

VQualA 2025 Challenge on Image Super-Resolution Generated Content Quality Assessment: Methods and Results

Yixiao Li*	Xin Li*	Chris Wei Zhou*	Shuo Xing	Hadi Amirpour
Xiaoshuai Hao	Guanghui Yue	Baoquan Zhao	Weide Liu	Xiaoyuan Yang
Zhenzhong Tu	Xinyu Li	Chuanbiao Song	Chenqi Zhang	Jun Lan
Huijia Zhu	Weiqiang Wang	Xiaoyan Sun	Shishun Tian	Dongyang Yan
Weixia Zhang	Junlin Chen	Wei Sun	Zhihua Wang	Zhuohang Shi
Zhizun Luo	Hang Ouyang	Tianxin Xiao	Fan Yang	Zhaowang Wu
		Kaixin Deng		

Abstract

This paper presents the ISRG-Q Challenge, built upon the Image Super-Resolution Generated Content Quality Assessment (ISRG-QA) dataset, and organized as part of the Visual Quality Assessment (VQualA) Competition at the ICCV 2025 Workshops. Unlike existing Super-Resolution Image Quality Assessment (SR-IQA) datasets, ISRG-QA places a greater emphasis on SR images generated by the latest generative approaches, including Generative Adversarial Networks (GANs) and diffusion models. The primary goal of this challenge is to analyze the unique artifacts introduced by modern super-resolution techniques and to evaluate their perceptual quality effectively. A total of 108 participants registered for the challenge, with 4 teams submitting valid solutions and fact sheets for the final testing phase. These submissions demonstrated state-of-the-art (SOTA) performance on the ISRG-QA dataset. The project is publicly available at: <https://github.com/Lighting-YXLI/ISRG-QA>.

1. Introduction

Super-resolution (SR) image quality assessment metrics aim to evaluate the perceptual quality of SR images from a human-centric perspective. This necessity arises from the

inherently ill-posed nature of the SR task, where a single low-resolution image may correspond to multiple plausible high-resolution reconstructions. As a result, SR faces a fundamental challenge in balancing fidelity (i.e., similarity to the ground truth) and naturalness (i.e., perceptual realism) [53, 55]. Moreover, SR images exhibit distinct distortion characteristics that differ significantly from those found in traditionally degraded images. Conventional distortions (e.g., blur, noise, and compression artifacts) [56] typically stem from information loss and lead to perceptual degradation. In contrast, SR methods often introduce enhancement-induced artifacts, including over-sharpened edges, hallucinated or false textures, and unnatural reconstruction patterns. Consequently, accurately assessing the perceptual quality of SR images is crucial [16, 54]—not only for evaluating but also for guiding the design and optimization of next-generation super-resolution algorithms.

To assess the SR-specific distortions, several SR-IQA datasets have been built, including the QADS [52], Waterloo [41], SISR-IQA [32], CVIU [23], RealSRQ [7], and SISAR [51]. The QADS [52] database contains 20 original HR references selected from the MDID database [28], and 980 SR images created by 21 SR algorithms, including 4 interpolation-based, 11 dictionary-based, and 6 DNN-based SR models, with upsampling factors equaling 2, 3, and 4. Each SR image is associated with the mean opinion score (MOS) collected from 100 individuals. In the CVIU [23] database, 1,620 SR images are produced by 9 SR approaches from 30 HR references. The HR reference images are selected from BSD200 according to the PSNR values. Six pairs of scaling factors (i.e., 2, 3, 4, 5, 6, 8) and kernel widths (i.e., 0.8, 1.0, 1.2, 1.6, 1.8, 2.0) are adopted, where a larger subsampling factor corresponds to a larger blur kernel width. Each image is rated by 50 subjects, and the mean of the median 40 scores is calculated for each image as the

*Yixiao Li (18335310648@163.com), Xin Li (xin.li@ustc.edu.cn), and Chris Wei Zhou (zhouw26@cardiff.ac.uk) are the challenge organizers. Shuo Xing, Hadi Amirpour, Xiaoshuai Hao, Guanghui Yue, Baoquan Zhao, Weide Liu, Xiaoyuan Yang, and Zhenzhong Tu are the technical supporters of the challenge. (Corresponding author: Chris Wei Zhou). The other authors are participants of the VQualA 2025 Challenge on Image Super-Resolution Generated Content Quality Assessment.

VQualA webpage: <https://vquala.github.io/>
The ISRG-QA dataset: <https://github.com/Lighting-YXLI/ISRG-QA>

MOS. The Waterloo [41] database involves 13 original HR references at 512×512 resolution, 39 low-resolution (LR) references, and 312 interpolated SR images generated by 8 interpolation algorithms, with upsampling factors of 2, 4, and 8. Subjective scores were collected from 30 participants aged 20–30 (17 males and 13 females). The SISR-IQA [32] database contains 15 ground-truth HR images selected from Set5, Set14, and BSD100, and 360 SR images reconstructed using 8 SR algorithms (e.g., DRCN [8] and VDSR [9]) with upsampling factors of 2, 3, and 4. LR images were generated via nearest neighbor interpolation. Subjective quality scores for all 360 SR images were collected from 16 participants, who were unaware of the HR references and SR methods. The RealSRQ [7] database contains 60 real-world HR references and 180 corresponding LR images at three scaling factors (2, 3, 4). A total of 1,620 SR images were generated using 10 SISR algorithms, including 5 non-deep methods and 5 deep models (SRCNN [5], CSCN [37], VDSR [9], SRGAN [10], and USRnet [43]). Subjective scores were collected from 60 participants (32 males and 28 females). The SISAR [51] database contains 12,600 SR images generated from 100 natural LR images. These images were processed using 10 SR algorithms or combinations, including 2 interpolation-based, 4 learning-based (SRCNN [5], VDSR [9], RCAN [50], SAN [3]), and 4 hybrid methods (e.g., SRCNN+BICUBIC). SR images were generated at 6 different scaling factors (i.e., 1.5, 2, 2.7, 3, 4, 3.6). Subjective scores were collected from 23 participants aged 20–30 with normal vision. Despite the progress in constructing SR-IQA databases, the super-resolution algorithms employed in these datasets have failed to keep pace with the rapid advancements in the field. For example, the most recent methods included are USRNet [43] (2020) and SAN [3] (2019), with SRGAN being the only generative adversarial network (GAN)-based approach represented. This lag significantly limits the effectiveness and generalizability of the resulting SR-IQA metrics in evaluating modern SR techniques.

