrogate: while it provides high-quality texture statistics, the synthesized high-frequency details are inevitably spatially misaligned with $\mathbf{I}_{SR}$. Applying local convolutions here would penalize realistic textures simply due to coordinate shifts.

To address this, we employ a Cross-Attention block. Its non-local feature matching capability allows patches in $\mathbf{F}^{(l)}_{SR}$ to attend to visually similar texture cues across $\mathbf{F}^{(l)}_{gen}$, making the perceptual assessment robust to spatial misalignment. The spatial dimensions are flattened into sequences $\mathbf{X}^{(l)}_{SR}, \mathbf{X}^{(l)}_{gen} \in \mathbb{R}^{N_l\times c_l}$ ($N_l = h_l \times w_l$), which are linearly projected into Queries, Keys, and Values through projections $\mathbf{W}_Q$, $\mathbf{W}_K$ and $\mathbf{W}_V$:

$$\mathbf{Q} = \mathbf{X}^{(l)}_{SR}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}^{(l)}_{gen}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}^{(l)}_{gen}\mathbf{W}_V. \tag{6}$$

The non-local texture correlation is computed via scaled dot-product attention:

$$\mathbf{A}^{(l)} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \tag{7}$$

The aligned features $\mathbf{A}^{(l)}$ are refined through a residual connection and a Feed-Forward Network (FFN) with Layer Normalization (LN):

$$\mathbf{P}^{(l)} = \left(\mathbf{X}^{(l)}_{SR} + \mathbf{A}^{(l)}\right) + \mathrm{FFN}\Big(\mathrm{LN}(\mathbf{X}^{(l)}_{SR} + \mathbf{A}^{(l)})\Big). \tag{8}$$

The sequence is then spatially averaged to form the perceptual descriptor $\mathbf{v}^{(l)}_{perc} \in \mathbb{R}^{c_l}$.

3.5 Quality Regression and Optimization

Quality Regression. The overall quality of the SR image is determined by integrating structural and perceptual properties. We concatenate the descriptors from selected deep layers (denoted by the set $\mathcal{L}$) to construct a unified global representation:

$$\mathbf{v}_{global} = \bigoplus_{l\in\mathcal{L}}\left(\mathbf{v}^{(l)}_{fid} \oplus \mathbf{v}^{(l)}_{perc}\right). \tag{9}$$

A Multi-Layer Perceptron regression head $\mathcal{R}_\phi$ maps this representation to the scalar quality score:

$$\hat{s} = \mathcal{R}_\phi(\mathbf{v}_{global}). \tag{10}$$

Optimization. For the training of our IQA metric, we adopt the standard Mean Squared Error (MSE) loss to minimize the absolute discrepancy between the predicted scores $\hat{s}_b$ and the ground-truth Mean Opinion Scores s_b within a batch of size B:

$$\mathcal{L} = \frac{1}{B}\sum_{b=1}^{B}(\hat{s}_b - s_b)^2. \tag{11}$$

Although rank-based losses (*e.g.*, PLCC or SRCC loss) are commonly used to supervise IQA models, we find that the MSE objective suffices here. The explicit decoupling of content and generated references provides a strong inductive bias, guiding the metric to align with human perception and obtain state-of-the-art performance.

Ranking-based Constraint for Tuning SR Models. Directly maximizing the scalar score of DualRefIQA as a fine-tuning objective may cause the restoration network to over-optimize and produce adversarial artifacts. To address this, we introduce a ranking-based constraint that enforces the network to learn relative perceptual ordering rather than hacking an absolute score, preventing adversarial degradation. Specifically, given a generated SR image $\hat{I}_{SR}$ and its corresponding heavily degraded LR input $\mathbf{I}_{LR}$, the network should inherently ensure that the perceptual quality of the restored output surpasses that of the degraded input by a certain margin. To enforce this correct ordinal relationship, the ranking constraint is formulated as a margin-ranking loss:

$$\mathcal{L}_{rank} = \max\left(0, \mathcal{R}_\phi(\mathbf{v}_{LR}) - \mathcal{R}_\phi(\mathbf{v}_{SR}) + m\right), \tag{12}$$

where $\mathcal{R}_\phi(\mathbf{v}_{SR})$ and $\mathcal{R}_\phi(\mathbf{v}_{LR})$ denote the predicted DualRefIQA scores for the SR output and the corresponding LR input, respectively, and m is a predefined margin.

