

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/190283/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Zheng, Bolun, Weng, Zijie, Zhang, Qianyu, Zhang, Hua, Yan, Chenggang, Jia, Xu and Zhou, Wei 2026. Perceptual quality assessment of real-world noise in ultra-high-definition images. IEEE Transactions on Image Processing 35 , pp. 10757-10772. 10.1109/tip.2026.3739609

Publishers page: <https://doi.org/10.1109/tip.2026.3739609>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



Perceptual Quality Assessment of Real-World Noise in Ultra-High-Definition Images

Bolun Zheng[✉], *Member, IEEE*, Zijie Weng[✉], Qianyu Zhang[✉], Hua Zhang[✉], Chenggang Yan[✉], Xu Jia[✉],
and Wei Zhou[✉], *Senior Member, IEEE*

Abstract—Accurately assessing the perceptual quality of ultra-high-definition (UHD) images degraded by real-world noise remains a formidable challenge in computer vision. Existing Blind Image Quality Assessment (BIQA) methods, typically tailored for synthetic distortions and low-resolution inputs, often fail to generalize to these complex real-world scenarios. To bridge this gap, we first establish a real-world Ultra-High-Definition Noisy Image Quality (UNIQ-5k) dataset, consisting of 4,759 UHD images with fine-grained perceptual annotations. Leveraging this foundation, we propose PINE, a novel noise-oriented BIQA method. To tackle the computational burden of UHD inputs, we adopt a Grid-based Fragment Sampling Strategy (GFSS) as an efficient preprocessing component, which extracts representative local features while preserving global semantic topology. Subsequently, PINE incorporates a dual-branch architecture to simulate the noise visibility characteristics of the Human Visual System (HVS): the Perceptual Spatial Attention Module (PSAM) exploits the texture masking effect to prioritize noise in smooth regions; the Perceptual Frequency Enhancement Module (PFEM) utilizes a Modified Noise-Aware Contrast Sensitivity Function (CSF) to explicitly emphasize medium-to-high-frequency noise components. Extensive experiments demonstrate that our method achieves state-of-the-art performance on the UNIQ-5k dataset and high rank consistency with PSNR and SSIM across the evaluated SIDD denoising results, showing that it may serve as an auxiliary no-reference indicator for restoration-oriented comparisons. The codes and models will be available at <https://github.com/zijieweng123/PINE.git>

Index Terms—Blind image quality assessment, real-world noise, ultra-high-definition, IQA database, denoising evaluation metric.

I. INTRODUCTION

NOWADAYS, the camera has evolved into a pivotal component of consumer electronic devices, establishing

ultra-high-definition (UHD) imaging as the mainstream standard [1]. Many manufacturers increasingly highlight camera performance as a keypoint, and the ability to suppress image noise serves as a crucial indicator [2]. However, the transition to modern UHD imaging has fundamentally transformed the theoretical characteristics of image noise. The substantial increase in resolution reshapes artifacts into fine-grained, high-frequency textures, while advanced image signal processing (ISP) can suppress simple additive noise, leaving behind complex, signal-dependent residual noise that defy traditional Gaussian modeling [3]. These subtle degradations become highly perceptible under the stringent viewing conditions of UHD displays, significantly degrading the user experience. Therefore, accurately assessing the perceptual quality of these images is essential for guiding the iterative optimization of camera systems.

Given the unavailability of reference images in authentic photography, Blind Image Quality Assessment (BIQA) becomes the only viable solution. Unfortunately, existing resources fall short of this need. Most IQA datasets are dominated by synthetic distortions or mixed distortions [4], [5], [6], [7]. In authentic datasets, subjective ratings reflect holistic quality, where noise is inevitably conflated with other degradations, preventing specific assessment. While synthetic datasets isolate noise, they rely on simplistic simulations that yield spatially uniform and exaggerated artifacts. These artificial patterns fail to capture the subtle, signal-dependent nature of modern imaging, creating a substantial domain gap [8]. Consequently, these datasets are unable to capture or assess the real noise introduced by nonlinear in-camera ISP algorithms, highlighting the need for a dedicated real-world noise dataset [9]. On the other hand, general-purpose BIQA models exhibit limited capability in recognizing and quantifying noise-specific degradation [10]. While general BIQA methods excel at overall quality scoring, they often lack the sensitivity required for camera benchmarking, where isolating specific noise artifacts is paramount. Their predictions on real noisy UHD images are frequently unstable, as they tend to be highly influenced by image content characteristics such as illumination and texture strength [11]. A motivating example is presented in Fig. 1. Existing metrics are easily misled by high luminance, erroneously assigning high scores to bright but noisy images. In contrast, our method surmounts this content bias, ensuring an accurate assessment consistent with human judgment. Furthermore, existing noise-specific BIQA methods face significant limitations. Many rely on natural scene statistics (NSS) features or synthetic data [12], [13], failing to generalize to complex real-world noise. Others depend on external metrics or denoising networks [14],

Received 26 December 2025; revised 24 August 2026; accepted 25 September 2026. Date of current version 9 October 2026. This work was supported by the National Natural Science Foundation of China under Grant 62371175. The associate editor coordinating the review of this article and approving it for publication was Prof. Jon Y. Hardeberg. (*Corresponding author: Qianyu Zhang.*)

Bolun Zheng is with the School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China, and also with Zhejiang Provincial Key Laboratory of Low Altitude Ubiquitous Networking Technology, Hangzhou 310018, China (e-mail: blzheng@hdu.edu.cn).

Zijie Weng, Qianyu Zhang, and Chenggang Yan are with the School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: zjweng@hdu.edu.cn; qyzhang@hdu.edu.cn; cgyan@hdu.edu.cn).

Hua Zhang is with the School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: zhangh@hdu.edu.cn).

Xu Jia is with the School of Future Technology and the School of Artificial Intelligence, Dalian University of Technology, Dalian 116024, China (e-mail: jiayushenyang@gmail.com).

Wei Zhou is with the School of Computational and Mathematical Sciences, Cardiff University, CF24 4AG Cardiff, U.K. (e-mail: ZhouW26@cardiff.ac.uk).

Digital Object Identifier 10.1109/TIP.2026.3739609

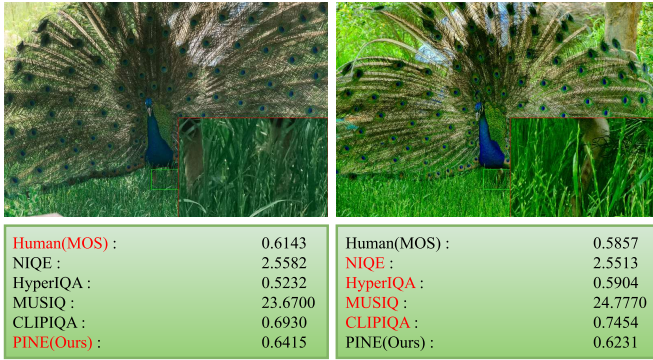


Fig. 1. Comparison between objective metrics and human preference on samples from the UNIQ-5k dataset. Notably, although the right image exhibits higher brightness than the left image, inspection of the zoomed-in patches reveals that the right image suffers from severe noise. Consequently, human subjects assign a higher Mean Opinion Score (MOS) to the left image. As shown by the predicted scores, PINE is less affected by the brightness difference than the compared metrics and produces a ranking consistent with the collected MOS.

which compromises evaluation purity and hinders the development of a unified framework aligned with human perception.

To address these challenges, we first construct a new real-world Ultra-High-Definition Noisy Image Quality (UNIQ-5k) dataset, comprising 4,759 UHD images captured in various real-world environments using diverse devices. Our UNIQ-5k emphasizes subjective ratings of noise visibility and perceptual quality, providing a reliable benchmark for modeling noise perception in realistic settings. Building upon UNIQ-5k, we propose a noise-oriented BIQA method, named Perceptual Image Noise Evaluator (PINE). PINE simulates HVS noise perception by introducing the Perceptual Spatial Attention Module (PSAM) to model texture masking via the Noise Visibility Function (NVF), and the Perceptual Frequency Enhancement Module (PFEM) to capture frequency-dependent noise sensitivity. Moreover, following the Grid Mini-patch Sampling (GMS) principle introduced in FAST-VQA [15], we adopt a Grid-based Fragment Sampling Strategy (GFSS) for UHD still images to reduce computational cost while preserving global structures and local noise details. Generally, the contributions of this work can be summarized as follows:

- We construct a large-scale real-world Ultra-High-Definition Noisy Image Quality (UNIQ-5k) dataset. As far as we know, this is the first benchmark dedicated to the subjective quality assessment of noise degradations in UHD images. Extensive experiments demonstrate that our UNIQ-5k dataset enhances the capability of existing methods to characterize real-world noise in UHD images.
- We propose PINE, a noise-oriented BIQA method inspired by HVS mechanisms. PSAM models noise masking via the NVF, while the proposed PFEM captures frequency-dependent noise sensitivity.
- Extensive experiments show that PINE performs strongly on UNIQ-5k and the evaluated public HR-IQA datasets, while exhibiting high algorithm-level rank consistency with PSNR and SSIM on the evaluated SIDD denoising results.

II. RELATED WORK

This section provides an overview of existing works related to our study. We first introduce the relevant IQA datasets. Then, we discuss representative BIQA methods, categorizing them into general models, High-Resolution (HR) strategies, and noise-specific approaches.

A. IQA Datasets

Early IQA research relied heavily on datasets constructed with artificial distortions. Pioneering benchmarks such as LIVE [16], CSIQ [17], TID2013 [4], and KADID-10k [5] introduced various synthetic artifacts (e.g., Gaussian blur, white noise, JPEG compression) into a set of high-quality reference images. While fundamental to the field, these datasets typically feature low-resolution images and fail to capture the complexity of signal-dependent noise inherent in modern ISP pipelines, exhibiting a significant domain gap from real-world imaging.

To address the domain gap, “in-the-wild” datasets containing user-generated content (UGC) were introduced. Notable examples include CLIVE [6], KonIQ-10k [7], and PaQ-2-PaQ [18]. These datasets capture authentically distorted images using various consumer devices, with subjective scores collected via crowdsourcing. However, a critical limitation persists: their image resolutions are predominantly low, lagging significantly behind the capture capabilities of contemporary UHD devices.

Recognizing the resolution bottleneck, researchers have begun to construct HR datasets. Early efforts like CID2013 [19] explored this direction but were limited by a small scale (480 images). SPAQ [20] significantly expanded the scale with 11,125 images captured by 66 smartphones; however, images were resized (longest side limited to 512 pixels) during subjective testing, rendering the scores less representative of the original HR perceptual quality. More recently, HRIQ [21] and UHD-IQA [22] provided authentically distorted images at 4K resolution. Nevertheless, these datasets generally cover mixed distortion types and lack specific focus on the perceptual characteristics of pure noise artifacts. Other UHD datasets focus on specific domains, such as portraits (PIQ23 [23]), underwater scenes (UIQD [24]), or night-time scenes (UHD-NID [25]).

As summarized in Tab. I, despite the proliferation of IQA databases, a benchmark specifically dedicated to authentic noise in UHD scenarios is conspicuously absent. Current datasets lack resolution or conflate noise with other distortions, while existing denoising datasets, though containing real noise, are ill-suited for quality assessment due to their limited scene diversity.

B. BIQA Methods

1) *General BIQA Methods*: Early BIQA methods mapped handcrafted statistical features to quality scores [26], [27], [28], [29]. However, these approaches struggle to characterize the complex, non-stationary, and signal-dependent nature of real-world noise, limiting their performance in realistic scenarios.