With the rapid advances of the generative methods in SR tasks, recent SR models have increasingly incorporated generative priors, particularly through GAN- and diffusion-based models [11, 15, 25, 40, 46]. While these methods have shown promising results, striking an effective balance between perceptual realism and reconstruction fidelity remains challenging. GAN-based approaches often yield high-fidelity metrics but may fail to capture vivid textures due to their unstable adversarial training and tendency toward over-optimization [39]. While diffusion models can produce detailed textures by leveraging powerful generative priors, their reliance on stochastic noise sampling and mismatches between prior and LR distributions can undermine pixel-level accuracy [42]. Consequently, the construction of subjective quality assessment datasets specifically involv-

ing SR images produced by the latest generative models is of paramount importance for facilitating the further refinement and advancement of SR techniques.

In conjunction with the ICCV 2025 Workshop, we present the ISRG-Q Challenge on image super-resolution generated content quality assessment. The goal of this challenge is to automatically evaluate the perceptual quality of super-resolved (SR) images, ensuring that the predicted scores align as closely as possible with human visual perception. This challenge is one of the VQualA 2025 Workshop associated challenges on: FIQA: Face Image Quality Assessment Challenge [24], ISRG-Q: Image Super-Resolution Generated Content Quality Assessment Challenge [17], EVQA-SnapUGC: Engagement Prediction for Short Videos Challenge [12], Visual Quality Comparison for Large Multimodal Models [57], DIQA: Document Image Enhancement Quality Assessment Challenge [6], GenAI-Bench AIGC Video Quality Assessment [2]. In the following sections, we describe the challenge in detail, present and analyze the results, and provide an overview of the participating methods.

2. VQualA 2025 Challenge on ISRG-Q

The VQualA 2025 Challenge on Image Super-Resolution Generated Content Quality Assessment (ISRG-Q) is the first challenge to be organized to advance the development of assessing super-resolved images, especially those generated by the latest GAN- and diffusion-based approaches. The details of the whole challenge are as follows:

2.1. ISRGen-QA database

ISRGen-QA is a super-resolution (SR) image quality assessment database that contains sufficient SR images generated by the latest generative models, including GAN- and diffusion-based methods. It consists of 720 super-resolved images at approximately 2K resolution ($2040 \times 1152 \sim 2040 \times 1440$), covering four typical upscaling factors ($\times 2$, $\times 3$, $\times 4$, and $\times 8$). A total of 15 advanced SR algorithms are used to generate the images, including 4 GAN-based (i.e., ESRGAN [34], Real-ESRGAN [35], BSRGAN [44], and SeD [11]), 5 diffusion-based (i.e., SR3 [26], IDM [4], SRDiff [13], CDFFormer [20], and SAM-DiffSR [33]), 4 transformer-based (i.e., SRNO [38], ATD-SR [45], SwinIR [18], and CAMixerSR [36]), 1 flow-based method (i.e., BFSR [30]), and 1 CNN-based method (i.e., EDSR [19]). The SR images are derived from 19 high-resolution (HR) reference images and 76 low-resolution (LR) reference images created via four down-sampling scales ($\times 2$, $\times 3$, $\times 4$, and $\times 8$). The HR references are selected from the DIV2K [1], and the utilized down-sampling method is the Bicubic. To ensure the reliability of perceptual quality annotations, subjective scores were collected from 23 human participants (11 female, 12 male, from 5

different countries and various ages), with anomaly filtering yielding valid scores from 21 participants. The dataset is divided into training (576 images, 80%), validation (72 images, 10%), and test (72 images, 10%) sets, facilitating reproducible model development and benchmarking.

2.2. Evaluation Protocol

This challenge utilized two metrics to measure the correlation of the quality predictions and the mean opinion scores (MOS), including Spearman rank-order correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC). SRCC and PLCC are employed to assess the monotonicity and accuracy of predictions, respectively. An ideal quality metric would have SRCC and PLCC values close to one. The final score used for ranking is computed by reweighting the above metrics as :

$$\text{Score} = 0.6 \times \text{SRCC} + 0.4 \times \text{PLCC}. \quad (1)$$

2.3. Challenge Phases

There are two phases in this challenge, i.e., the development and testing phases. The details are as follows:

2.3.1. Development Phase:

In the development phase, we release 576 SR images and their corresponding high-resolution and low-resolution reference images in our ISRGen-QA dataset to support each team in developing their algorithms. Moreover, we release 72 SR images without their MOS scores for validation. Each participant can upload their quality predictions of the validation set to the challenge platform (Codalab: <https://codalab.lisn.upsaclay.fr/competitions/22924>). Then they can obtain the corresponding final score, SRCC, and PLCC. In the development phases, we received 193 submissions from 12 teams in total.

2.3.2. Testing Phase:

In the testing phases, we release 72 SR images and their corresponding high-resolution and low-resolution reference images in our ISRGen-QA dataset for testing. The final ranking is achieved with the score in Eq. 1. In the test stage, 5 teams submitted their final results to the challenge platform. At the end of this competition, we received the fact sheets and source codes from 4 teams, which are utilized for the final ranking.

3. Challenge Results

The main results from the 4 participating teams (Team MCV, Team ydy, Team QA-Veteran, and Team 2077 Agent) are summarized in Table 1, as well as the detailed information on their methods. Figures 1 and 2 present illustrations of the performance achieved by the submitted methods.

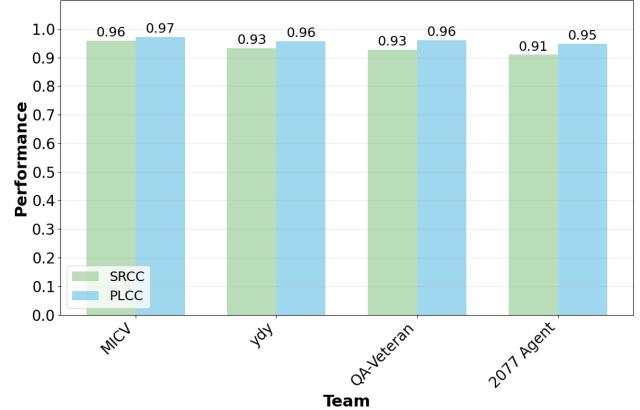


Figure 1. The performance of the methods submitted by different teams on the testing set.