3.6 Construction of the Dual-Reference SR Quality (DSRQ) Dataset

Existing datasets lack the **dual-reference structure** required by our paradigm (*i.e.*, paired content references and generated references). Thus we construct the **Dual-Reference Super-Resolution Quality (DSRQ)** dataset. We follow the pipeline adopted by DiffIQA and extended via a three-step synthesis pipeline: **1) Content References:** Since DiffIQA only provides pristine HR images and

degraded SR outputs, we synthesize their corresponding LR counterparts to serve as content references. We design a blind degradation method, for each 512×512 pristine image, we sequentially apply: (i) Gaussian blur with a randomly selected kernel size from $\{7, 9, 11, 13, 15\}$ and $\sigma \in [0.2, 3.0]$; (ii) a 4× downsampling operation using a randomly chosen interpolation method (nearest, linear, cubic, or area); (iii) Additive White Gaussian Noise (AWGN) with a noise level uniformly sampled from $[0, 25]$; and (iv) JPEG compression with a Quality Factor ranging from $[30, 95]$. The degraded images are then upsampled back to 512×512 via bicubic interpolation, forming severely degraded LR inputs.

1) Generated References: To obtain pseudo-HR references that provide idealized high-frequency Percetpual Anchors, we use a powerful Vision-Language Generative Model (*i.e.*, `gemini-3.1-flash-image-preview`). We condition generation process purely on LR inputs, guided by a carefully crafted text prompt[41] to prevent semantic drift. An excerpt of the prompt is as follows: *"Please upscale this image to a high-quality 1024×1024 resolution. Perform super-resolution processing to significantly enhance clarity, remove pixelation, and sharpen textures, while strictly preserving the original content, composition, and colors unchanged. Do not alter the subject's features or hallucinate new elements."* This explicit constraint forces the generative model to produce highly realistic textures without modifying the underlying structures of an LR image.

2) Aggregation of Continuous MOS: To provide fine-grained supervision for the quality regression task, we aggregate the original discrete, categorical subjective annotations from DiffIQA. The raw dataset collects ratings from three independent annotators using a three-level scale (*Positive*, *Similar*, and *Negative*). We quantitatively map these categorical labels to numeric values (1.0, 0.5, and 0.0, respectively) and calculate their arithmetic mean. This transforms the discrete labels into a relatively continuous Mean Opinion Score (MOS) distributed within $[0, 1]$. Through this rigorous process, DSRQ comprises over 170,000 meticulously aligned image triplets, each containing a distorted SR output, a content reference, and a generated reference.

4 Experiments

4.1 Experimental Setup

Datasets. To comprehensively evaluate our proposed method, we conduct experiments on the **DSRQ** dataset and the **ISRGen-QA** [11] dataset. **1) DSRQ**. We evaluate on the DSRQ dataset (Section 3), split into train/val/test sets (8:1:1) at the pristine-image level to prevent data leakage. The dataset comprehensively supports three evaluation paradigms for a given distorted SR image (*dist*): **No-Reference** (using only *dist*), **Full-Reference** (using *dist* and pristine HR), and our **DualRefIQA** (using *dist*, LR, and generated reference), which enables rigorous assessment without relying on pristine HR counterparts. **2) ISRGen-QA**, which focuses on severe perceptual distortions and complex structural artifacts unique to modern generative models. To verify the robustness and adaptability of our feature-space proxy extraction to new generative distributions, we employ the ISRGen-QA dataset for a comprehensive evaluation. Following a similar paradigm, we synthesize the corresponding generated references for the ISRGen-QA samples.

To ensure a fair comparison, both our DualRefIQA and the data-driven baseline metrics (denoted with the -FT suffix) are explicitly fine-tuned on the ISRGen-QA training set before being quantitatively evaluated on the unseen testing split. This setup allows us to validate whether the metrics can effectively learn and adapt to novel artifacts.

Evaluation Criteria. To comprehensively quantify the performance of the Image Quality Assessment (IQA) models, we employ four standard correlation metrics widely adopted in the community. Specifically, we utilize Spearman's Rank Correlation Coefficient (SRCC) and Kendall's Rank Correlation Coefficient (KRCC) to measure monotonic rank correlation. Pearson's Linear Correlation Coefficient (PLCC) is used to evaluate linear correlation, while Root Mean Squared Error (RMSE) is adopted to quantify the absolute prediction error. For SRCC, PLCC, and KRCC, values closer to 1 indicate better performance, whereas a lower RMSE implies higher prediction accuracy. Prior to computing the PLCC and RMSE, we map the predicted scores to the same scale as the subjective Mean Opinion Scores (MOS) using a standard five-parameter non-linear logistic regression function.