TABLE I
COMPARISON OF EXISTING IQA DATASETS. HR DENOTES
HIGH-RESOLUTION. RESOLUTIONS FOR HR DATASETS
REPRESENT TYPICAL VALUES

Dataset	Source / Distorted Images	Image Resolution (px)	Target Scenario
LIVE [16]	29 / 779	768 × 512	Synthetic
CSIQ [17]	30 / 866	512 × 512	Synthetic
TID2013 [4]	25 / 3000	512 × 384	Synthetic
KADID-10k [5]	81 / 10125	512 × 384	Synthetic
CLIVE [6]	1162	500 × 500	In-the-wild
KonIQ-10k [7]	10073	1024 × 768	In-the-wild
PaQ-2-PaQ [18]	39810	variable	In-the-wild
CID2013 [19]	480	1600 × 1200	HR (General)
SPAQ [20]	11125	3016 × 4032	HR (General)
HR1Q [21]	1120	2880 × 2160	HR (General)
UHD-IQA [22]	6073	3840 × 2160	HR (General)
PIQ23 [23]	5116	1900 × 2400	HR (Portrait)
UIQD [24]	5369	3840 × 2160	HR (Underwater)
UHD-NID [25]	1600	5616 × 3744	HR (Night-time)

To overcome the limited representation capability of hand-crafted features, researchers have increasingly shifted towards data-driven approaches. With the rapid advancement of deep learning, CNN-based BIQA methods have achieved significant breakthroughs [30], [31], [32], [33]. In recent years, Transformer [34] architectures have been introduced into BIQA tasks. For instance, AHIQ [35] combined Vision Transformer (ViT) and CNN to capture both global and local features. Similarly, MANIQA [36] applied ViT to image patches, while CFANet [37] demonstrated the performance gains of replacing traditional CNN backbones with a Swin Transformer. LoDA [38] efficiently adapts a frozen, large-scale pre-trained ViT by extracting multi-scale local distortion features with a CNN and injecting them into ViT tokens through cross-attention.

To better simulate HVS perception, researchers have explored hybrid spatial-frequency representations. MSGW [39] applied gradient filters to wavelet approximation subbands to capture different types of distortions. LDNet [40] integrated degradation priors and semantic features through parallel DPLMs. Yang et al. [41] utilized wavelet and curvelet transforms to capture distortions.

Recent studies have also improved BIQA from the perspectives of multimodal reasoning and data efficiency. DeQA [42] bridges continuous human quality ratings and the discrete outputs of multimodal large language models by discretizing score distributions into soft labels. In contrast to designing a new quality predictor, PGCS [43] is an IQA-oriented coreset selection method. It partitions a multi-task perceptual latent space and adaptively samples within each partition to preserve the coverage and local densities of distortion, quality, and semantic attributes, thereby reducing training-data redundancy and annotation noise.

Despite recent progress, applying general BIQA methods to HR images remains computationally expensive and prone to performance degradation [44].

2) *BIQA Methods for HR Images*: To address the limitations of general methods and meet the demands of modern

UHD photography, researchers have shifted their attention toward quality prediction specifically for HR scenarios. For instance, Korhonen et al. [45] employed RNNs for patch-level spatial pooling, while Liao et al. [46] introduced compression sampling to handle HR images. MobileIQA [47] optimized input resolution to preserve fine details. LAR-IQA [48] adopted a dual-branch design: the authenticity branch processes downsampled images, while the synthetic branch processes cropped patches. UIQA [49] evaluated UHD image quality from three complementary perspectives, and achieved first place in the AIM 2024 UHD-IQA Challenge [50]. SAMA [51] improves efficient high-resolution quality assessment through a scaling-and-masking strategy that preserves multi-scale local details and global semantics within a fixed-size input. GLIA [52] likewise addresses detail loss during input resizing through a dual-stream design: a resized-image stream retains global semantics, while grid-sampled fragments from the original-resolution image preserve local distortions; a lightweight adapter then enables bidirectional global-local interaction. However, these methods have not been optimized for real-world noise assessment and lack specific modeling of HVS mechanisms related to noise perception.

3) *BIQA Methods for Noise Distortion*: Given the limitations of general models, some research has focused explicitly on noise artifacts. Zhai et al. [12] introduced a dual-model framework combining noise estimation with active inference to handle varying noise intensities. Yang et al. [13] proposed FMIQA based on frequency mapping, while Osust [53] proposed a noise-oriented BIQA approach based on local gradient features and a visual saliency model. Lu [14] developed a data-driven denoising quality prediction method that fuses existing IQA metrics with features extracted from denoising networks.

Unfortunately, due to the lack of dedicated real-world noise datasets, these methods are typically validated on synthetic noise subsets. This limitation restricts their practical applicability and prevents a systematic understanding of how noise in real-world imaging scenarios affects human perceptual quality.

III. DATASET CREATION

A. Data Collection

Image noise is influenced by shooting conditions, environmental lighting, imaging devices, ISP algorithms, and subsequent encoding processes [54]. To better match the constructed dataset with human visual perception, we take the final RGB images, which are directly presented to end users, as the basis for data collection. The UNIQ-5k dataset contains 4,759 images derived from four complementary components.

Part I of the dataset comprises images exclusively captured by our team using various imaging devices, including professional DSLR cameras, flagship smartphones, and entry-level smartphones, with the specific device distribution detailed in Fig. 2 (a). Part II augments the dataset scale with images sourced from the public domain, primarily from Pixabay.com. To ensure the authenticity of the images and expand the variety of shooting equipment, we maintain strict selection criteria, requiring images to be authentically captured photographs with explicit shooting equipment metadata and a minimum

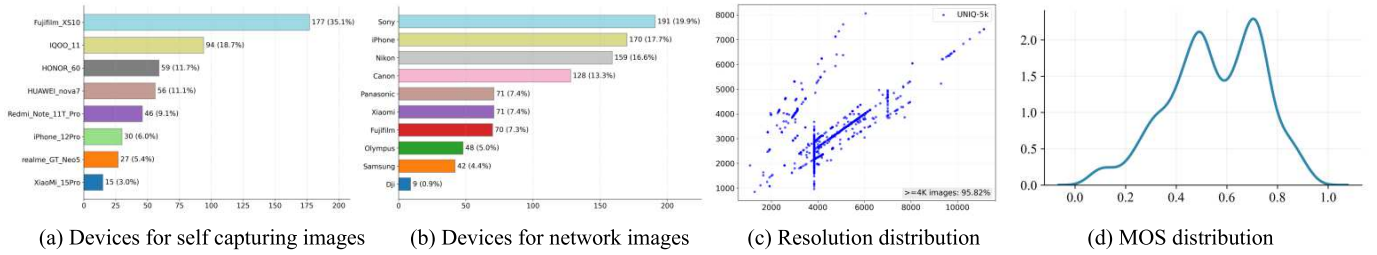


Fig. 2. (a) The distribution of specific camera models used to capture the self-collected portion of the UNIQ-5k dataset. (b) The distribution of labeled shooting equipment for the images acquired from the network. (c) A scatter plot showing the distribution of image resolutions, where the x-axis represents the width and the y-axis represents the height, highlighting the UHD nature of the UNIQ-5k dataset. (d) The distribution of MOS, where the x-axis represents the MOS and the y-axis represents the corresponding density.



(a) Representative scenes in the UNIQ-5k dataset are organized by three non-mutually exclusive scene attributes: environment (indoor (ID) / outdoor (OD)), illumination (daytime (DT) / nighttime (NT)), and texture complexity (flat-textured area (FT) / high-density textured area (HT)). Thus, one image may carry labels from multiple attributes, e.g., outdoor and daytime.



(b) Level 1: Extremely poor (c) Level 2: Poor (d) Level 3: Fair (e) Level 4: Good (f) Level 5: Excellent

Fig. 3. Overview of the UNIQ-5k dataset and the rating criteria. (a) Representative scenes contained in the dataset. (b)–(f) Visual samples corresponding to the proposed five-level subjective noise assessment scale, representing Levels 1 through 5, respectively.

resolution of 4K. The distribution of devices identified in these web-sourced images is presented in Fig. 2 (b). Finally, Parts III and IV incorporate high-quality images from the UHD-IQA dataset and the SPAQ dataset, respectively, from which we exclude samples exhibiting significant artificial degradation such as severe blurring or overexposure.

As shown in Fig. 3 (a), the entire dataset offers extensive content diversity, encompassing a wide range of scenes from various geographical locations. We characterize scene diversity along three attribute dimensions: environment, illumination condition, and texture complexity. These attributes are not mutually exclusive, and a single image may be associated with multiple attributes. In addition, the dataset includes various themes such as architecture, portraits, animals, and oceans. To ensure the authenticity of the captured image quality perceived by end-users and to preserve the original diversity stemming from various shooting devices, we do not crop or scale the collected images in any way. As shown in Fig. 2 (c), the images in our dataset cover a wide spectrum of resolutions, with over 95% of the content exceeding the UHD-1 (4K) standard. This high fidelity is essential for assessing noise.

B. Subjective Quality Marking

To establish a reliable ground truth for our dataset, we conducted a rigorous subjective quality assessment experiment

following the ITU-R guidelines [55]. The assessment was conducted under normal illumination at a viewing distance of approximately one to two screen heights. Each image was independently rated by the same 21 participants, all with normal or corrected-to-normal vision. During the evaluation, the images were displayed at their original 100% scale to prevent the loss of high-frequency noise details caused by downsampling. Given that the resolution of the test images exceeded the physical resolution of the display monitor, we implemented an interactive viewing mechanism. Participants were instructed to pan and scan across the entire image area to meticulously inspect local noise artifacts before assigning a score.

To ensure the reliability of the assessment and prevent arbitrary judgments, we incorporated a region-level annotation protocol as a validation mechanism. Specifically, upon detecting noise artifacts, participants were required to circle the affected regions and categorize the local noise intensity as “Slight”, “Noticeable”, or “Severe”. This procedure provided verifiable evidence of the participants’ attention and created an audit trail for their decision-making process. The visual examples in Fig. 3 (b)–(f) were used as calibration anchors before the experiment.

For each participant–image pair, the annotated regions are represented by three mutually exclusive severity masks

TABLE II

SUMMARY OF INTER-OBSERVER RELIABILITY AND MOS PRECISION. VALUES IN PARENTHESES ARE MEASURED ON THE NORMALIZED [0, 1] SCALE

Statistic	Value
Initial / retained observers	21 / 20
Mean pairwise SRCC	0.7346
Median individual-to-LOO-MOS SRCC	0.8426
ICC(2,1)	0.7309
ICC(2,20)	0.9819
Mean within-image standard deviation	0.4204 (0.1051)
Average standard error of MOS	0.0940 (0.0235)
Average 95% CI half-width	0.1967 (0.0491)

corresponding to Slight, Noticeable, and Severe noise. Let a_{sl} , a_{no} , and a_{se} denote the proportions of the total image area labeled as Slight, Noticeable, and Severe, respectively. For each severity class, its spatial coverage a is categorized as absent ($a = 0$), localized ($0 < a \leq 0.25$), extensive ($0.25 < a \leq 0.75$), or dominant ($a > 0.75$). The five-level image rating is determined jointly by noise severity and spatial coverage. A severity class present in the image, i.e., with $a > 0$, produces a candidate rating according to a fixed mapping rule: Slight noise yields Level 4 for localized or extensive coverage and Level 3 for dominant coverage; Noticeable noise yields Level 3 for localized coverage and Level 2 for extensive or dominant coverage; and Severe noise yields Level 2 for localized coverage and Level 1 for extensive or dominant coverage. When multiple severity classes are present, the lowest candidate level, representing the worst perceived quality, is selected as the final image-level rating. If no noise region is annotated, the image is assigned Level 5.

C. Subjective Quality Analysis

Initially, 21 subjective ratings were collected for each image to compute a preliminary MOS. To identify inconsistent observers, we computed the SRCC between each observer's ratings and the corresponding leave-one-observer-out MOS. Based on these correlations, one observer with an individual-to-LOO-MOS SRCC below 0.7 was excluded. The final MOS was then recomputed from the 20 valid ratings and normalized to the range [0, 1].