3.1. Results Analysis

As presented in Table 1, all participating teams demonstrated exceptional performance, achieving overall scores exceeding 0.9, which indicates strong alignment with human perceptual quality judgments. A consistent pattern emerges across all teams: PLCC scores consistently exceed SRCC values (ranging from 0.9476 to 0.9714 for PLCC versus 0.9096 to 0.9588 for SRCC). This discrepancy reveals important insights about the methods' characteristics:

- **Linear relationship capture:** The higher PLCC scores indicate that all methods excel at capturing linear correlations between predicted and ground truth quality scores.
- **Rank-order consistency:** The relatively lower SRCC scores suggest some challenges in maintaining perfect monotonic rank consistency across the entire quality spectrum.
- **Practical implications:** While methods may occasionally misorder samples with similar quality levels, they maintain strong overall quality prediction accuracy.

Figure 1 shows the performance histogram comparing PLCC and SRCC scores across the 4 participating teams. The consistently high performance scores, tightly clustered between 0.91 and 0.97, highlight the effectiveness of the submitted super-resolution quality assessment methods.

Furthermore, Figure 2 shows the scatter plots of predicted scores versus the MOS for all 4 team methods on the testing set. The curves are obtained through fourth-order polynomial nonlinear fitting. We can observe that the predicted scores obtained by the top-performing team methods demonstrate higher correlations with the MOS values, as evidenced by the tighter clustering of data points around the fitted curves.

Table 1. Quantitative results from the VQualA 2025 Image Super-Resolution Generated Content Quality Assessment Challenge, including detailed information on the methods used by the 4 participating teams. The best performances are highlighted in bold. Note that **GFlops** are calculated relative to **Input Size**.

Rank	Team	Leader	Overall	SRCC \uparrow	PLCC \uparrow	Params. (M)	Input Size	GFlops (G)	Ensemble	Extra Data
1	MICV	Chuanbiao Song	0.9638	0.9588	0.9714	6	(448, 448, 3)	1000	\times	\times
2	ydy	Shishun Tian	0.9429	0.9333	0.9572	161	(128, 128, 3)	6	\times	\times
3	QA-Veteran	Weixia Zhang	0.9409	0.9277	0.9608	375.32	(1280, 1280, 3)	428.83	\times	\times
4	2077 Agent	Zhuohang Shi	0.9248	0.9096	0.9476	91.56	(2040, 1152, 3)	322.73	\times	\times

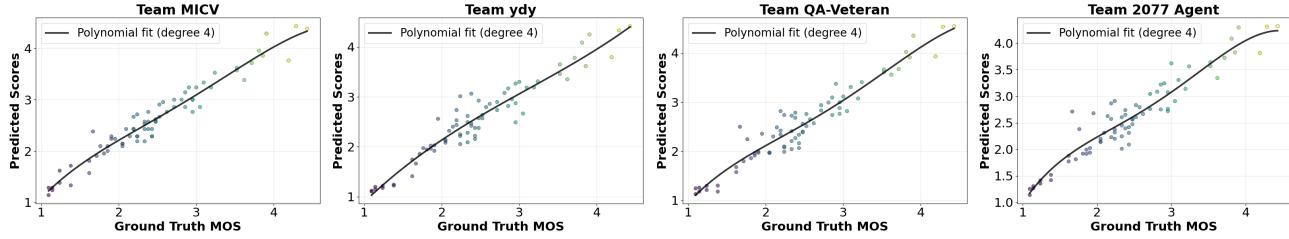


Figure 2. Scatter plots of predicted scores versus MOS for all participating teams on the testing set. The curves are obtained by a fourth-order polynomial nonlinear fitting.

4. Teams and Methods

4.1. MICV Team

The MICV team proposes the Hybrid Vision Transformer (ViT) and Convolutional Neural Network (CNN) for SR Image Quality Assessment [14]. The proposed method focuses solely on the SR images as input, leveraging the ViT to capture global dependencies and the CNN to extract spatial features. The architecture of the method is illustrated in Figure 3. To adaptively model the visual features of SR images, they introduce a multi-stage attention mechanism that enhances hierarchical feature fusion through self-attention and transposed self-attention.

Specifically, the ViT takes the SR image as input and generates high-level image tokens. These tokens are first processed by a self-attention module to capture global contextual relationships. Subsequently, the output is transposed and fed into a second self-attention layer, enabling the model to learn dependencies along alternative spatial directions. The output of the transposed self-attention is then reverted to its original dimensionality and passed through a third self-attention layer, further refining cross-scale interactions. Finally, the enriched feature map is encoded through multiple convolution layers, linear layers, and a sigmoid activation function to predict the quality score.

By eliminating the need for LR & HR priors, the model achieves a more streamlined architecture while maintaining strong performance in quality estimation. This design not only reduces computational complexity and GPU memory, but also avoids potential biases introduced by reference im-

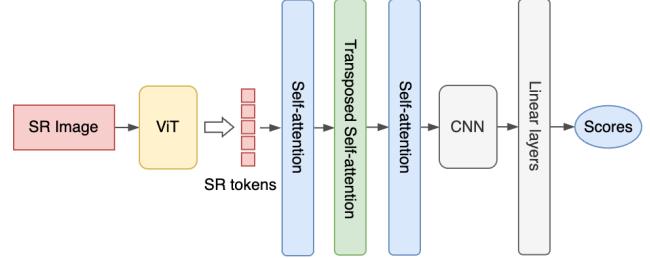


Figure 3. Model architecture of the proposed Hybrid Vision Transformer and Convolutional Neural Network for Super-Resolution Image Quality Assessment.

age dependencies.