Implementation Details. Our DualRefIQA model is implemented using the PyTorch framework. For feature extraction, we employ the pre-trained Swin-Transformer (specifically, the `swinv2_base_window8_256` variant) as the shared visual backbone. During both training and evaluation, the generated Super-Resolution images ($\mathbf{I}_{SR}$) and Content References ($\mathbf{I}_{LR}$) are resized to 512×512, while the Generated References ($\mathbf{I}_{gen}$) are processed at a higher resolution of 1024×1024 and downsampled losslessly via our SDA module. The model is optimized using the AdamW optimizer with a weight decay of 1×10^{-3} and an initial learning rate of 1×10^{-4}. We train the regression head for 100 epochs with a batch size of 4. The experiments are conducted on an NVIDIA A100 GPU, and the Mean Squared Error (MSE) is strictly adopted as the optimization objective without relying on complex rank-based losses.

4.2 Comparison with State-of-the-Art Methods

To demonstrate the superiority of our proposed DualRefIQA, we conduct comprehensive comparisons with several state-of-the-art (SOTA) IQA methods. We broadly categorize these baselines into two groups: **No-Reference (NR) Metrics** and **Full-Reference (FR) Metrics**. For fairness, models denoted with the **-FT** suffix indicate that they have been explicitly fine-tuned on the DSRQ dataset. Specifically, for the FR-FT metrics, the fine-tuning process utilizes the realistically unavailable ground-truth High-Resolution (HR) images as references.

Performance on DSRQ. As shown in Table 1, quantitative results on the DSRQ dataset validate the effectiveness of our approach. Among all NR metrics, our DualRefIQA achieves the highest correlation scores across all three indicators (*e.g.*, SRCC of **0.6913**), significantly outperforming state-of-the-art vision-language-based assessors like MANIQA-FT and TOPIQ_NR-FT, even when these baselines have been thoroughly fine-tuned on the target distribution. On the other hand, while the heavily optimized FR metrics (such as AFINE-FT and A-FINE) achieve higher correlation bounds, it is crucial to emphasize that they inherently depend on idealistic ground-truth HR images as references, which is unavailable in practical SR tasks. Remarkably, despite operating entirely without true

Table 1: Quantitative comparison of state-of-the-art IQA methods on DSRQ and ISRGen-QA datasets. For NR metrics, bold and underline indicate the best and second-best performance, respectively. Note that traditional FR metrics utilize inaccessible ground-truth HR images, which only serve as an upper bound for NR metrics. (↑: higher is better, ↓: lower is better).

Category	Method	DSRQ Dataset				ISRGen-QA			
		SRCC ↑	PLCC ↑	KRCC ↑	RMSE ↓	SRCC ↑	PLCC ↑	KRCC ↑	RMSE ↓
NR Metrics	NIQE [15]	0.1468	0.1544	0.1059	0.3615	0.2131	0.3978	0.1455	0.7384
	HyperIQA [20]	0.0666	0.1279	0.0467	0.3629	0.5016	0.4598	0.3799	1.9875
	MUSIQ [8]	0.1586	0.1978	0.1128	0.3587	0.2120	0.3190	0.1561	0.7628
	MANIQA [32]	0.0245	0.0549	0.0174	0.3653	0.5035	0.5294	0.3681	0.6828
	MANIQA-FT	0.6559	0.6518	0.4993	0.2726	0.5314	0.5597	0.3790	0.6631
	CLIPIQA [21]	0.2359	0.2559	0.1694	0.3537	0.2206	0.4040	0.1443	0.7362
	TOPIQ_NR [3]	0.0570	0.1362	0.0399	0.3625	0.5399	0.6386	0.4058	0.6193
	TOPIQ_NR-FT	0.6430	0.6386	0.4870	0.2766	0.7151	0.7537	0.5495	0.5289
	TReS [6]	0.1185	0.1757	0.0842	0.3602	0.4519	0.6162	0.3210	0.6339
	TReS-FT	0.3956	0.4321	0.2917	0.3242	0.3656	0.4564	0.2609	0.7161
	QualiCLIP [1]	0.0504	0.0814	0.0363	0.3647	0.1304	0.3292	0.1160	0.7600
	MACLIP [12]	0.1423	0.1841	0.1017	0.3596	0.2426	0.4517	0.1515	0.7181
	MACLIP-FT	0.2080	0.2297	0.1508	0.3499	0.0372	0.4606	0.0277	0.7144
	ARNIQA [2]	0.1900	0.2356	0.1377	0.3556	0.0225	0.2314	0.0171	0.7830
	ARNIQA-FT	0.6043	0.5998	0.4532	0.2876	0.5391	0.5321	0.3893	0.7428
	AFINE_NR	0.0387	0.0771	0.0276	0.3648	0.2625	0.2881	0.1820	0.7707
	Ours (DualRefIQA)	**0.6913**	**0.6922**	**0.5343**	**0.2716**	**0.7773**	**0.8324**	**0.6067**	**0.4461**
FR Metrics	PSNR	0.0083	0.0463	0.0069	0.3655	0.6854	0.7884	0.5354	0.4951
	MS-SSIM [27]	0.0859	0.1334	0.0586	0.3626	0.7167	0.8443	0.5672	0.4313
	SSIM [26]	0.1013	0.1409	0.0704	0.3622	0.7465	0.8634	0.5884	0.4061
	VIF [19]	0.0807	0.1323	0.0550	0.3628	0.7363	0.8211	0.5825	0.4594
	FSIM [36]	0.0465	0.1213	0.0304	0.3632	0.7042	0.8082	0.5471	0.4740
	VSI [34]	0.0454	0.1160	0.0299	0.3634	0.6948	0.7967	0.5448	0.4864
	LPIPS [37]	0.2247	0.2564	0.1592	0.3537	*0.8100*	*0.8840*	*0.6520*	*0.3762*
	DISTS [5]	0.3212	0.3442	0.2305	0.3435	*0.8386*	*0.8767*	*0.6861*	*0.3871*
	AHIQ [9]	0.0062	0.0023	0.0057	0.3659	0.2293	0.4337	0.1207	0.7252
	TOPIQ_FR	0.0397	0.1075	0.0251	0.3638	0.7356	0.7781	0.5672	0.5056
	AFINE_FR	0.0865	0.1055	0.0615	0.3638	0.7483	0.8129	0.5789	0.4687
	AFINE-FT	*0.7198*	*0.7175*	*0.5592*	*0.2549*	0.7596	0.8180	0.5931	0.4502
	A-FINE [4]	*0.7578*	*0.7552*	*0.5950*	*0.2398*	0.3195	0.5608	0.2197	0.6664
	Ours (DualRefIQA)	0.6913	0.6922	0.5343	0.2716	0.7773	0.8324	0.6067	0.4461