We further assessed inter-observer consistency, rating reliability, and MOS precision based on the 20 valid ratings. The mean pairwise SRCC, median individual-to-LOO-MOS SRCC, ICC(2,1), ICC(2,20), mean within-image standard deviation, average standard error, and average 95% confidence-interval half-width are summarized in Tab. II. The strong observer correlations and ICC(2,1), together with the near-unity ICC(2,20) and the small normalized confidence-interval half-width, indicate substantial inter-observer agreement and high reliability of the aggregated MOS labels.

The dataset was divided into two subsets: approximately 80% for training and 20% for validation. The MOS distribution is illustrated in Fig. 2 (d). Statistically, the scores exhibit a bimodal distribution with two dominant peaks concentrated approximately at 0.5 and 0.7. Most samples are concentrated within the central range, whereas relatively few samples occur

near the two endpoints. This pattern may be associated with two factors: First, the two modes may reflect the relatively frequent occurrence of medium- and moderately high-quality images in the collected real-world data. Second, subjective ratings may be affected by central tendency bias, also referred to as end-point aversion. Assigning extreme quality levels generally requires stronger and more unambiguous perceptual evidence, which may lead to a greater concentration of ratings in the intermediate range. While UNIQ-5k covers diverse scene content, imaging devices, and noise conditions, its distribution may retain moderate imbalances across scene categories and device models, with relatively fewer samples at extreme noise levels.

IV. PROPOSED METHOD

The overall framework of PINE is shown in Fig. 4. It consists of four parts: image preprocessing stage, frequency-domain feature extraction stage, semantic feature extraction stage, and feature fusion and regression stage. The pipeline starts with the image preprocessing stage, where the adopted Grid-based Fragment Sampling Strategy (GFSS) transforms UHD inputs into structurally preserved patches. The core feature extraction process adopts a coarse-to-fine dual-pathway architecture. One pathway focuses on low-level distortion perception: it utilizes the Perceptual Spatial Attention Module (PSAM) to construct a visibility map that prioritizes perceptually significant noise, which then modulates the Perceptual Frequency Enhancement Module (PFEM) to enhance medium-to-high-frequency details often suppressed by resizing. The other pathway focuses on high-level semantic abstraction via the Swin Transformer, ensuring the model understands the underlying image content. By effectively fusing these multi-scale frequency features with deep semantic embeddings, our model enables accurate image-quality prediction. The inference procedure of PINE is summarized in Algorithm 1.

A. Image Preprocessing Stage

The high and variable image resolutions in UNIQ-5k pose a substantial computational challenge for IQA models. Processing full-resolution images directly imposes a high computational burden, and inconsistent image resolutions complicate the determination of a suitable batch size. Consequently, an important issue arises: how to convert UHD images into compact representations that can effectively represent the quality of the original image? Motivated by the observation that signal-dependent camera-noise statistics are approximately homogeneous within small regions of slowly varying intensity [56], [57], we employ a grid-based fragment sampling strategy similar to that of FAST-VQA [15] for UHD image noise assessment, referred to as GFSS in our framework. By sampling fragments on a uniform grid, this strategy extracts representative local noise features while maintaining the global semantic topology. Unlike random cropping, which disrupts spatial context, this strategy preserves the relative spatial arrangement of the sampled fragments, ensuring that the global structural layout is effectively retained for downstream processing.

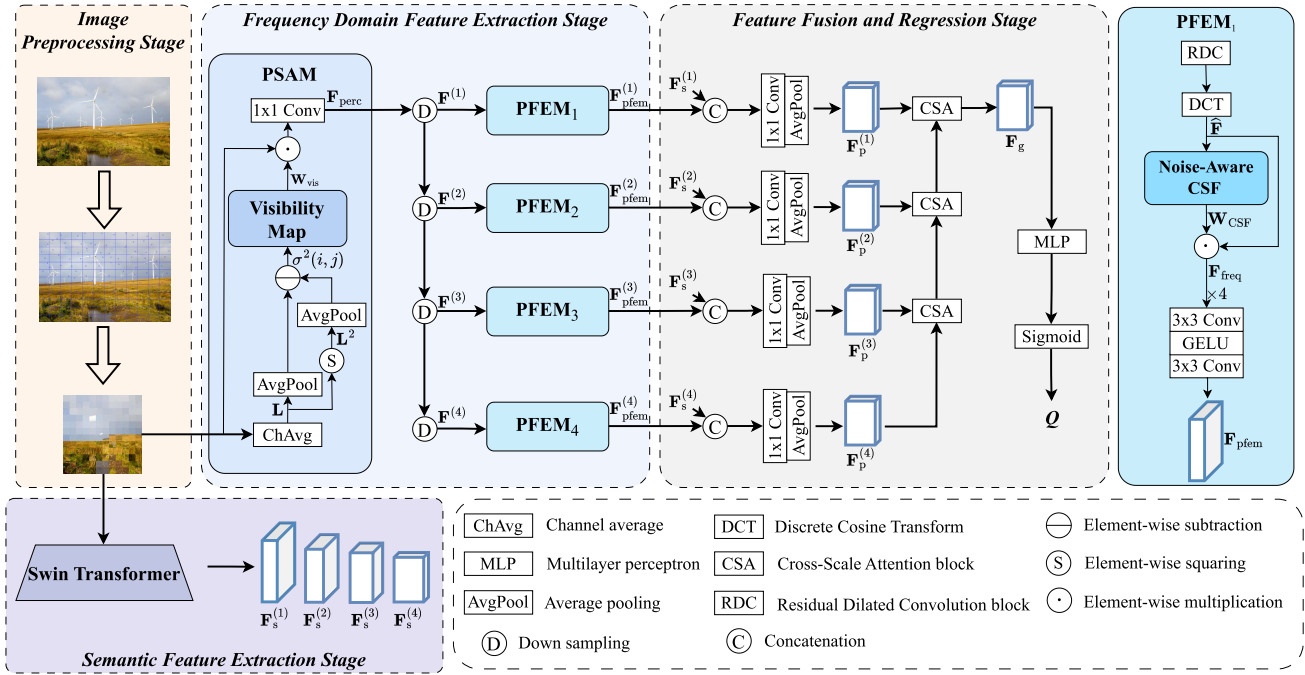


Fig. 4. Architecture overview of the proposed PINE. After image preprocessing, the frequency branch employs the Perceptual Spatial Attention Module (PSAM) and the Perceptual Frequency Enhancement Module (PFEM) to enhance noise-related details, while the semantic branch leverages a Swin Transformer for content understanding. These features are fused to predict the final quality score.

Algorithm 1 Forward Procedure of PINE

Require: UHD image x ; grid parameters N and f ; spatial-attention parameters w and D ; frequency weight matrix $\mathbf{W}_{\text{CSF}}$.

Ensure: Predicted quality score Q .

- 1: Divide x into $N \times N$ non-overlapping cells and sample one $f \times f$ fragment from each cell.
- 2: Concatenate the fragments according to their grid positions to obtain $x_f \in \mathbb{R}^{Nf \times Nf \times 3}$.
- 3: Compute the luminance map $\mathbf{L}$ and local variance σ^2 using an average-pooling window of size $w \times w$.
- 4: Generate the visibility map $\mathbf{W}_{\text{vis}} = 1/(1 + \theta\sigma^2)$, where $\theta = D/(\max(\sigma^2) + \epsilon)$.
- 5: Obtain spatially weighted features $\mathbf{F}_{\text{perc}} = \text{Conv}_{1 \times 1}(x_f \odot \mathbf{W}_{\text{vis}})$.
- 6: Apply Residual Dilation Convolution (RDC), block-wise Discrete Cosine Transform (DCT), modified CSF weighting, and downsampling to obtain $\{\mathbf{F}_{\text{pfem}}^{(i)}\}_{i=1}^4$.
- 7: Extract multi-scale semantic features $\{\mathbf{F}_s^{(i)}\}_{i=1}^4$ from the Swin Transformer branch.
- 8: Fuse $\{\mathbf{F}_{\text{pfem}}^{(i)}\}_{i=1}^4$ and $\{\mathbf{F}_s^{(i)}\}_{i=1}^4$ with 1×1 convolutions, adaptive pooling, and Cross-Scale Attention (CSA) blocks.
- 9: Regress the fused representation with a Multilayer Perceptron (MLP) followed by a sigmoid activation to obtain Q .

Specifically, given a UHD image $x \in \mathbb{R}^{H \times W \times 3}$, the image x is first uniformly divided into $N \times N$ non-overlapping patches. To reduce the spatial dimension while preserving the

pristine noise characteristics, we perform a crop operation within each patch instead of resizing. For each patch, a sub-patch $P_{i,j}$ of size $f \times f$ is extracted using a random offset. Finally, all cropped sub-patches $P_{i,j}$ are concatenated according to their original grid positions (i, j) to reconstruct a new structurally preserved fragment image $x_f \in \mathbb{R}^{H' \times W' \times 3}$, where $H' = W' = N \cdot f$. We set $N = 12$ and $f = 32$, resulting in a 384×384 input that matches the input resolution of the Swin Transformer-Base branch without additional resizing.

B. Frequency-Domain Feature Extraction Stage

Since image noise often manifests as signal interference in the frequency domain [40], modeling in the frequency domain can yield a more accurate characterization of image noise degradation. According to [58] and [59], however, the HVS exhibits non-uniform sensitivity across varying domains. In the spatial domain, noise visibility is governed by the texture masking effect, where artifacts are perceptually amplified in smooth regions but suppressed in textured areas. In the frequency domain, the Contrast Sensitivity Function (CSF) indicates that human sensitivity peaks at medium frequencies and decays at high frequencies. This biological decay poses a conflict for noise assessment, as noise energy is predominantly concentrated in the high-frequency band [41]. To reconcile biological plausibility with the physical characteristics of noise, we propose a dual-perspective approach: the Perceptual Spatial Attention Module (PSAM) captures spatially adaptive texture masking, while the Perceptual Frequency Enhancement Module (PFEM) employs a Modified Noise-Aware CSF strategy to explicitly attend to medium-to-high-frequency noise

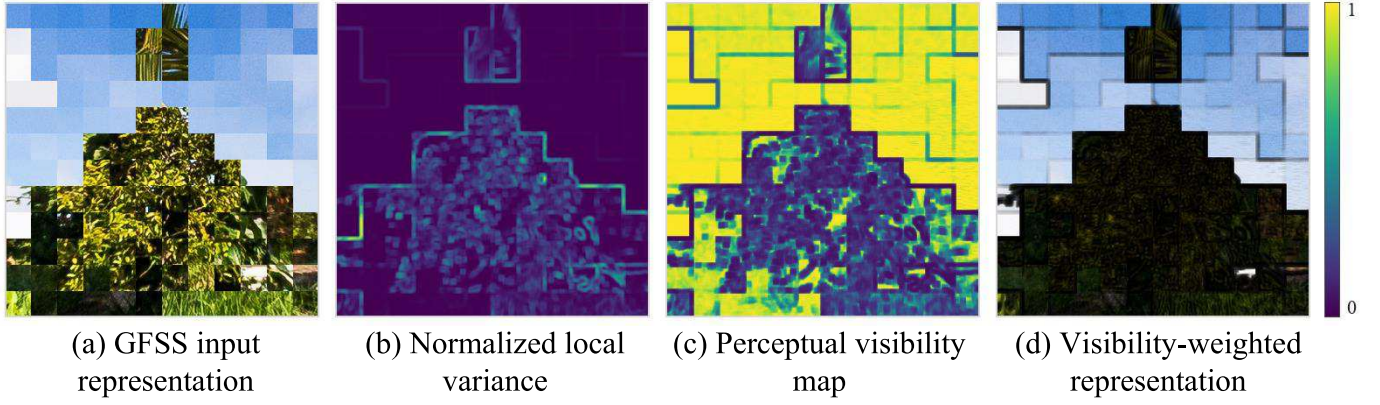


Fig. 5. Visualization of the Perceptual Spatial Attention Module (PSAM) on the GFSS representation. (a) The GFSS input representation. (b) The normalized local variance map, where brighter responses indicate stronger local texture activity or structural discontinuities. (c) The perceptual visibility map generated from the local variance according to the NVF formulation. (d) The visibility-weighted representation after spatial modulation. The visibility map assigns larger weights to smooth regions and smaller weights to textured regions, object boundaries, and GFSS-induced fragment discontinuities, reflecting perceptual visibility under texture masking.

components, ensuring that fine-grained artifacts are not attenuated by standard perceptual models.