Training Setup This task aims to learn models to predict the MOS of 21 participants for SR images. The training data consists of 576 SR images from the ISRGen-QA dataset, with their corresponding low- and high-resolution counterparts, covering upscaling levels of $\times 2$, $\times 3$, $\times 4$, and $\times 8$. During the data pre-processing phase, the SR images are directly used as input, with no reference to LR or HR counterparts. Data augmentation is implemented through random horizontal flipping and random cropping operations to enhance model robustness. The resulting cropped images maintain a resolution of 448x448 pixels, with any entirely black images being re-cropped to ensure data validity. The AdamW optimizer is employed with an initial learning rate of 1e-5, and learning rate scheduling is conducted using co-

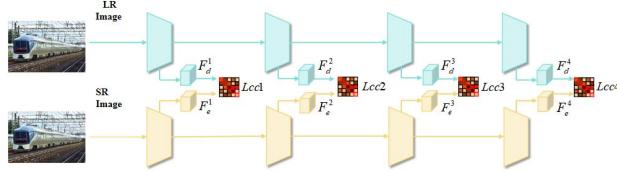


Figure 4. Model architecture of the proposed Cross-Covariance Loss Calculation in team ydy.

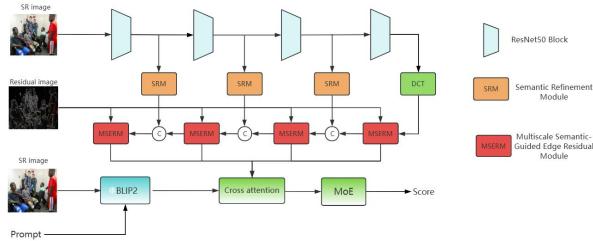


Figure 5. Model framework of the proposed BLIP-2 Assisted Residual-Guided Quality Assessment for Super-Resolution Images.

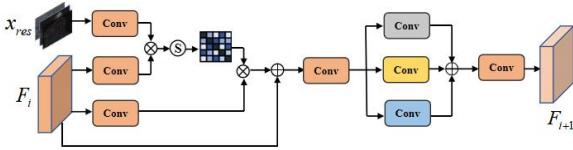


Figure 6. Model framework of the proposed SRM: Semantic Refinement Module in team ydy.

sine annealing. The loss function was defined as a weighted sum of the PLCC loss and SRCC loss with equal weighting ratios of 1:1. The training is conducted on 8 NVIDIA A100 GPUs with a batch size of 4 for 200 epochs.

Testing Details During the validation phase, 72 SR images are utilized to evaluate the effectiveness of the trained model. Following the same pre-processing procedure as in the training phase, only the SR images are processed, with no involvement of LR or HR counterparts. Subsequently, center cropping is performed to extract 448×448 resolution images, which are then directly input into the network for quality score prediction. Subsequently, center cropping is performed to extract input resolution images, which are then processed by the network for quality score prediction.

4.2. ydy Team

The ydy team proposes the BLIP-2 Assisted Residual-Guided Quality Assessment for Super-Resolution Images [29], which is a hybrid quality assessment network tailored for SR images, integrating semantic information from

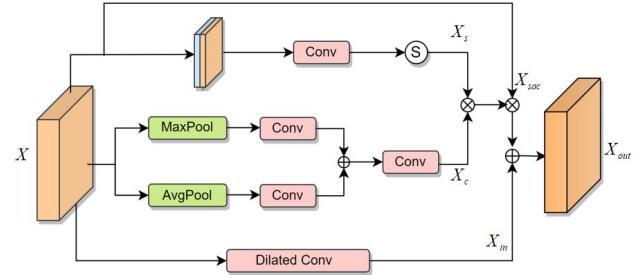


Figure 7. Model framework of the proposed MSERM: Multiscale Semantic-Guided Edge Residual Module in team ydy.

BLIP-2 features and residual guidance from precomputed residual maps. The architecture builds upon a dual-branch backbone with ResNet50 encoders for both super-resolved and low-resolution images. The model is designed to mitigate SR-specific artifacts, such as edge distortions and texture inconsistencies, through multi-level residual guidance and semantic alignment in SR-IQA tasks. First, hierarchical features from the SR image and its corresponding low-quality reference are extracted using two parallel ResNet50 backbones, as shown in Figure 4. Their feature-level similarity is supervised via a cross whitening loss, ensuring consistency at multiple depths. The residual image—capturing high-frequency differences between SR and LR—is processed through a multi-stage pooling operation to generate hierarchical residual cues. These cues are fused with the backbone features via a series of channel-wise Self-Attention blocks, allowing residual signals to guide perceptual feature enhancement, as shown in Figure 5. The Semantic Refinement Module (SRM, as shown in Figure 6) enhances semantic representation between the encoder and decoder by integrating spatial, channel, and semantic cues. It captures global context via dilated convolution and refines features through spatial and channel attention, whose outputs are fused and combined with semantic features to guide decoding. And the Multiscale Semantic-guided Edge Residual Module (MSERM, as shown in Figure 7) leverages semantic-aware edge residuals to refine features using an attention mechanism, followed by multiscale enhancement via parallel dilated convolutions. Together, SRM and MSERM ensure semantically rich, edge-sensitive, and context-aware feature representations. Finally, a Mixture-of-Experts (MoE) gating mechanism is applied, where a set of projection experts is weighted dynamically based on the gated attention output to produce the final quality prediction.

Training Setup The proposed model was implemented using PyTorch 2.0.0 and Python 3.9, and trained on a single NVIDIA RTX 3090 GPU. The training was conducted

for 100 epochs with a batch size of 16, requiring approximately 8 hours in total. They used the official ISRG-Q training dataset, which includes 576 SR images along with their corresponding LR and HR reference images and MOS as supervision signals. For each image, 30 training and 30 test patches were randomly sampled. Semantic features extracted from BLIP-2 were precomputed and stored as ‘.pt’ files for training. The optimization was performed using Adam with a fixed learning rate of 1e-4 and a weight decay of 5e-4. During training, they applied standard data augmentation techniques including random horizontal flipping, random cropping to 128×128 patches, and normalization using ImageNet statistics. The model was trained with a multi-objective loss function, combining an L1 loss between predicted scores and MOS, a Cross-Covariance Loss (CCL) between multilevel SR and LR features, and a cosine similarity loss between the visual-semantic outputs and the precomputed BLIP-2 query vectors.