HR references (utilizing only our generated proxies), our method surpasses a vast majority of traditional and recently proposed FR metrics (such as LPIPS, DISTS, and A-FINE). This compellingly proves that our feature-space generated references successfully substitute ground-truth semantics to reliably guide the quality assessment of generative textures.

Adaptability on ISRGen-QA Dataset. To further demonstrate the robustness and adaptability of our proposed method to varied generative distributions and degradation types, we conduct experiments on the ISRGen-QA dataset. For a fair comparison, all data-driven baselines denoted with the -FT suffix, alongside our DualRefIQA, are explicitly fine-tuned on the ISRGen-QA training set. As shown in Table 1, DualRefIQA achieves a remarkable SRCC of 0.7773 and PLCC of 0.8324, establishing a new SOTA performance and significantly outperforming all competing NR metrics. While FR metrics like LPIPS and DISTS achieve marginally higher correlations (*e.g.*, SRCC of 0.8100) on ISRGen-QA, it is critical to note that they strictly require inaccessible ground-truth HR images. In contrast, our DualRefIQA operates entirely blindly without GT, yet achieves a highly competitive SRCC of 0.7773, significantly closing the gap between reference-free methods and idealistic Full-Reference upper bounds. Notably, even with sufficient fine-tuning, complex Vision-Language Models (*e.g.*, MANIQA-FT and MACLIP-FT) struggle to effectively capture the severe generative hallucinations specific to this dataset. These results compellingly validate that explicitly decoupling content and quality evaluation via dual-reference proxies provides a robust inductive bias, enabling our architecture to exceptionally adapt to artifacts introduced by SR methods based on generative models.

4.3 Ablation Studies

To deeply understand the contribution of each proposed component, we conduct extensive ablation studies on the DSRQ dataset. Specifically, we evaluate the impact of the dual-reference paradigm, the asymmetric network architecture, the Spatial De-Align (SDA) module, and the hierarchical feature selection strategy. The quantitative results are summarized in Table 2.

Effectiveness of the Dual-Reference Paradigm. We first decouple our framework to verify the necessity of integrating both



(a) HR (b) Generated Super-Resolution images from DiffIQA

Method	GT_MOS	PSNR	SSIM	LPIPS	HyperIQA	MANIQA	CLIPIQA	Ours
01	0.6667	23.6303	0.7093	0.2608	0.6534	0.6425	0.7330	0.3313
02	0.6667	24.8934	0.7263	0.2819	0.6603	0.6236	0.6936	0.3535
03	0.0000	28.6721	0.7401	0.3146	0.3640	0.2391	0.4630	0.3079
04	0.0000	26.8860	0.8122	0.1280	0.5263	0.3725	0.5485	0.3108
05	0.1667	25.1045	0.7455	0.2381	0.6450	0.6035	0.7369	0.3125
06	0.0000	31.3323	0.8523	0.2217	0.3750	0.3246	0.3677	0.3013