1) *Perceptual Spatial Attention Module*: The PSAM functions as a spatial attention mechanism to prioritize perceptually significant regions. First, the structure-preserving fragment input x_f is converted into a luminance map $\mathbf{L}$ by averaging the RGB channels. We then estimate local texture activity using local variance statistics. Let $\mu(i, j)$ and $\sigma^2(i, j)$ denote the local mean and variance computed within a $w \times w$ window, where $w = 7$.

$$\begin{aligned} \mu(i, j) &= \text{AvgPool}_w(\mathbf{L}), \\ \sigma^2(i, j) &= \text{AvgPool}_w(\mathbf{L}^2) - \mu(i, j)^2. \end{aligned} \quad (1)$$

Based on the texture-masking effect, we derive the noise visibility map $\mathbf{W}_{\text{vis}}$. High values in $\mathbf{W}_{\text{vis}}$ correspond to smooth regions with high noise sensitivity, whereas low values indicate textured areas with stronger masking. The weighting function is formulated as:

$$\mathbf{W}_{\text{vis}}(i, j) = \frac{1}{1 + \theta \cdot \sigma^2(i, j)}, \quad (2)$$

where θ serves as a sensitivity tuning parameter, adaptively defined by the dynamic range of the patch:

$$\theta = \frac{D}{\max(\sigma^2) + \epsilon}, \quad (3)$$

where $D \in (0, 100]$ is a scaling constant, with $D = 75$ used by default, and ϵ ensures numerical stability.

Finally, the input x_f is modulated by this visibility map and projected into a high-dimensional feature space:

$$\mathbf{F}_{\text{perc}} = \text{Conv}_{1 \times 1}(x_f \odot \mathbf{W}_{\text{vis}}), \quad (4)$$

where $\odot$ denotes element-wise multiplication, generating the spatially weighted feature $\mathbf{F}_{\text{perc}} \in \mathbb{R}^{H' \times W' \times C}$.

As illustrated in Fig. 5, PSAM yields a spatial weighting pattern that is consistent with the NVF formulation on the GFSS representation. The normalized local variance is high in regions containing dense vegetation, palm textures, object boundaries, and GFSS-induced fragment discontinuities, while

smooth sky regions exhibit relatively low responses. Following the inverse relationship in Eq. 2, these low-variance regions receive higher perceptual visibility weights, whereas texture-dominated and edge-dominated regions are assigned lower weights due to stronger masking effects. The resulting visibility-weighted representation demonstrates that PSAM modulates the input according to perceptual distortion visibility, preserving smooth and sensitive regions while suppressing highly textured content before feature projection.

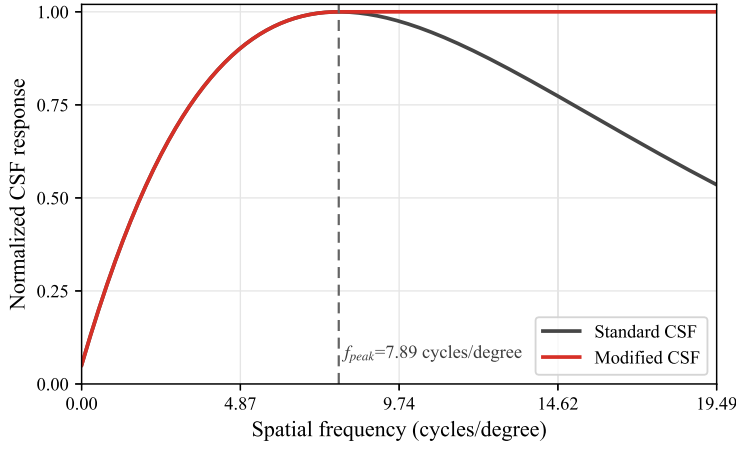
2) *Perceptual Frequency Enhancement Module*: To capture noise patterns in the frequency domain, the PFEM first employs a Residual Dilation Convolution (RDC) block [40] to expand the receptive field. It then applies a $b \times b$ block-wise Discrete Cosine Transform (DCT), where $b = 8$, to obtain the frequency representation $\hat{\mathbf{F}}$. To incorporate perceptually motivated frequency weighting without attenuating noise-related high-frequency information, we introduce a high-frequency-preserving Contrast Sensitivity Function (CSF). Conventional CSF models, such as the Mannos-Sakrison model [60], exhibit reduced sensitivity at high spatial frequencies and may therefore weaken noise-related components. We retain the original CSF response up to its peak frequency and fix the response at the peak value for higher frequencies.

To construct the frequency-weighting matrix, we first map each DCT coefficient index to its corresponding spatial frequency. For a $b \times b$ DCT block, the radial spatial frequency in cycles/pixel at coefficient position (u, v) is defined as:

$$f_r(u, v) = \sqrt{\left(\frac{u}{2b}\right)^2 + \left(\frac{v}{2b}\right)^2}, \quad u, v \in \{0, \dots, b-1\}. \quad (5)$$

The factor $2b$ arises from the spatial frequencies of the DCT-II basis functions, ensuring consistent frequency mapping for different DCT block sizes. The radial frequency is then converted from cycles/pixel to cycles/degree using a pixels-per-degree factor: $f_\theta = f_r \times \text{PPD}$. We use $\text{PPD} = 30$ as a nominal model parameter and evaluate its sensitivity at 24, 30, and 36 pixels per degree.

The standard M-S CSF is defined as $S(f) = 2.6(0.0192 + 0.114f)e^{-(0.114f)^{1.1}}$. Let f_{peak} denote the frequency where $S(f)$



(a) Standard and modified CSF responses.

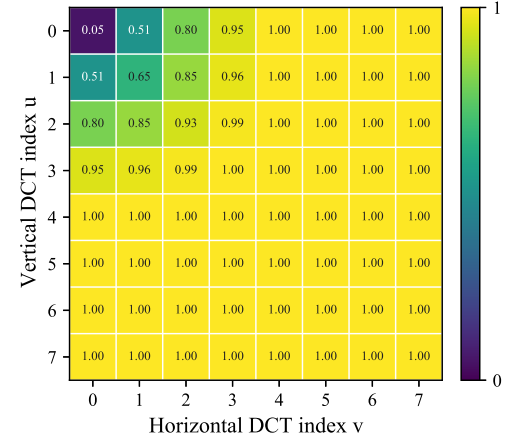
(b) Normalized 8×8 DCT weighting matrix induced by the modified CSF.

Fig. 6. Modified Noise-Aware CSF and its corresponding DCT weighting matrix. The modified response follows the standard M-S CSF up to its peak and remains constant at higher frequencies.

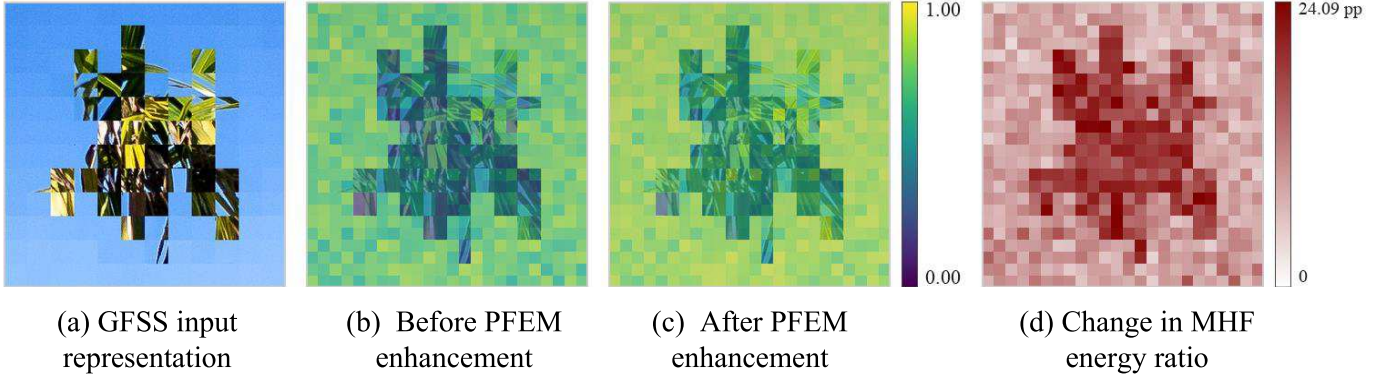


Fig. 7. Visualization of PFEM on the GFSS representation. (a) GFSS input. (b) and (c) Local medium-to-high-frequency (MHF) energy ratios before and after PFEM frequency weighting, displayed using the same $[0, 1]$ color scale. (d) Change in the local MHF energy ratio, measured in percentage points, where darker regions indicate a larger increase.

is maximized. We define our Modified Noise-Aware CSF, $\tilde{S}(f)$, as a piecewise function:

$$\tilde{S}(f_\theta) = \begin{cases} S(f_\theta), & \text{if } f_\theta \leq f_{peak} \\ S(f_{peak}), & \text{if } f_\theta > f_{peak} \end{cases} \quad (6)$$

As shown in Fig. 6, the modified response follows the standard CSF up to $f_{peak} = 7.89$ cycles/degree and remains constant thereafter. The resulting normalized weighting matrix assigns lower weights to DC and low-frequency coefficients while preserving medium-to-high-frequency coefficients.

The weighting matrix $\mathbf{W}_{CSF}$ is constructed by normalizing $\tilde{S}(f_\theta)$ by its maximum response, yielding weights in $[0, 1]$. The enhanced frequency features are then obtained via:

$$\mathbf{F}_{freq} = \hat{\mathbf{F}} \odot \mathbf{W}_{CSF}. \quad (7)$$

Finally, to match the spatial resolution of the semantic branch, $\mathbf{F}_{freq}$ undergoes a convolution block with downsampling to yield the final output $\mathbf{F}_{pfem}$.

As shown in Fig. 7, PFEM redistributes the spectral emphasis toward medium-to-high-frequency (MHF) components through spatially shared, frequency-dependent DCT

weighting. Figs. 7(b) and (c), visualized using the same normalized color scale, show the local MHF energy ratios before and after PFEM frequency weighting, respectively. The corresponding change in the local MHF energy ratio is shown in Fig. 7(d), where darker regions indicate a larger increase in percentage points. Larger changes are observed in structurally detailed regions, whereas smoother regions exhibit smaller changes. Since the PFEM parameters are shared across spatial locations, these variations primarily reflect differences in local spectral composition. The visualization is consistent with the intended role of PFEM in preventing excessive attenuation of MHF components and increasing their relative contribution. Together with Fig. 5, these results illustrate the complementary functions of the two modules: PSAM models spatial variations in noise visibility under texture masking, whereas PFEM performs frequency-selective weighting of potentially noise-related components.

C. Semantic Feature Extraction Stage

While frequency-domain analysis effectively characterizes the physical intensity of noise, the subjective perception of

image quality is governed by a complex interplay between signal degradation and semantic context. Noise perceptibility is not uniform across the image; it is modulated by visual saliency and content importance. For instance, the HVS directs preferential attention to key semantic regions (e.g., faces or salient objects). Consequently, noise artifacts located in these salient areas are more perceptually salient and detrimental to quality scores than those in non-salient backgrounds, even if their physical amplitudes are identical. Therefore, relying solely on low-level frequency features is insufficient to capture this content-dependent quality degradation. To bridge this gap, we introduce the Swin Transformer [61] to extract semantic features, leveraging its hierarchical architecture to model long-range dependencies.