Testing Details During testing, the proposed model followed the same patch-wise strategy used during training. Each super-resolved image was divided into 30 patches of size 128×128, and a quality score was predicted for each patch. The final image-level score was obtained by averaging the patch-level predictions. As a test-time preprocessing step, they applied center cropping to extract representative image regions. This helps standardize input distributions and reduce prediction variance. BLIP-2 multimodal features were pre-extracted and loaded during inference. All testing was conducted on an NVIDIA RTX 3090 GPU.

4.3. QA-Veteran Team

The QA-Veteran team proposes Blind Super-resolution Quality Assessment based on a Resolution-adaptive Vision-language Model [49]. Image super-resolution aims to recover a high-resolution (HR) image from an LR input. This is a typical ill-posed problem, as multiple plausible HR outputs may correspond to the same LR input. Therefore, to reliably evaluate the quality of super-resolved images, they argue that a no-reference (NR) or blind IQA approach should be adopted. Compared with natural images (e.g., photos taken by a camera), super-resolved images exhibit two distinct characteristics:

- **Algorithm-dependence:** The quality scores of super-resolved images are highly influenced by the specific super-resolution algorithm used. Therefore, an IQA model must be capable of effectively capturing the degradation artifacts introduced by different algorithms.
- **High resolution** (e.g., 2K, 4K, etc.): This poses a challenge when applying conventional image preprocessing methods, such as aggressively downsampling the image before feeding it into the model. Such downsampling may obscure the fine-grained features that are crucial

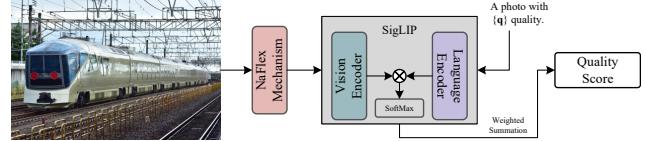


Figure 8. The diagram of the proposed Blind Super-resolution Quality Assessment based on a Resolution-adaptive Vision-language Model.

for distinguishing the quality differences between high-resolution super-resolved images, ultimately reducing the effectiveness of the quality assessment.

To address the above challenges, the proposed method is built upon **SigLIP2-NaFlex** [31]. On one hand, the model benefits from large-scale image-text pretraining, which enables it to learn rich image representations and better capture the distortion characteristics introduced by different super-resolution algorithms. On the other hand, the **NaFlex** mechanism in SigLIP2 preserves the original *aspect ratio* and *resolution* of super-resolved images as much as possible, allowing the model to retain fine-grained quality cues that are crucial for distinguishing between high-resolution super-resolved images.

As shown in Figure 8, given an input super-resolved image \mathbf{x} , they leverage the vision encoder within the SigLIP-2 model to obtain a visual embedding vector. Inspired by [48], they design a textual template: “*a photo with {c} quality*”, where $c \in \mathcal{C} = \{1, 2, 3, 4, 5\} = \{\text{“bad”}, \text{“poor”}, \text{“fair”}, \text{“good”}, \text{“perfect”}\}$. They then use the language encoder within the SigLIP-2 model to encode the textual embedding of all entries in the textual template. They compute the cosine similarity of the visual embedding and all the textual entries and apply a softmax function to obtain the probability distribution of \mathbf{x} over five quality levels $\hat{p}(c|\mathbf{x})$. They relate c to a scalar quality score \hat{q} by $\hat{q}(\mathbf{x}) = \sum_{c=1}^C \hat{p}(c|\mathbf{x}) \times c$, where $C = 5$ is the number of quality levels.

Training Setup They build their model on the SigLIP2-base-patch16-NaFlex. During training, they randomly choose the maximum number of patches from 4624, 5184, and 5776. They train the model using a single NVIDIA A5880-ada GPU, using the AdamW optimizer [22] with a decoupled weight decay regularization of 10^{-3} . The initial learning rate is set to 5×10^{-6} , which is scheduled by a cosine annealing rule [21]. They optimize the model for 6 epochs with a mini-batch size of 12. They use a combination of fidelity loss [47], PLCC loss, and L1 loss as the loss function.

Testing Details During inference, they fix the maximum number of patches to 4624. The NaFlex mechanism will

adaptively preprocess the input image.

4.4. 2077 Agent Team

The 2077 Agent team proposes the Ultra-High-Resolution Image Quality Assessment[27]. SR images refer to images converted from LR to HR through super-resolution reconstruction techniques, with pixel density ranging from tens to hundreds of times that of traditional low-resolution images, thus belonging to the category of ultra-high-resolution images. Taking 8K super-resolution images as an example, their resolution of 33 million pixels significantly enhances the capability of detail representation. However, the quality assessment of super-resolution images faces significant challenges:

- **Massive Data Volume:** The enormous data size of SR images results in prohibitively high computational and storage costs in traditional pixel-level evaluation algorithms, making effective training and deployment difficult.
- **Architectural Limitations:** Existing frameworks based on CNN and Transformers are primarily designed for low-resolution images, struggling to adapt to the scale characteristics of SR images.

Traditional high-resolution image quality assessment methods typically employ downsampling or patch-based cropping strategies, but these approaches have notable limitations:

- **Downsampling** inevitably leads to loss of fine details, compromising the accuracy of the assessment.
- **Patch-based Evaluation** divides an image into blocks and calculates an average score as the overall quality metric, implicitly assuming uniform contributions from all blocks. However, this assumption often fails in practice. For instance, when an SR image contains entirely black anomalous blocks, human subjective evaluation would consider such blocks significantly detrimental to overall quality, whereas average scoring underestimates their negative impact, creating a substantial discrepancy between algorithmic results and human perception. Consequently, novel algorithms are urgently needed to address these technical bottlenecks in SR image quality assessment.

To overcome these issues, this study breaks away from the conventional MOS evaluation paradigm for entire images. Instead, it dynamically allocates weights based on the varying contributions of different regions to subjective quality perception: texture-rich detail regions are assigned higher weights, while smooth background regions receive lower weights. The final weighted regional evaluations are globally fused via a Score Transformer, generating MOS scores that better align with human visual perception.