(c) Ground Truth MOS of SR images and their predicted quality scores

Figure 3: Visual comparison on DSRQ. Green, yellow, and red denote the best, second-best, and worst results. DualRefIQA produces scores more consistent with human MOS than existing methods.

low-level structure and high-level generated references. As shown in the first section of Table 2, utilizing only the LR content reference (*Content-only*) yields suboptimal correlations, as it ignores the synthesized naturalness of generative textures. Conversely, relying solely on the generated HR reference (*Quality*-only) leads to a severe lack of structural awareness, failing to penalize semantic drift. By incorporating both references into a cohesive feature-space proxy system, our **DualRefIQA** significantly boosts correlation metrics, confirming that SR evaluation with a balanced content–quality assessment aligns with human preferences.

Impact of SDA module. To evaluate the necessity of preserving high-frequency statistics from the generative texture proxy, we conduct an ablation by completely removing the SDA module (using standard bicubic downsampling to align resolutions). As shown in the second section of Table 2, the performance degrades dramatically across all correlation indicators when the SDA module is omitted. This substantial drop clearly demonstrates that standard downsampling operations irreversibly destroy critical high-frequency perceptual cues. By leveraging the PixelUnshuffle operation in our proposed SDA module, DualRefIQA successfully retains the rich texture statistics of the generated reference, establishing a robust perceptual upper bound without spatial misalignment penalties.

Analysis of Feature Selection Strategy. Lastly, we investigate the choice of abstraction levels extracted from the Swin Transformer backbone. Shallow layers (Stages 1 & 2) primarily encode low-level noise and edges, which are excessively rigid for generative outputs, while exclusively using the deepest layer (Stage 4) loses spatial structural details. We observe that, accumulating features from all four stages does not yield the best performance, likely due to feature redundancy and the math-washing effect of conflicting gradients. By selectively extracting intermediate-to-deep semantic features (**Stages 3 & 4**), DualRefIQA strikes the perfect balance between abstraction and structural preservation, delivering maximum accuracy with optimized computational cost.

Table 2: Ablation studies on the DSRQ dataset evaluating the effectiveness of the dual-reference paradigm, SDA module, and feature selection strategy. Bold indicates the default optimal configuration of our DualRefIQA (↑: higher is better, ↓: lower is better).

Configuration / Variant	SRCC ↑	PLCC ↑	KRCC ↑
1. Dual-Reference Paradigm (Model Variant)			
w/o LR Content Ref (*Texture-only*)	0.6555	0.6549	0.5237
w/o GenHR Texture Ref (*Fidelity-only*)	0.6481	0.6421	0.4988
DualRefIQA (Both Refs)	**0.6913**	**0.6922**	**0.5343**
2. Spatial De-Align (SDA) module			
w/o SDA module	0.5847	0.5752	0.3112
DualRefIQA (Both Refs)	**0.6913**	**0.6922**	**0.5343**
3. Feature Selection Strategy (Swin Transformer Layers)			
Shallow Layers Only (Stages 1 & 2)	0.5643	0.4792	0.3740
Deepest Layer Only (Stage 4)	0.5910	0.5556	0.4327
All Layers (Stages 1 to 4)	0.5641	0.4645	0.3652
Middle-Deep Layers (Stages 3 & 4)	**0.6913**	**0.6922**	**0.5343**

4.4 Visual Comparisons

To intuitively demonstrate the effectiveness of our proposed DualRefIQA, we provide a visual comparison of SR images alongside its predicted quality scores in Figure 3. Specifically, we select a challenging case containing severe generative hallucinations, misaligned structural artifacts, and unnatural textures. For each sample, we compare our predicted scores against representative NR and FR baselines, using the ground-truth Mean Opinion Score (MOS) as the authoritative reference.

As illustrated in Figure 3, traditional quality assessment metrics frequently misjudge the perceptual naturalness of generated SR outputs. In scenarios where generative models synthesize excessive or semantically incorrect hallucinated textures (*e.g.*, unrealistic structural distortions that deviate from the low-resolution content), baseline metrics often assign erroneously high scores, as they fail to decouple content fidelity from texture naturalness. In contrast, by leveraging the asymmetric alignment of our feature-space proxies, DualRefIQA accurately identifies these misaligned artifacts and penalizes the hallucinatory regions. Thus, our predicted scores maintain a robust and monotonic alignment with human MOS, confirming that our dual-reference paradigm reliably mimics human perceptual judgment in ill-posed generative evaluation tasks.