Specifically, we employ a Swin Transformer-Base model pre-trained on ImageNet-22K [62] to extract multi-scale semantic features. Let $\mathcal{F}_{\text{swin}}(\cdot)$ denote the backbone mapping function. The extraction process is formulated as:

$$\{\mathbf{F}_s^{(1)}, \mathbf{F}_s^{(2)}, \mathbf{F}_s^{(3)}, \mathbf{F}_s^{(4)}\} = \mathcal{F}_{\text{swin}}(x_f), \quad (8)$$

where $\mathbf{F}_s^{(i)}$ represents the semantic feature map output from the i -th stage of the Swin Transformer, capturing contextual information from coarse to fine scales.

D. Feature Fusion and Regression Stage

To leverage the complementary nature of low-level visual artifacts and high-level semantic content, we design a hierarchical feature fusion and regression stage.

Specifically, at each scale i , we first concatenate the frequency-domain feature with the corresponding semantic feature along the channel dimension. The combined feature is then passed through a 1×1 convolution layer and an adaptive average-pooling layer to produce the preliminary fused features, denoted as $\{\mathbf{F}_p^{(i)}\}_{i=1}^4$:

$$\mathbf{F}_p^{(i)} = \text{AvgPool}(\text{Conv}_{1 \times 1}(\text{Concat}(\mathbf{F}_{\text{pfem}}^{(i)}, \mathbf{F}_s^{(i)}))). \quad (9)$$

To further enhance cross-level information interaction, we introduce the Cross-Scale Attention (CSA) block [37], which progressively fuses high-level semantic context into lower-level representations in a top-down manner, yielding the final global representation $\mathbf{F}_g$:

$$\mathbf{F}_g = \text{CSA}(\mathbf{F}_p^{(1)}, \text{CSA}(\mathbf{F}_p^{(2)}, \text{CSA}(\mathbf{F}_p^{(3)}, \mathbf{F}_p^{(4)}))). \quad (10)$$

Finally, $\mathbf{F}_g$ is mapped to the predicted quality score Q through a Multilayer Perceptron (MLP) followed by a Sigmoid activation function.

V. EXPERIMENTS

A. Implementation Details

1) *Datasets*: In the experimental section, comprehensive evaluations are primarily conducted on the proposed UNIQ-5k dataset to establish the baseline performance of the model. To further verify the capability and accuracy of the proposed PINE as a specialized metric for image denoising tasks, we extend our validation to the SIDD Benchmark [63]. Specifically, we apply various denoising algorithms to SIDD

images and utilize PINE to assess the quality scores of the denoised results. Finally, to further evaluate PINE on additional public HR-IQA benchmarks, we conduct experiments on UHD-IQA [22], SPAQ [20], and HRIQ [21].

2) *Evaluation Metrics*: For the UNIQ-5k dataset, we report Spearman's Rank-order Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), Kendall's Rank-order Correlation Coefficient (KRCC), and Root Mean Square Error (RMSE), while KRCC is omitted for the HR datasets. SRCC and KRCC measure the prediction monotonicity [64], [65], PLCC quantifies the linear correlation [66], and RMSE reflects the overall prediction accuracy [67]. Regarding the SIDD benchmark, given the absence of subjective scores, we validate our metric by analyzing the consistency between its average predicted scores and standard full-reference metrics, including PSNR and SSIM.

3) *Training Details*: Our method is implemented using the PyTorch framework and trained on a single NVIDIA GeForce RTX 3090 GPU. We employ the AdamW optimizer with an initial learning rate of 3×10^{-5} and a weight decay of 10^{-2} . The AdamW parameters are set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rate is adjusted using a cosine annealing scheduler, with the hyperparameters set to $T_{\max} = 50$ and $\eta_{\min} = 0$. The training process spans 100 epochs with a batch size of 24 and uses the mean squared error (MSE) loss. For the main comparison, all trainable models are trained using the same validation split and random seed. To assess stochastic robustness, our PINE model is additionally trained in ten independent runs with different random seeds.

B. Performance Comparison

1) *Performance Comparison on UNIQ-5k*: To conduct a comparative analysis, we benchmark the proposed method against a set of representative BIQA algorithms. These include the traditional handcrafted methods (BRISQUE [26], NIQE [27]), a wide range of deep learning-based models (DeepIQA [30], HyperIQA [31], DBCNN [32], CFANet-NR [37], LoDA [38], DP-IQA [68], and SIND [69]). Additionally, we compare with vision-language models (CLIPQA+ [70], QualiCLIP+ [71]), models specialized for HR or sampling-based quality assessment (UIQA [49], MobileIQA [47], and SAMA [51]), an IQA-oriented coreset selection method (PGCS [43]), and an MLLM-based quality-scoring model (DeQA [42]).

The quantitative results on the UNIQ-5k dataset are summarized in Tab. III. Our method achieves state-of-the-art performance across all evaluation metrics, securing an SRCC of 0.9482, a PLCC of 0.9507, a KRCC of 0.8057, and an RMSE of 0.0560. The small standard deviations demonstrate that the performance is stable with respect to random initialization. This superiority stems from our specific design for UHD noise, whereas existing baselines exhibit the following critical limitations: Handcrafted metrics (BRISQUE, NIQE) fail to capture the complex, signal-dependent nature of real-world ISP noise due to their limited feature capacity. General deep models like HyperIQA are inherently designed to extract high-level semantic features for generic quality assessment. Even when retrained on noise datasets, their backbone networks tend

TABLE III

COMPARISON OF DIFFERENT BIQA METHODS ON THE UNIQ-5k DATASET. THE MEAN AND STD. ROWS FOR PINE SUMMARIZE TEN INDEPENDENT RUNS WITH DIFFERENT RANDOM SEEDS. FOR COMPLEXITY EVALUATION, ALL METHODS USE THE SAME NATIVE-RESOLUTION UHD IMAGES WITH THEIR DEFAULT INFERENCE PREPROCESSING; TIME AND PEAK MEMORY ARE MEASURED WITH BATCH SIZE 1 IN FP32 ON A SINGLE NVIDIA RTX 3090 GPU. THE BEST AND SECOND-BEST RESULTS ARE **BOLD** AND UNDERLINED, RESPECTIVELY

Method	UNIQ-5k				Computational complexity			
	SRCC $\uparrow$	PLCC $\uparrow$	KRCC $\uparrow$	RMSE $\downarrow$	Params (M) $\downarrow$	FLOPs (G) $\downarrow$	Time (s) $\downarrow$	Peak memory (MiB) $\downarrow$
BRISQUE [26]	0.1441	0.0919	0.0988	0.2941	—	—	0.1967	—
NIQE [27]	0.3741	0.2050	0.2566	0.3842	—	—	0.2251	—
DeepIQA [30]	0.4951	0.4845	0.3435	0.1570	5.23	4487.82	0.5013	10456.00
HyperIQA [31]	0.8382	0.8130	0.6413	0.1138	27.37	429.07	0.1397	897.27
DBCNN [32]	0.7524	0.7680	0.5635	0.1749	<u>15.31</u>	193.86	0.3386	<u>179.25</u>
CFANet-NR [37]	0.7182	0.7320	0.5333	0.2251	45.19	80.75	0.3582	249.72
DP-IQA [68]	0.8134	0.8410	0.6282	0.0971	1064.67	2088.31	0.3389	4743.83
SIND [69]	0.7678	0.7927	0.5775	0.1094	149.62	45.12	0.3324	1373.32
CLIPQA+ [70]	0.6641	0.7196	0.4848	0.2285	102.02	23.84	<u>0.1184</u>	413.58
QualiCLIP+ [71]	0.6347	0.6984	0.4601	0.1637	102.12	<u>93.67</u>	0.2959	429.12
MobileIQA [47]	0.8050	0.8284	0.6358	0.1007	96.72	121.94	0.1778	697.68
UIQA [49]	0.8854	0.8929	0.7087	0.0855	82.85	81.82	0.2936	409.29
LoDA [38]	0.9132	0.9185	0.7461	0.0707	118.23	54.75	0.2597	694.00
SAMA [51]	<u>0.9185</u>	<u>0.9248</u>	<u>0.7549</u>	<u>0.0694</u>	28.40	8.71	0.3922	141.36
PGCS [43]	0.8754	0.8838	0.6965	0.0840	135.74	305.31	0.2228	644.33
DeQA [42]	0.8320	0.8518	0.6488	0.0936	8197.86	1982.47	0.6215	15788.32
PINE (Ours)	0.9482	0.9507	0.8057	0.0560	153.10	158.51	0.1056	752.64
PINE (Mean)	0.9479	0.9503	0.8055	0.0562	—	—	—	—
PINE (Std.)	0.0010	0.0010	0.0017	0.0004	—	—	—	—

TABLE IV

AVERAGE QUALITY SCORES OF TRADITIONAL DENOISING ALGORITHMS ON THE SIDD BENCHMARK EVALUATED BY VARIOUS IQA METRICS. THE LEARNING-BASED BIQA BASELINES ARE EVALUATED USING UNIQ-5K-RETRAINED WEIGHTS FOR A FAIR COMPARISON. FOR EACH METRIC (ROW), THE BEST AND SECOND-BEST DENOISING RESULTS ARE MARKED IN **BOLD** AND UNDERLINED, RESPECTIVELY

Type	Method	Denoising Methods				Range
		Bilateral Filter [72]	NLM [73]	BM3D [74]	Guided Filter [75]	
Traditional	PSNR $\uparrow$	31.7982	29.9924	25.2270	25.5452	(0, $+\infty$)
	SSIM [76] $\uparrow$	0.8930	<u>0.7968</u>	0.6154	0.5591	[0, 1]
BIQA	BRISQUE [26] $\downarrow$	49.6651	13.3155	<u>26.3604</u>	31.3616	[0, 150]
	NIQE [27] $\downarrow$	10.5844	8.0310	10.3853	<u>9.7406</u>	[0, 100]
	NRQM [77] $\uparrow$	2.5304	<u>4.6491</u>	3.0314	5.7279	[0, 10]
	DeepIQA [30] $\uparrow$	0.4604	0.4698	0.5099	<u>0.4752</u>	[0, 1]
	HyperIQA [31] $\uparrow$	0.4433	0.4569	0.4262	<u>0.4511</u>	[0, 1]
	DBCNN [32] $\uparrow$	0.5178	<u>0.3992</u>	0.3121	0.3509	[0, 1]
	CLIPQA+ [70] $\uparrow$	0.4629	0.5379	<u>0.4722</u>	0.4593	[0, 1]
	QualiCLIP+ [71] $\uparrow$	0.4692	0.4617	0.4362	0.4218	[0, 1]
	CFANet-NR [37] $\uparrow$	<u>0.3904</u>	0.5569	0.3755	0.3337	[0, 1]
	LoDA [38] $\uparrow$	0.2462	<u>0.2408</u>	0.2263	0.1845	[0, 1]
	SAMA [51] $\uparrow$	<u>0.2684</u>	0.3387	0.2095	0.2097	[0, 1]
	PINE (Ours)$\uparrow$	0.3526	<u>0.3516</u>	0.2880	0.2890	[0, 1]

to conflate fine-grained noise artifacts with image textures due to the lack of explicit frequency-aware mechanisms. Consequently, they struggle to decouple noise visibility from complex image content. Even HR-specialized models struggle. MobileIQA necessitates downsampling due to resolution caps, inadvertently acting as a low-pass filter that suppresses key noise artifacts. Similarly, while SAMA and UIQA handle high resolution effectively, their lack of noise-specific inductive bias limits their discriminability for fine-grained noise patterns. By addressing these bottlenecks via the adopted GFSS and dual-branch perception mechanism, our model achieves robust and accurate assessment.