Thus, this study proposes **UltraR-IQA**, an ultra-high-resolution image quality assessment algorithm, as shown in

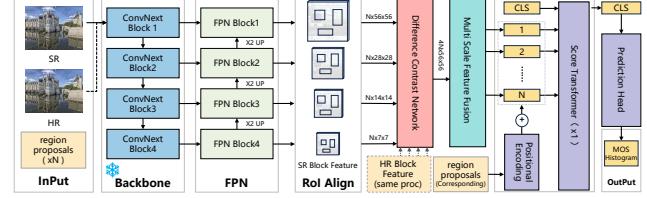


Figure 9. UltraR-IQA Network Architecture: Inputting SR/HR and pre-generated candidate boxes, freezing ConvNext Base backbone parameters, randomly selecting N candidate boxes for ROI Align cropping with the corresponding HR/SR cropping and the same processing; feeding same-layer features into difference Contrast network, fusing cross-layer features via multi-scale network to obtain latent scores, embedding candidate positions with position encoding, and outputting a 5-dimensional frequency distribution.

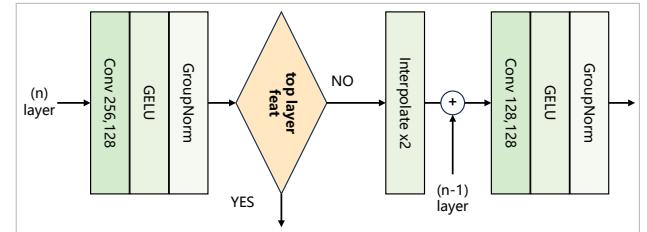


Figure 10. FPN Module Architecture: Taking multi-level ConvNext features as input and outputting 256-channel features; non-highest level inputs undergo upsampling, addition with upper-level features, and subsequent convolution.

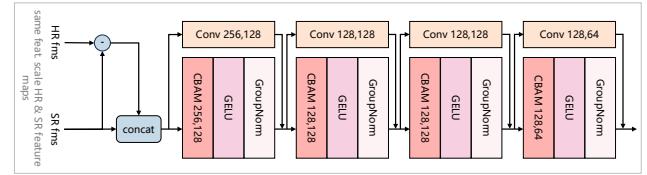


Figure 11. Difference Contrast Network Architecture: Inputting same-layer candidate box features, where “-” denotes feature subtraction and “concat” represents channel-wise concatenation.

Figure 9. By constructing a multi-stage processing system, it enables quantitative quality comparisons between SR images and HR images across multiple spatial scales and abstraction levels. The algorithmic architecture comprises five core modules:

- **Candidate Region Generation:** Candidate Region Generation extracts key candidate regions from HR images using the Selective Search algorithm, laying the foundation for subsequent fine-grained analysis due to its ability to generate high-quality candidate regions with varying scales and aspect ratios, crucial for capturing diverse visual features in SR and HR images.
- **Multi-Scale Feature Extraction:** Combines the

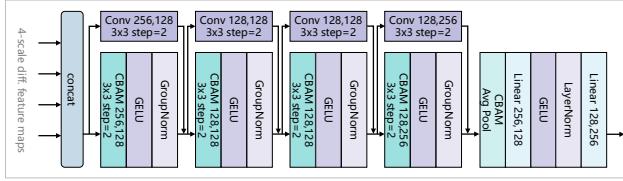


Figure 12. Multi-Scale Feature Fusion Architecture: Employing convolutional downsampling with padding 1 and stride 3; taking multi-scale difference features as input and outputting latent scores for candidate regions.

ConvNeXt-Base backbone network with a Feature Pyramid Network (FPN, as shown in Figure 10) to efficiently extract multi-scale features from HR and SR images.

- **Difference Contrast Network:** For each candidate region, a difference contrast network (as shown in Figure 11) that integrates residual connections and Convolutional Block Attention Modules (CBAM) precisely computes feature discrepancies.
- **Multi-Scale Feature Fusion:** As shown in Figure 12, this module generates regional latent scores through a multi-scale fusion strategy, emphasizing the complementary nature of features at different levels.
- **Score Prediction:** Leverages a Score Transformer model incorporating Fourier and geometric positional encoding to aggregate regional latent scores and output a 5-level MOS frequency histogram distribution, mapping feature differences to subjective quality scores.

For scenarios involving gigapixel ultra-high-resolution images with limited local computational resources, this method supports decomposing images into multiple sub-regions for serial processing. By dynamically allocating region-specific weights via an attention mechanism, it achieves efficient and accurate image quality assessment, effectively mitigating the underestimation of outlier impacts caused by average scoring. When computational resources are sufficient, the entire image can also be processed directly, with both approaches yielding approximately equivalent evaluation results.

Training Setup This study pioneers a minimum-variance constrained approach for constructing MOS frequency histograms, identifying scoring combinations that maximize alignment with target MOS values while minimizing variance. According to competition protocols, MOS scores are independently assigned by 21 evaluators. Assuming integer ratings within the 1-5 range, the algorithm employs five-layer nested loops to exhaustively traverse all possible scoring combinations across evaluators, filtering sets where the mean score equals the target value. Subsequently, variance is computed for each qualifying combination, with lower variance indicating more concentrated score distribu-

tions. The minimal-variance combination is selected as the optimal solution, forming the foundational data for MOS frequency histogram construction. Computational verification confirms that the constructed MOS frequency histogram exhibits an expectation error of 2.56×10^{-15} relative to ground truth, with each MOS score corresponding to a unique optimal combination. Finally, the model employs a KL-divergence loss function, deriving the ultimate MOS prediction by computing the mathematical expectation of this 5-dimensional MOS frequency histogram.

The framework utilizes a pre-trained ConvNeXt-Base backbone network (supporting Tiny variants). Optimization employs the AdamW optimizer with initial learning rate 1×10^{-4} and weight decay coefficient 0.1. Training employs an actual batch size of 1 with gradient accumulation over 4 steps (effective batch size=4) for 100 epochs. Learning rate scheduling follows the StepLR strategy, specifically halving the rate every 25 epochs to enable dynamic optimization.

During image preprocessing, HR and SR images undergo normalization before SR images are overlaid onto HR counterparts using the top-left corner as the alignment origin for dimensional unification. Candidate regions meeting specified criteria are pre-generated via the Selective Search algorithm (persisted before training). Training data augmentation incorporates random horizontal flipping (50% probability) and 90° interval rotation. All training procedures were executed on a single NVIDIA RTX 2080 Ti GPU.