4.5 Application: DualRefIQA-Guided SR Fine-Tuning

A robust Image Quality Assessment (IQA) metric should not only accurately evaluate the perceptual quality of images but also serve as an effective optimization objective to guide the training of Image Restoration models. To demonstrate the practical utility of DualRefIQA, we integrate it into the training pipeline of a representative Super-Resolution network, RCAN [39], as a perceptual loss.

Experimental Setup. We adopt the standard 4× SR setting and fine-tune the models on the training split of the DSRQ dataset. The baseline models are optimized using the pixel-wise $\mathcal{L}_1$ (L1 loss) loss and a combination of $\mathcal{L}_1$ and $\mathcal{L}_{LPIPS}$ (LPIPS loss). In contrast, our proposed models are fine-tuned by incorporating our DualRefIQA

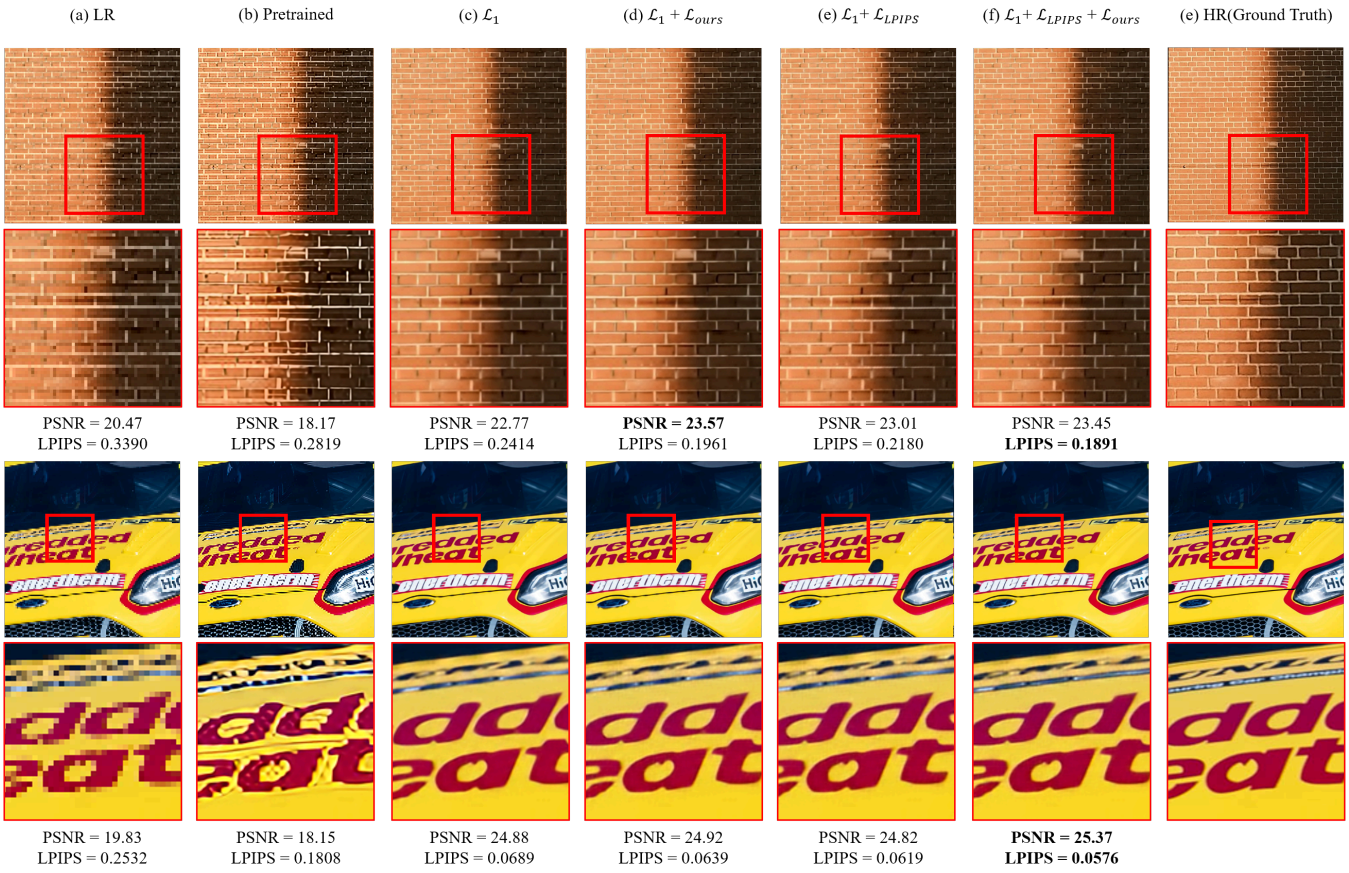


Figure 4: Visual comparison of the Baseline model trained with the combination of L1 loss($\mathcal{L}_1$), LPIPS loss($\mathcal{L}_{LPIPS}$) and Dual-RefIQA loss ($\mathcal{L}_{ours}$). The model optimized with DualRefIQA can reconstruct sharper and more perceptually realistic textures compared to the baseline.