2) *Computational Complexity*: The computational comparison is reported in Tab. III. PINE contains 153.10 M parameters and requires 158.51 GFLOPs for one GFSS-processed input. Its parameter count is lower than those of DP-IQA and DeQA and is comparable to that of SIND, but is higher than those of the remaining learning-based baselines. A substantial portion of this model size arises from the Swin Transformer-Base semantic backbone. We retain this backbone to preserve its hierarchical semantic modeling capability, since the backbone ablation in Tab. VII shows that replacing it with lighter CNN alternatives reduces prediction accuracy. Thus, the current design accepts a larger parameter count in exchange

TABLE V

AVERAGE QUALITY SCORES OF DEEP LEARNING-BASED DENOISING ALGORITHMS ON THE SIDD BENCHMARK EVALUATED BY VARIOUS IQA METRICS. THE LEARNING-BASED BIQA BASELINES ARE EVALUATED USING UNIQ-5K-RETRAINED WEIGHTS FOR A FAIR COMPARISON. FOR EACH METRIC (ROW), THE BEST AND SECOND-BEST DENOISING RESULTS ARE MARKED IN **BOLD** AND UNDERLINED, RESPECTIVELY

Type	Method	Denoising Methods						Range
		DnCNN [78] (TIP'17)	CBDNet [79] (CVPR'19)	DeamNet [80] (CVPR'21)	Restormer [81] (CVPR'22)	Xformer [82] (ICLR'24)	MaIR [83] (CVPR'25)	
Traditional	PSNR↑	26.2086	38.2967	39.3951	39.9282	<u>39.8905</u>	39.8304	(0, +∞)
	SSIM [76]↑	0.4804	0.9343	0.9430	0.9469	0.9469	<u>0.9465</u>	[0, 1]
BIQA	BRISQUE [26]↓	17.3598	<u>48.3807</u>	62.9313	64.1872	63.9696	65.0216	[0, 150]
	NIQE [27]↓	8.9350	<u>13.0647</u>	15.5945	16.8116	16.9284	16.9389	[0, 100]
	NRQM [77]↑	6.6158	<u>2.6355</u>	2.5598	2.5182	2.5133	2.5111	[0, 10]
	DeepIQA [30]↑	0.4984	0.4642	0.4666	0.4664	0.4663	0.4665	[0, 1]
	HyperIQA [31]↑	0.3909	0.5824	0.6234	0.6331	0.6339	<u>0.6334</u>	[0, 1]
	DBCNN [32]↑	0.2327	0.6366	0.6618	0.6599	<u>0.6617</u>	0.6586	[0, 1]
	CLIPQA+ [70]↑	0.4377	0.5647	0.5320	<u>0.5360</u>	0.5356	0.5356	[0, 1]
	QualiCLIP+ [71]↑	0.4488	0.4966	0.4930	0.4901	0.4902	0.4899	[0, 1]
	CFANet-NR [37]↑	0.3364	0.5983	0.7156	<u>0.7208</u>	0.7211	0.7183	[0, 1]
	LoDA [38]↑	0.1997	0.2911	0.3001	0.3048	0.3051	<u>0.3050</u>	[0, 1]
	SAMA [51]↑	0.1906	0.5487	0.6957	<u>0.7100</u>	0.7108	<u>0.7087</u>	[0, 1]
	PINE (Ours) ↑	0.1958	0.6185	0.7482	0.7533	<u>0.7532</u>	0.7531	[0, 1]

TABLE VI

COMPARISON OF DIFFERENT BIQA METHODS ON THREE PUBLIC HR DATASETS. THE BEST AND SECOND-BEST RESULTS ARE **BOLD** AND UNDERLINED, RESPECTIVELY. “—” INDICATES UNAVAILABLE RESULTS. SINCE DEQA REQUIRES BOTH THE MOS AND THE CORRESPONDING RATING STANDARD DEVIATION (STD) FOR EACH IMAGE DURING TRAINING, IT CANNOT BE TRAINED ON UHD-IQA, WHICH PROVIDES ONLY MOS LABELS. THEREFORE, ITS AVERAGE PERFORMANCE ACROSS THE THREE DATASETS IS NOT REPORTED TO ENSURE A FAIR COMPARISON

Method	UHD-IQA [22]			SPAQ [20]			HRIQ [21]			Average		
	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓
BRISQUE [26]	0.0226	0.0417	0.1530	0.5213	0.5110	0.2272	0.1410	0.0866	0.3384	0.2283	0.2131	0.2395
NIQE [27]	0.1521	0.1725	0.3599	0.6861	0.7085	0.4026	0.5921	0.5947	0.2447	0.4768	0.4919	0.3357
DBCNN [32]	0.6504	0.6542	0.0590	0.9153	0.9191	0.0856	0.8694	0.8737	0.1181	0.8117	0.8157	0.0876
HyperIQA [31]	0.4859	0.3273	0.0644	0.6814	0.6696	0.2087	0.6983	0.6939	0.1978	0.6219	0.5636	0.1570
DeepIQA [30]	0.2397	0.2374	0.0725	0.8457	0.8481	0.1189	0.5734	0.5690	0.2019	0.5529	0.5515	0.1311
CLIPQA+ [70]	0.7431	0.7324	0.1078	0.8943	0.8981	0.0920	<u>0.8975</u>	0.8909	0.1104	0.8450	0.8405	0.1034
QualiCLIP+ [71]	0.7570	0.7521	0.0786	0.9022	0.9049	0.0983	0.8802	0.8781	0.1159	0.8465	0.8450	0.0976
CFANet-NR [37]	0.6862	0.6882	0.0558	0.9165	0.9182	0.0862	0.8483	0.8568	0.1123	0.8170	0.8211	0.0848
MobileIQA [47]	0.7757	0.7831	0.0439	0.9181	0.9190	0.0806	0.8666	0.8737	0.1054	0.8535	0.8586	0.0766
LoDA [38]	0.6847	0.6871	0.0482	0.9258	0.9273	0.0770	0.8240	0.8375	0.1299	0.8115	0.8173	0.0850
SAMA [51]	0.7371	0.7448	0.0474	0.9265	0.9301	0.0777	0.7618	0.7659	0.1626	0.8084	0.8136	0.0959
UIQA [49]	0.8169	0.8238	0.0402	0.9279	0.9315	<u>0.0759</u>	0.8959	<u>0.9115</u>	0.0967	0.8802	0.8889	0.0709
DP-IQA [68]	0.6428	0.6510	0.0588	0.9219	0.9245	0.0798	0.8570	0.8442	0.1002	0.8072	0.8066	0.0796
SIND [69]	0.6021	0.6074	0.0549	0.9177	0.9210	0.0811	0.8197	0.8206	0.1346	0.7798	0.7830	0.0902
PGCS [43]	0.5902	0.6029	0.0546	0.9165	0.9201	0.0813	0.8561	0.8643	0.1102	0.7876	0.7957	0.0820
DeQA [42]	—	—	—	0.9301	0.9349	0.0747	0.8319	0.8398	0.1290	—	—	—
PINE (Ours)	0.8226	0.8285	<u>0.0425</u>	<u>0.9297</u>	0.9302	0.0795	0.9125	0.9151	<u>0.0973</u>	0.8883	0.8913	<u>0.0731</u>

for stronger semantic representation and quality-prediction performance. In terms of operation count, PINE requires fewer FLOPs than DeepIQA, HyperIQA, DBCNN, and DP-IQA, while several recent methods achieve lower FLOPs. Nevertheless, PINE records the shortest measured inference time of 0.1056 s per UHD image and a peak inference memory of 752.64 MiB, while achieving the best prediction accuracy. This low latency is facilitated by GFSS, which converts each native-resolution UHD image into a single fixed-size 384×384 representation and avoids repeated full-resolution or multi-crop inference. Overall, PINE favors prediction accuracy and low inference latency at the cost of a relatively large model and substantial operation count, while maintaining moderate peak memory consumption. Developing a more compact variant without substantially degrading noise-perception performance remains an important direction for future work.

3) *Validation as a Denoising Metric*: To further evaluate the ability of PINE to serve as an objective metric under real-world noise conditions, we conduct extensive denoising performance assessments on the SIDD Benchmark [63]. As a widely recognized benchmark in the field, SIDD provides a rigorous platform for measuring the effectiveness of denoising algorithms on real-world noisy images. We employ a diverse set of denoising algorithms, including four traditional methods (Bilateral Filter [72], NLM [73], BM3D [74], Guided Filter [75]) and six deep learning-based methods (DnCNN [78], CBDNet [79], DeamNet [80], Restormer [81], Xformer [82], MaIR [83]). Using their official pre-trained models, we process the noisy images from the SIDD Benchmark to generate denoised outputs with varying quality levels.

Tab. IV and Tab. V present the average quality scores predicted by various IQA metrics for traditional and deep

TABLE VII

ABLATION RESULTS ON THE UNIQ-5k DATASET. “SEM.” DENOTES THE SEMANTIC BRANCH, “SWIN TRANS.” DENOTES THE SWIN TRANSFORMER, AND “TIME (S)” IS THE AVERAGE INFERENCE TIME FOR ONE UHD IMAGE ON A SINGLE NVIDIA RTX 3090 GPU. THE BEST AND SECOND-BEST RESULTS ARE **BOLD** AND UNDERLINED, RESPECTIVELY

Group	Setting	Modules					UNIQ-5k				
		GFSS	Sem.	PFEM	PSAM	Backbone	SRCC $\uparrow$	PLCC $\uparrow$	KRCC $\uparrow$	RMSE $\downarrow$	Time (s) $\downarrow$
Component	Semantic branch only	✓	✓	–	–	Swin Trans.	0.9069	0.9045	0.7538	0.0833	0.0256
	Frequency branch only / w/o semantic branch	✓	–	✓	✓	–	<u>0.9390</u>	<u>0.9469</u>	<u>0.7947</u>	0.0616	0.0839
	Frequency branch w/o PSAM	✓	–	✓	–	–	0.9252	0.9211	0.7815	0.0639	0.0713
	Full model w/o PSAM	✓	✓	✓	–	Swin Trans.	0.9290	0.9353	0.7776	<u>0.0587</u>	0.0906
	Full model w/o PFEM	✓	✓	–	✓	Swin Trans.	0.9160	0.9146	0.7814	0.0717	<u>0.0348</u>
	Full PINE	✓	✓	✓	✓	Swin Trans.	0.9482	0.9507	0.8057	0.0560	0.1056
Backbone	Full model with ResNet50	✓	✓	✓	✓	ResNet50	<u>0.9257</u>	<u>0.9266</u>	<u>0.7667</u>	<u>0.0679</u>	0.0986
	Full model with MobileNetV3	✓	✓	✓	✓	MobileNetV3	0.9112	0.9178	0.7439	0.0725	0.1005
	Full model with Swin Transformer	✓	✓	✓	✓	Swin Trans.	0.9482	0.9507	0.8057	0.0560	0.1056
Sampling	Resize	Resize	✓	✓	✓	Swin Trans.	0.8422	0.8616	0.6602	0.0914	0.1124
	Random Crop ($\times 25$)	Crop	✓	✓	✓	Swin Trans.	<u>0.9298</u>	<u>0.9326</u>	<u>0.7726</u>	<u>0.0667</u>	2.6275
	Adaptive Resize + Crop	Resize+Crop	✓	✓	✓	Swin Trans.	0.8437	0.8618	0.6648	0.0921	0.1046
	GFSS	✓	✓	✓	✓	Swin Trans.	0.9482	0.9507	0.8057	0.0560	<u>0.1056</u>