Testing Details During the inference phase, the operational workflow mirrors the training procedure. The system executes multiple iterations of candidate region sampling and inference operations (default: 20 iterations) based on precomputed candidate regions, subsequently averaging the predictions across iterations to enhance result stability. Furthermore, input images consistently retain their original resolution without modification.

5. Acknowledgments

We thank the VQualA 2025 sponsors: Snap Research, INTSIG, and TAOBAO & TMALL Group.

References

- [1] 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*, 2017. 2
- [2] Ying Chen, Huasheng Wang, Pengxiang Xiao, Yukang Ding, Enpeng Liu, Chris Wei Zhou, and et al. VQualA 2025 Challenge on GenAI-Bench AIGC Video Quality Assessment:

- Methods and Results. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, pages 1–11, 2025. 2
- [3] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-Order Attention Network for Single Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 2
- [4] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit Diffusion Models for Continuous Super-Resolution. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 10021–10030, 2023. 2
- [5] Yunxing Gao, Hengjian Li, Jiwen Dong, and Guang Feng. A deep convolutional network for medical image super-resolution. In *2017 Chinese Automation Congress (CAC)*, pages 5310–5315, 2017. 2
- [6] Fan Huang, Xiongkuo Min, Zhichao Ma, Xiaohong Liu, Chris Wei Zhou, Guangtao Zhai, and et al. VQualA 2025 Document Image Quality Assessment Challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, pages 1–8. 2
- [7] Qiuping Jiang, Zhentao Liu, Ke Gu, Feng Shao, Xinfeng Zhang, Hantao Liu, and Weisi Lin. Single Image Super-Resolution Quality Assessment: A Real-World Dataset, Subjective Studies, and an Objective Metric. *IEEE Transactions on Image Processing*, 31:2279–2294, 2022. 1, 2
- [8] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-Recursive Convolutional Network for Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1637–1645, 2016. 2
- [9] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016. 2
- [10] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 2
- [11] Bingchen Li, Xin Li, Hanxin Zhu, Yeying Jin, Ruoyu Feng, Zhizheng Zhang, and Zhibo Chen. SeD: Semantic-Aware Discriminator for Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25784–25795, 2024. 2
- [12] Dasong Li, Sizhuo Ma, Hang Hua, Wenjie Li, Jian Wang, Chris Wei Zhou, Fengbin Guan, Xin Li, Zihao Yu, Yiting Lu, Ru-Ling Liao, Yan Ye, Zhibo Chen, Wei Sun, Linhan Cao, Yuqin Cao, Weixia Zhang, Wen Wen, Kaiwei Zhang, Zijian Chen, Fangfang Lu, Xiongkuo Min, Guangtao Zhai, Erjia Xiao, Lingfeng Zhang, Zhenjie Su, Hao Cheng, Yu Liu, Renjing Xu, Long Chen, Xiaoshuai Hao, Zhenpeng Zeng, Jianjin Wu, Xuxu Wang, Qian Yu, Bo Hu, Weiwei Wang, Pinxin Liu, Yunlong Tong, Luchuan Song, Jinxi He, Jiaru Wu, and Hanjia Lyu. VQualA 2025 Challenge on Engagement Prediction for Short Videos: Methods and Results. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, pages 1–13, 2025. 2
- [13] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueling Chen. SRDiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 2
- [14] Xinyu Li, Chuanbiao Song, Chenqi Zhang, Jun Lan, Huijia Zhu, Weiqiang Wang, and Xiaoyan Sun. Hybrid Vision Transformer and Convolutional Neural Network for Super-Resolution Image Quality Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, 2025. 4
- [15] Xingyuan Li, Zirui Wang, Yang Zou, Zhixin Chen, Jun Ma, Zhiying Jiang, Long Ma, and Jinyuan Liu. DifIISR: A Diffusion Model with Gradient Guidance for Infrared Image Super-Resolution. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 7534–7544, 2025. 2
- [16] Yixiao Li, Xiaoyuan Yang, Jun Fu, Guanghui Yue, and Wei Zhou. Deep bi-directional attention network for image super-resolution quality assessment. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 1
- [17] Yixiao Li, Xin Li, Chris Wei Zhou, Shuo Xing, Hadi Amirpour, Xiaoshuai Hao, Guanghui Yue, Baoquan Zhao, Weide Liu, Xiaoyuan Yang, Zhengzhong Tu, and et al. VQualA 2025 Challenge on Image Super-Resolution Generated Content Quality Assessment: Methods and Results. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, pages 1–10, 2025. 2
- [18] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image Restoration Using Swin Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 2
- [19] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017. 2
- [20] Qingguo Liu, Chenyi Zhuang, Pan Gao, and Jie Qin. CDFormer: When Degradation Prediction Embraces Diffusion Model for Blind Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7455–7464, 2024. 2
- [21] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*, 2017. 6
- [22] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. 6
- [23] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for