loss ($\mathcal{L}_{ours}$) as an additional objective. The overall training objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{dist} + \lambda \cdot \mathcal{L}_{ours}, \tag{13}$$

where the distortion loss $\mathcal{L}_{dist}$ can be either $\mathcal{L}_1$ or $\mathcal{L}_1 + \mathcal{L}_{LPIPS}$, and λ is a balancing coefficient empirically set to 0.001. Note that during training, we also seamlessly introduce a ranking-based constraint into $\mathcal{L}_{ours}$ to better preserve the relative perceptual quality ordering.

Quantitative Results. The quantitative comparison on the DSRQ validation set is shown in Table 3. Baselines optimized exclusively with $\mathcal{L}_1$ or $\mathcal{L}_1 + \mathcal{L}_{LPIPS}$ struggle to balance structural fidelity and perceptual quality. In contrast, our DualRefIQA loss mitigates this by leveraging the generated reference as a Perceptual Anchor, guiding RCAN to synthesize realistic high-frequency details without severe structural distortions. Furthermore, to prevent adversarial artifacts caused by directly optimizing absolute IQA scores, we introduce a ranking-based constraint into $\mathcal{L}_{ours}$. This ensures the network learns the correct relative order of perceptual fidelity instead of hacking values.

Table 3: Quantitative comparison of SR fine-tuning w/ and w/o DualRefIQA loss on the DSRQ validation set (↑: higher is better, ↓: lower is better).

Method	Loss	PSNR↑	SSIM↑	L1↓	LPIPS↓	IQA↑
RCAN	*Zero-shot*	19.79	0.6304	0.0690	0.3335	0.2080
Baseline	$\mathcal{L}_1$	25.88	0.7742	0.0345	0.2558	0.2109
Ours	$\mathcal{L}_1 + \lambda\mathcal{L}_{ours}$	**26.01**	**0.7786**	**0.0339**	**0.2470**	**0.2393**
Baseline	$\mathcal{L}_1 + \mathcal{L}_{LPIPS}$	25.83	0.7725	0.0347	0.1932	0.2166
Ours	$\mathcal{L}_1 + \mathcal{L}_{LPIPS} + \lambda\mathcal{L}_{ours}$	**25.94**	**0.7755**	**0.0342**	**0.1812**	**0.2421**

5 Conclusion

We present DualRefIQA, a dual-reference IQA paradigm for generative model-based SR methods that decouples content and quality evaluation via an LR Content Anchor and a generated Perceptual Anchor. We design an asymmetric dual-branch architecture with local CNNs for fidelity and cross-attention for texture assessment, and construct DSRQ, a large-scale dataset with generated references and human MOS. DualRefIQA achieves state-of-the-art correlation with human perception and serves as an effective optimization objective for SR fine-tuning.

References

[1] Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini. 2024. Quality-Aware Image-Text Alignment for Opinion-Unaware Image Quality Assessment. *arXiv preprint arXiv:2403.11176* (2024).

[2] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. 2024. ARNIQA: Learning Distortion Manifold for Image Quality Assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 189–198.

[3] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2024. TOPIQ: A Top-down Approach From Semantics to Distortions for Image Quality Assessment. *IEEE Transactions on Image Processing* 33 (2024), 2404–2418.

[4] Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. 2025. Toward Generalized Image Quality Assessment: Relaxing the Perfect Reference Quality Assumption. *arXiv preprint arXiv:2503.11221* (2025).

[5] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. 2020. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 5 (2020), 2567–2581.

[6] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. 2022. No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3209–3218.

[7] S Alireza Golestaneh and Lina J Karam. 2016. Reduced-Reference Quality Assessment Based on the Entropy of DWT Coefficients of Locally Weighted Gradient Magnitudes. *IEEE Transactions on Image Processing* 25, 11, 5293–5303.

[8] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5148–5157.

[9] Shanshan Lao, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang. 2022. Attentions Help CNNs See Better: Attention-based Hybrid Image Quality Assessment Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1140–1149.

[10] Christian Ledig, Lucas Theis, Ferenc Huszár, et al. 2017. Photo-Realistic Single Image Super-resolution Using a Generative Adversarial Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 312–315.