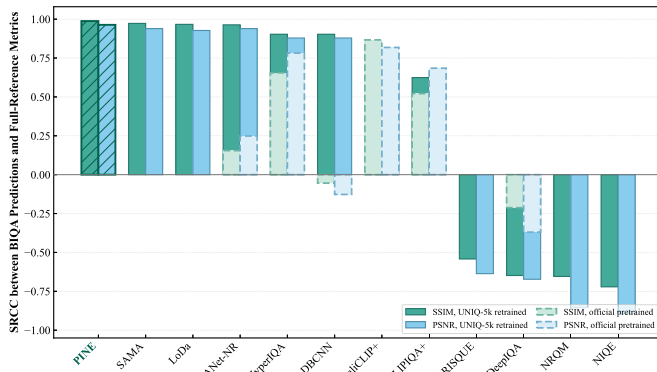


Fig. 8. SRCC between BIQA-predicted average scores and the full-reference PSNR and SSIM criteria across ten denoising algorithms on the SIDD benchmark. Solid bars denote results obtained using UNIQ-5k-retrained weights, whereas light-colored bars with dashed outlines denote results obtained using the official pre-trained weights.

learning-based denoising algorithms, respectively. For a fair comparison, all learning-based BIQA baselines in these tables are evaluated using weights retrained on UNIQ-5k, whereas the traditional metrics are evaluated in their original forms. We assess the ranking consistency between the predicted quality scores and the full-reference denoising criteria by computing the SRCC between the average BIQA scores and the PSNR/SSIM values across the ten denoising algorithms, as summarized in Fig. 8. PINE achieves the highest SSIM-based SRCC of 0.9878 and the highest PSNR-based SRCC of 0.9636. Among the retrained learning-based baselines, SAMA, LoDa, and CFANet-NR exhibit the strongest overall correlations with PSNR and SSIM, followed closely by HyperIQA. DeepIQA is a notable exception: its correlations remain negative after retraining, indicating that its predictions are poorly aligned with the denoising ranking. Traditional statistical metrics (e.g., BRISQUE, NIQE, NRQM) tend to misinterpret the smoothing effects of modern denoising as blur distortions, resulting in severe misalignment and negative correlations.

4) *Impact of the UNIQ-5k Dataset:* To further substantiate the necessity and contribution of the proposed

UNIQ-5k dataset, we investigate whether the suboptimal performance of existing BIQA methods could be attributed to the domain gap in their training data. To isolate the effect of the training domain, we compare the official pre-trained weights of the learning-based BIQA baselines with their counterparts retrained on UNIQ-5k. The results reported in Tab. IV and Tab. V use the UNIQ-5k-retrained weights for the main comparison. Many official BIQA models are trained on generic in-the-wild IQA datasets, such as KonIQ-10k [7], whose images contain diverse, coexisting distortions. Such training encourages holistic quality prediction, but does not explicitly require the model to resolve the fine-grained visibility differences caused by real-world sensor and ISP noise. This mismatch creates a domain gap for denoising-oriented assessment.

The effect of retraining is summarized in Fig. 8: solid bars denote the UNIQ-5k-retrained weights, whereas light-colored bars with dashed outlines denote the official pre-trained weights. UNIQ-5k retraining substantially improves both PSNR- and SSIM-based SRCC for CFANet-NR, HyperIQA, and DBCNN. DeepIQA is an exception: it remains strongly negatively correlated with both criteria after retraining, indicating that UNIQ-5k adaptation does not improve its denoising-method ranking in this experiment. Thus, the benefit of UNIQ-5k is model-dependent, but the overall trend indicates that noise-oriented training substantially improves the alignment of generic BIQA models with denoising quality.

The score curves in Fig. 9, where the denoising methods are ordered by ascending PSNR, provide a complementary view of this effect. For most learning-based models, retraining reduces the pronounced non-monotonic variations observed with the official weights and produces a more coherent progression across denoising methods. The remaining failure case of DeepIQA further suggests that dataset adaptation alone cannot eliminate all model-specific limitations. Overall, these results support the role of UNIQ-5k in narrowing the domain gap between generic perceptual-quality assessment and noise-oriented denoising assessment.

5) *Evaluation on Additional Public HR-IQA Benchmarks:* To further evaluate PINE beyond the noise-specific

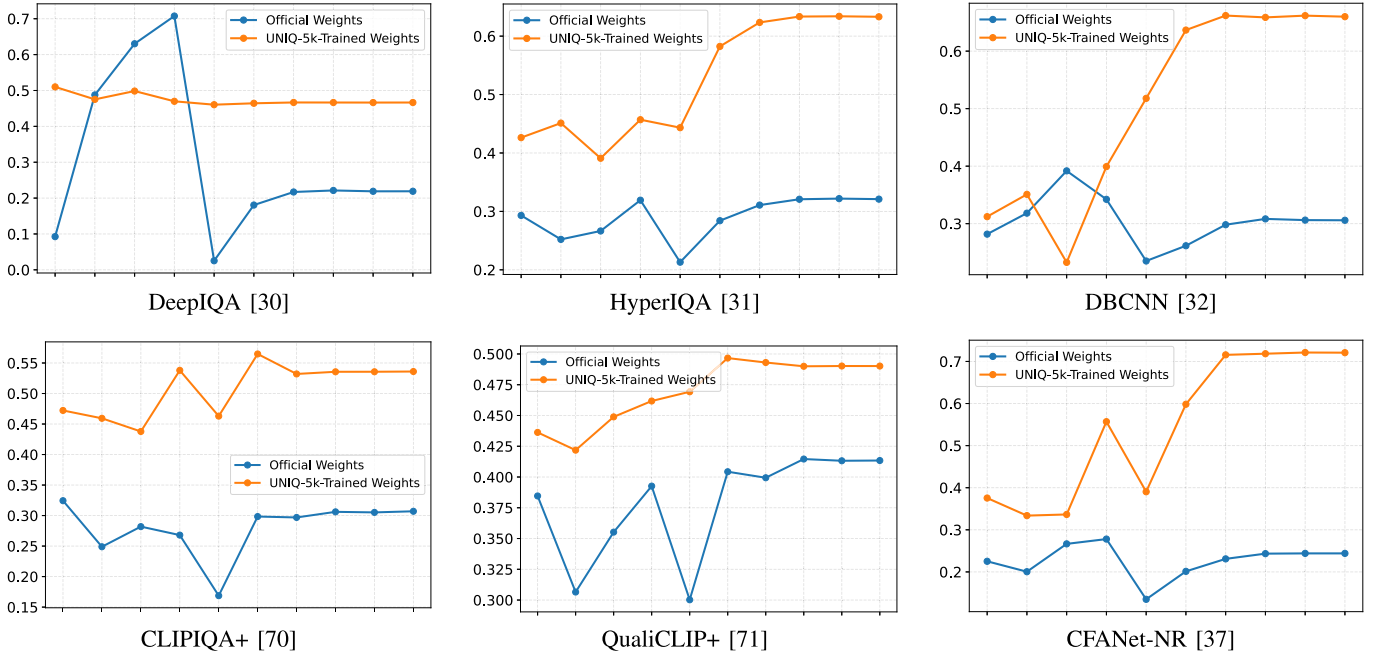


Fig. 9. Comparison of predicted quality scores on denoised images using official pre-trained weights (blue curve) and weights trained on the proposed UNIQ-5k dataset (orange curve). The x-axis represents ten denoising algorithms sorted by ascending PSNR: BM3D, Guided Filter, DnCNN, NLM, Bilateral Filter, CBDNet, DeamNet, MaIR, Xformer, and Restormer. The y-axis denotes the average quality score predicted by the BIQA model for the denoised images produced by each algorithm.

UNIQ-5k setting, we conduct experiments on three additional public HR-IQA benchmarks: UHD-IQA, SPAQ, and HRIQ. These datasets contain authentic mixed distortions and provide complementary evaluation under the selected HR-IQA settings.

As summarized in Tab. VI, PINE achieves the highest SRCC and PLCC on UHD-IQA and HRIQ. On SPAQ, PINE remains competitive, obtaining an SRCC of 0.9297. Among the methods with results reported on all three datasets, PINE obtains the highest average SRCC and PLCC, reaching 0.8883 and 0.8913, respectively.

These results indicate that the representations learned by PINE remain effective on the three evaluated public HR-IQA benchmarks. Although PINE is designed primarily for noise-oriented quality assessment, the adopted GFSS enables efficient processing of HR inputs while retaining native-resolution details within the sampled fragments and their coarse spatial organization. In addition, PSAM and PFEM enhance the model’s sensitivity to spatially varying and frequency-domain artifacts, which may contribute to its competitive performance on the evaluated mixed-distortion benchmarks. Collectively, these results suggest that PINE has applicability beyond noise-specific assessment within the evaluated HR-IQA settings.

C. Ablation Study

To verify the contribution of each design choice, we conduct four groups of controlled ablation experiments on the UNIQ-5k dataset: component analysis, including PFEM stage placement; backbone replacement; input sampling comparison; and hyperparameter sensitivity analysis. The component,

TABLE VIII

ABLATION OF PFEM STAGES ON THE UNIQ-5K DATASET. THE STAGES ARE ADDED CUMULATIVELY FROM PFEM₁ TO PFEM₄. THE BEST AND SECOND-BEST RESULTS ARE **BOLD** AND UNDERLINED, RESPECTIVELY

PFEM ₁	PFEM ₂	PFEM ₃	PFEM ₄	SRCC $\uparrow$	PLCC $\uparrow$	KRCC $\uparrow$	RMSE $\downarrow$	Time (s) $\downarrow$
✓				0.8968	0.9021	0.7669	0.0820	0.0526
✓	✓			0.9126	0.9182	0.7703	0.0713	0.0769
✓	✓	✓		<u>0.9315</u>	<u>0.9346</u>	0.7854	0.0649	0.0905
✓	✓	✓	✓	0.9482	0.9507	0.8057	0.0560	0.1056

backbone, and sampling configurations are reported in Tab. VII; the hyperparameter results are reported in Tab. IX.

1) *Component Isolation*: To investigate the contribution of each feature branch, we evaluate two model variants using only the semantic or frequency branch. The frequency-only variant outperforms the semantic-only variant, suggesting that local spectral distortions provide more direct and discriminative cues for real-world noise assessment, whereas semantic features mainly offer indirect contextual information. Nevertheless, the full model achieves the best overall performance, demonstrating the complementarity of the two branches: frequency features characterize noise-related distortions, while semantic context helps determine their perceptual relevance with respect to scene content.

We further evaluate the contributions of Perceptual Spatial Attention Module (PSAM) and Perceptual Frequency Enhancement Module (PFEM). Removing PSAM reduces performance in both the full model and the frequency-only variant, indicating that texture-aware weighting helps suppress responses in complex regions while preserving visible noise in

TABLE IX

HYPERPARAMETER SENSITIVITY ON THE UNIQ-5K VALIDATION SPLIT. EACH CONFIGURATION WAS TRAINED ONCE USING THE SAME FIXED RANDOM SEED. † DENOTES THE DEFAULT SETTING. THE BEST AND SECOND-BEST RESULTS WITHIN EACH PARAMETER GROUP ARE HIGHLIGHTED IN **BOLD** AND UNDERLINED, RESPECTIVELY

Parameter	Value	SRCC↑	PLCC↑	KRCC↑	RMSE↓
GFSS (N, f)	(8, 48)	0.9440	0.9456	0.7971	0.0586
	(12, 32) [†]	0.9482	0.9507	0.8057	0.0560
	(16, 24)	<u>0.9458</u>	<u>0.9479</u>	<u>0.8007</u>	<u>0.0575</u>
PSAM w	3	0.9459	0.9485	0.8018	0.0573
	5	0.9458	0.9492	0.8006	0.0568
	7 [†]	0.9482	0.9507	0.8057	0.0560
	9	<u>0.9476</u>	<u>0.9493</u>	<u>0.8047</u>	<u>0.0566</u>
	11	0.9394	0.9445	0.7897	0.0599
NVF D	25	0.9483	<u>0.9499</u>	<u>0.8053</u>	<u>0.0562</u>
	50	0.9458	0.9486	0.8024	0.0570
	75 [†]	<u>0.9482</u>	0.9507	0.8057	0.0560
	100	0.9468	0.9488	0.8038	0.0568
PFEM b	4	<u>0.9470</u>	<u>0.9498</u>	<u>0.8039</u>	<u>0.0564</u>
	8 [†]	0.9482	0.9507	0.8057	0.0560
	16	0.9456	0.9483	0.8003	0.0571
PFEM PPD	24	0.9474	<u>0.9497</u>	0.8034	0.0564
	30 [†]	0.9482	0.9507	0.8057	0.0560
	36	<u>0.9481</u>	<u>0.9497</u>	<u>0.8056</u>	<u>0.0563</u>

smooth regions. Removing PFEM also causes a performance drop, showing the benefit of enhancing perceptually relevant frequency components. As shown in Tab. VIII, applying PFEM to more stages improves accuracy but increases inference time, as shallow and deep stages capture local noise patterns and broader spectral changes, respectively. These results show that PSAM and PFEM provide complementary spatial and frequency cues.