- single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 1
- [24] Sizhuo Ma, Wei-Ting Chen, Qiang Gao, Jian Wang, Chris Wei Zhou, Wei Sun, Weixia Zhang, Linhan Cao, Jun Jia, Xiangyang Zhu, Dandan Zhu, Xiongkuo Min, Guangtao Zhai, Baoying Chen, Xiongwei Xiao, Jishen Zeng, Wei Wu, Tiexuan Lou, Yuchen Tan, Chunyi Song, Zhiwei Xu, MohammadAli Hamidi, Hadi Amirpour, Mingyin Bai, Jiawang Du, Zhenyu Jiang, Zilong Lu, Ziguang Cui, Zongliang Gan, Xinpeng Li, Shiqi Jiang, Chenhui Li, Changbo Wang, Weijun Yuan, Zhan Li, Yihang Chen, Yifan Deng, Ruting Deng, Zhanglu Chen, Boyang Yao, Shuling Zheng, Feng Zhang, Zhiheng Fu, Abhishek Joshi, Aman Agarwal, Rakhil Immidisetti, Ajay Narasimha Mopidevi, Vishwajeet Shukla, Hao Yang, Ruikun Zhang, Liyuan Pan, Kaixin Deng, Hang Ouyang, Fan Yang, Zhizun Luo, Zuochang Shi, Songning Lai, Weilin Ruan, and Yutao Yue. VQualA 2025 Challenge on Face Image Quality Assessment: Methods and Results. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, pages 1–10, 2025. 2
- [25] Brian B Moser, Stanislav Frolov, Federico Raue, Sebastian Palacio, and Andreas Dengel. Dynamic Attention-Guided Diffusion for Image Super-Resolution. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 451–460, 2025. 2
- [26] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image Super-Resolution via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 2
- [27] Zuochang Shi, Zhizun Luo, Hang Ouyang, Tianxin Xiao, Fan Yang, Zhaochang Wu, and Kaixin Deng. Ultra-High-Resolution Image Quality Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, 2025. 7
- [28] Wen Sun, Fei Zhou, and Qingmin Liao. MDID: A multiply distorted image database for image quality assessment. *Pattern Recognition*, 61:153–168, 2017. 1
- [29] Shishun Tian and Dongyang Yan. BLIP-2 Assisted Residual-Guided Quality Assessment for Super-Resolution Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025. 5
- [30] Li-Yuan Tsao, Yi-Chen Lo, Chia-Che Chang, Hao-Wei Chen, Roy Tseng, Chien Feng, and Chun-Yi Lee. Boosting Flow-based Generative Super-Resolution Models via Learned Prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26005–26015, 2024. 2
- [31] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features, 2025. 6
- [32] Wenfei Wan, Jinjian Wu, Guangming Shi, Yongbo Li, and Weisheng Dong. Super-Resolution Quality Assessment: Subjective Evaluation Database and Quality Index Based on Perceptual Structure Measurement. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 1, 2
- [33] Chengcheng Wang, Zhiwei Hao, Yehui Tang, Jianyuan Guo, Yujie Yang, Kai Han, and Yunhe Wang. SAM-DiffSR: Structure-Modulated Diffusion Model for Image Super-Resolution. *arXiv preprint arXiv:2402.17133*, 2024. 2
- [34] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 0–0, 2018. 2
- [35] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training Real-World Blind Super-Resolution With Pure Synthetic Data. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 1905–1914, 2021. 2
- [36] Yan Wang, Yi Liu, Shijie Zhao, Junlin Li, and Li Zhang. CAMixerSR: Only Details Need More “Attention”. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25837–25846, 2024. 2
- [37] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep Networks for Image Super-Resolution With Sparse Prior. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 370–378, 2015. 2
- [38] Min Wei and Xuesong Zhang. Super-Resolution Neural Operator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18247–18256, 2023. 2
- [39] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 25456–25467, 2024. 2
- [40] Jiarui Yang, Tao Dai, Yufei Zhu, Naiqi Li, Jinmin Li, and Shu-Tao Xia. Diffusion Prior Interpolation for Flexibility Real-World Face Super-Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9211–9219, 2025. 2
- [41] Hojatollah Yeganeh, Mohammad Rostami, and Zhou Wang. Objective Quality Assessment of Interpolated Natural Images. *IEEE Transactions on Image Processing*, 24(11):4651–4663, 2015. 1, 2
- [42] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling Up to Excellence: Practicing Model Scaling for Photo-Realistic Image Restoration In the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25669–25680, 2024. 2
- [43] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep Unfolding Network for Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3217–3226, 2020. 2
- [44] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. In *Proceedings of the IEEE/CVF*

- Computer Vision and Pattern Recognition Conference*, pages 4791–4800, 2021. 2
- [45] Leheng Zhang, Yawei Li, Xingyu Zhou, Xiaorui Zhao, and Shuhang Gu. Transcending the Limit of Local Window: Advanced Super-Resolution Transformer with Adaptive Token Dictionary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, 2024. 2
- [46] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. RankSRGAN: Generative Adversarial Networks With Ranker for Image Super-Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 2
- [47] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021. 6
- [48] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14071–14081, 2023. 6
- [49] Weixia Zhang, Junlin Chen, Wei Sun, and Zhihua Wang. Blind Super-resolution Quality Assessment based on a Resolution-adaptive Vision-language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, 2025. 6
- [50] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *Proceedings of the European Conference on Computer Vision*, pages 286–301, 2018. 2
- [51] Tiesong Zhao, Yuting Lin, Yiwen Xu, Weiling Chen, and Zhou Wang. Learning-Based Quality Assessment for Image Super-Resolution. *IEEE Transactions on Multimedia*, 24:3570–3581, 2022. 1, 2
- [52] Fei Zhou, Rongguo Yao, Bozhi Liu, and Guoping Qiu. Visual Quality Assessment for Super-Resolved Images: Database and Method. *IEEE Transactions on Image Processing*, 28(7):3528–3541, 2019. 1
- [53] Wei Zhou and Zhou Wang. Quality Assessment of Image Super-Resolution: Balancing Deterministic and Statistical Fidelity. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 934–942, 2022. 1
- [54] Wei Zhou, Qiuping Jiang, Yuwang Wang, Zhibo Chen, and Weiping Li. Blind quality assessment for image superresolution using deep two-stream convolutional networks. *Information Sciences*, 528:205–218, 2020. 1
- [55] Wei Zhou, Zhou Wang, and Zhibo Chen. Image Super-Resolution Quality Assessment: Structural Fidelity Versus Statistical Naturalness. In *2021 13th International conference on quality of multimedia experience (QoMEX)*, pages 61–64. IEEE, 2021. 1
- [56] Wei Zhou, Hadi Amirpour, Christian Timmerer, Guangtao Zhai, Patrick Le Callet, and Alan C Bovik. Perceptual Visual Quality Assessment: Principles, Methods, and Future Directions. *arXiv preprint arXiv:2503.00625*, 2025. 1
- [57] Hanwei Zhu, Haoning Wu, Zicheng Zhang, Lingyu Zhu, Yixuan Li, Peilin Chen, Shiqi Wang, Chris Wei Zhou, Linhan Cao, Wei Sun, Xiangyang Zhu, Weixia Zhang, Yucheng Zhu, Jing Liu, Dandan Zhu, Guantao Zhai, Xiongkuo Min, Zhichao Zhang, Xinyue Li, Shubo Xu, Anh Dao, Yifan Li, Hongyuan Yu, Jiaojiao Yi, Yiding Tian, Yupeng Wu, Feiran Sun, Jiao Lijuan, and Song Jiang. VQualA 2025 Challenge on Visual Quality Comparison for Large Multimodal Models: Methods and Results. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, pages 1–11, 2025. 2