[11] Yixiao Li, Xin Li, Chris Wei Zhou, Shuo Xing, Hadi Amirpour, Xiaoshuai Hao, Guanghui Yue, Baoquan Zhao, Weide Liu, Xiaoyuan Yang, et al. 2025. Vquala 2025 Challenge on Image Super-resolution Generated Content Quality Assessment: Methods and Results. *arXiv preprint arXiv:2509.06413* (2025).

[12] Zhicheng Liao, Dongxu Wu, Zhenshan Shi, Sijie Mai, Hanwei Zhu, Lingyu Zhu, Yuncheng Jiang, and Baoliang Chen. 2026. Beyond Cosine Similarity: Magnitude-Aware CLIP for No-Reference Image Quality Assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 6934–6942.

[13] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. 2024. DiffBIR: Toward Blind Image Restoration with Generative Diffusion Prior. In *Proceedings of the European Conference on Computer Vision*. 430–448.

[14] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing* 21, 12 (2012), 4695–4708.

[15] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters* 20, 3 (2013), 209–212.

[16] Abdul Rehman and Zhou Wang. 2012. Reduced-Reference Image Quality Assessment by Structural Similarity Estimation. *IEEE Transactions on Image Processing* 21, 8 (2012), 3378–3389.

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, et al. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10674–10685.

[18] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image Super-resolution via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4713–4726.

[19] H.R. Sheikh and A.C. Bovik. 2006. Image Information and Visual Quality. *IEEE Transactions on Image Processing* 15, 2 (2006), 430–444.

[20] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. 2020. Blindly Assess Image Quality in the Wild Guided by a Self-adaptive Hyper Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3667–3676.

[21] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023. Exploring CLIP for Assessing the Look and Feel of Images. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2555–2563.

[22] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7794–7803.

[23] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021. Towards Real-world Blind Face Restoration with Generative Facial Prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9168–9178.

[24] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 2021. Real-ESRGAN: Training Real-world Blind Super-resolution with Pure Synthetic Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1905–1914.

[25] Xintao Wang, Ke Yu, Shixiang Wu, et al. 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *Proceedings of the European Conference on Computer Vision Workshpos*. 63–79.

[26] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

[27] Z. Wang, E.P. Simoncelli, and A.C. Bovik. 2003. Multiscale Structural Similarity for Image Quality Assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, Vol. 2. 1398–1402.

[28] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. 2024. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. *arXiv preprint arXiv:2312.17090* (2024).

[29] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. 2024. One-step Effective Diffusion Network for Real-world Image Super-resolution. *Advances in Neural Information Processing Systems* (2024), 92529–92553.

[30] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. 2024. SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 25456–25465.

[31] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. 2020. Learning Texture Transformer Network for Image Super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5791–5800.

[32] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. 2022. MANIQA: Multi-dimension Attention Network for No-reference Image Quality Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1191–1200.

[33] Tao Yang, Rongyuan Wu, Shuai Li, Zhengqiang Zhang, and Lei Zhang. 2023. Pixel-Aware Stable Diffusion for Realistic Image Super-Resolution and Personalized Stylization. In *arXiv preprint arXiv:2308.14469*.

[34] Lin Zhang, Ying Shen, and Hongyu Li. 2014. VSI: A Visual Saliency-Induced Index for Perceptual Image Quality Assessment. *IEEE Transactions on Image Processing* 23, 10 (2014), 4270–4281.

[35] Lin Zhang, Lei Zhang, and Alan C Bovik. 2015. A Feature-Enriched Completely Blind Image Quality Evaluator. *IEEE Transactions on Image Processing* 24, 8 (2015), 2579–2591.

[36] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. 2011. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing* 20, 8 (2011), 2378–2386.

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 586–595.

[38] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. 2023. Blind Image Quality Assessment via Vision-language Correspondence: A Multitask Learning Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14071–14081.

[39] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image Super-resolution Using Very Deep Residual Channel Attention Networks. In *Proceedings of the European Conference on Computer Vision*. 294–310.

[40] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. 2022. Towards Robust Blind Face Restoration with Codebook Lookup Transformer. *Advances in Neural Information Processing Systems* (2022), 30599–30611.

[41] Jialong Zuo, Haoyou Deng, Hanyu Zhou, Jiaxin Zhu, Yicheng Zhang, Yiwei Zhang, Yongxin Yan, Kaixing Huang, Weisen Chen, Yongtai Deng, et al. 2025. Is Nano Banana Pro a Low-Level Vision All-Rounder? A Comprehensive Evaluation on 14 Tasks and 40 Datasets. *arXiv preprint arXiv:2512.15110* (2025).