2) *Backbone Replacement*: To evaluate the backbone design, we replace the Swin Transformer with ResNet50 [84] and MobileNetV3 [85], while keeping GFSS, PSAM, PFEM, and the training protocol unchanged. Both CNN-based variants show lower performance than the Swin Transformer. This may be attributed to the fragmented nature of GFSS inputs. Because the input consists of spatially disjoint fragments, the resulting artificial boundaries disrupt the local continuity assumption essential for CNNs. In contrast, the Swin Transformer, leveraging its windowed self-attention mechanism, demonstrates greater robustness to such non-contiguous spatial structures.

3) *Comparison of Input Sampling Strategies*: We compare the adopted GFSS preprocessing scheme with direct resizing, random cropping with prediction averaging, and adaptive resizing followed by center cropping. GFSS performs best among all four accuracy metrics, followed by random cropping, whereas direct resizing and adaptive resizing with center cropping yield similar but metric-dependent results. Notably, even when GFSS is replaced with 25 random crops whose predictions are averaged, PINE obtains an SRCC of 0.9298, a PLCC of 0.9326, a KRCC of 0.7726, and an RMSE of 0.0667, still outperforming all competing methods in Tab. III. This result indicates that PINE’s strong performance

is not solely attributable to GFSS, but also benefits from the proposed noise-perception architecture.

4) *Hyperparameter Sensitivity Analysis*: We further examine the sensitivity of PINE to the adopted GFSS configuration and the principal hyperparameters of the HVS-inspired modules. For GFSS, we vary the grid-fragment pair (N, f) while enforcing $N \cdot f = 384$, thereby keeping the input resolution of the Swin Transformer fixed. We then independently vary the PSAM local window size w , the NVF scaling constant D , the PFEM DCT block size b , and the pixels-per-degree (PPD) value. In each experiment, only one parameter is changed, with all other settings held at their default values. The results are reported in Tab. VI.

The default configuration achieves the best overall performance across the tested parameter groups. PINE remains stable under moderate variations in the GFSS configuration and the parameters of NVF and PFEM, while an overly large PSAM window causes a more noticeable performance drop. Overall, these results demonstrate that PINE is robust to reasonable hyperparameter variations.

VI. CONCLUSION

This paper tackles the challenge of assessing real-world noise in ultra-high-definition (UHD) photography. We first introduce the UNIQ-5k dataset, a dedicated benchmark that encapsulates the complexity of authentic noise patterns, bridging the gap left by synthetic and low-resolution datasets. Building on this foundation, we propose PINE, a novel HVS-aware quality assessment method. By combining the adopted GFSS preprocessing with the proposed HVS-inspired perception modules (PSAM and PFEM), PINE performs fine-grained perceptual modeling on a fixed-size representation, thereby keeping the computational cost of the prediction network fixed after fragment extraction, largely independent of the native UHD resolution. The current model nevertheless remains computationally heavy, and its accuracy-complexity trade-off is reported explicitly.

Experimental results validate that PINE achieves state-of-the-art performance on the UNIQ-5k dataset and competitive performance on the evaluated public HR-IQA benchmarks. On the evaluated SIDD denoising results, PINE exhibits high rank-order consistency with the PSNR- and SSIM-based criteria, suggesting its potential as an auxiliary no-reference proxy for restoration-oriented denoising evaluation. Future work will focus on expanding the diversity of the UNIQ-5k dataset and developing lightweight variants of PINE to enable real-time on-device quality monitoring.

REFERENCES

- [1] V. Blahnik and O. Schindelbeck, “Smartphone imaging technology and its applications,” *Adv. Opt. Technol.*, vol. 10, no. 3, pp. 145–232, Jun. 2021.
- [2] Y. Lu and S.-W. Jung, “Progressive joint low-light enhancement and noise removal for raw images,” *IEEE Trans. Image Process.*, vol. 31, pp. 2390–2404, 2022.
- [3] H. C. Karaimer and M. S. Brown, “A software platform for manipulating the camera imaging pipeline,” in *Proc. Eur. Conf. Comput. Vis. Cham*, Switzerland: Springer, 2016, pp. 429–444.
- [4] N. Ponomarenko et al., “Color image database TID2013: Peculiarities and preliminary results,” in *Proc. Eur. Workshop Vis. Inf. Process. (EUVIP)*, Jun. 2013, pp. 106–111.

- [5] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–3.
- [6] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.
- [7] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4041–4056, 2020.
- [8] Y. Zhou et al., "When AWGN-based denoiser meets real noises," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13074–13081.
- [9] S. Kousha, A. Maleky, M. S. Brown, and M. A. Brubaker, "Modeling sRGB camera noise with normalizing flows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17442–17450.
- [10] C. Deng, S. Wang, A. C. Bovik, G.-B. Huang, and B. Zhao, "Blind noisy image quality assessment using sub-band kurtosis," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1146–1156, Mar. 2020.
- [11] K. Zhang et al., "Practical blind image denoising via Swin-conv-UNet and data synthesis," *Mach. Intell. Res.*, vol. 20, no. 6, pp. 822–836, Dec. 2023.
- [12] G. Zhai, A. Kaup, J. Wang, and X. Yang, "A dual-model approach to blind quality assessment of noisy images," *APSIPA Trans. Signal Inf. Process.*, vol. 4, no. 1, p. 4, 2015.
- [13] G. Yang, Y. Liao, Q. Zhang, D. Li, and W. Yang, "No-reference quality assessment of noise-distorted images based on frequency mapping," *IEEE Access*, vol. 5, pp. 23146–23156, 2017.
- [14] S. Lu, "No-reference image denoising quality assessment," in *Proc. Sci. Inf. Conf. Cham*, Switzerland: Springer, 2019, pp. 416–433.
- [15] H. Wu et al., "FAST-VQA: Efficient end-to-end video quality assessment with fragment sampling," in *Computer Vision—(ECCV)*, Cham, Switzerland: Springer, 2022, pp. 538–554.
- [16] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
- [17] E. C. Larson and D. M. Chandler, "Most apparent distortion: A dual strategy for full-reference image quality assessment," in *Image quality and system performance VI*, vol. 7242. Bellingham, WA, USA: SPIE, 2009, Art. no. 72420S.
- [18] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3575–3585.
- [19] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Häkkinen, "CID2013: A database for evaluating no-reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 390–402, Jan. 2015.
- [20] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3674–3683.
- [21] H. Huang, Q. Wan, and J. Korhonen, "High resolution image quality database," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 3105–3109.
- [22] V. Hosu, L. Agnolucci, O. Wiedemann, D. Iso, and D. Saupe, "UHD-IQA benchmark database: Pushing the boundaries of blind photo quality assessment," in *Proc. Eur. Conf. Comput. Vis. Cham*, Switzerland: Springer, 2024, pp. 467–482.
- [23] N. Chahine, A.-S. Calarasanu, D. Garcia-Civiero, T. Cayla, S. Ferradans, and J. Ponce, "An image quality assessment dataset for portraits," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9968–9978.
- [24] Y. Liu, B. Zhang, R. Hu, K. Gu, G. Zhai, and J. Dong, "Underwater image quality assessment: Benchmark database and objective method," *IEEE Trans. Multimedia*, vol. 26, pp. 7734–7747, 2024.
- [25] J. Shi, F. Shao, C. Tian, H. Chen, L. Xu, and Y.-S. Ho, "Progressive bidirectional feature extraction and enhancement network for quality evaluation of night-time images," *IEEE Trans. Multimedia*, vol. 26, pp. 1690–1705, 2024.
- [26] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
- [27] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Nov. 2012.
- [28] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.
- [29] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.
- [30] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.
- [31] S. Su et al., "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3667–3676.
- [32] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.
- [33] H. Zheng, H. Yang, J. Fu, Z.-J. Zha, and J. Luo, "Learning conditional knowledge distillation for degraded-reference image quality assessment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10222–10231.
- [34] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
- [35] S. Lao, "Attentions help CNNs see better: Attention-based hybrid image quality assessment network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2022, pp. 1140–1149.
- [36] S. Yang et al., "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1191–1200.
- [37] C. Chen et al., "TOPIQ: A top-down approach from semantics to distortions for image quality assessment," *IEEE Trans. Image Process.*, vol. 33, pp. 2404–2418, 2024.
- [38] K. Xu et al., "Boosting image quality assessment through efficient transformer adaptation with local feature enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2662–2672.
- [39] M. Mobini and M. R. Faraji, "Multi-scale gradient wavelet-based image quality assessment," *Vis. Comput.*, vol. 40, no. 12, pp. 8713–8728, Dec. 2024.
- [40] H. Zhang et al., "Learning degradation priors for reliable no-reference image quality assessment," *J. Vis. Commun. Image Represent.*, vol. 102, Jun. 2024, Art. no. 104189.
- [41] H. Yang et al., "Blind image quality assessment with small training sets by local and global features," *KSII Trans. Internet Inf. Syst.*, vol. 19, no. 4, p. 1244, 2025.
- [42] Z. You, X. Cai, J. Gu, T. Xue, and C. Dong, "Teaching large language models to regress accurate image quality scores using score distribution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 14483–14494.
- [43] A. Nema, H. Zhu, and W. Lin, "Holistic coreset selection for data efficient image quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2025, pp. 2432–2437.
- [44] O. Wiedemann, V. Hosu, S. Su, and D. Saupe, "Konx: Cross-resolution image quality assessment," *Qual. User Exper.*, vol. 8, no. 1, Dec. 2023, Art. no. 8.
- [45] J. Korhonen, Y. Su, and J. You, "Consumer image quality prediction using recurrent neural networks for spatial pooling," 2021, *arXiv:2106.00918*.
- [46] R. Liao, C. Hui, L. Yuan, H. Zhu, and F. Jiang, "Image quality assessment with compressed sampling," 2024, *arXiv:2404.17170*.
- [47] Z. Chen et al., "MobileIQA: Exploiting mobile-level diverse opinion network for no-reference image quality assessment using knowledge distillation," in *Proc. Eur. Conf. Comput. Vis. Cham*, Switzerland: Springer, 2024, pp. 1–17.
- [48] N. Jamshidi Avnaki, A. Ghildyal, N. Barman, and S. Zadtootaghaj, "LAR-IQA: A lightweight, accurate, and robust no-reference image quality assessment model," 2024, *arXiv:2408.17057*.
- [49] W. Sun et al., "Assessing UHD image quality from aesthetics, distortions, and saliency," in *Proc. Eur. Conf. Comput. Vis. Cham*, Switzerland: Springer, 2024, pp. 109–126.
- [50] V. Hosu et al., "AIM 2024 challenge on UHD blind photo quality assessment," in *Proc. Eur. Conf. Comput. Vis. Cham*, Switzerland: Springer, 2024, pp. 261–286.
- [51] Y. Liu, Y. Quan, G. Xiao, A. Li, and J. Wu, "Scaling and masking: A new paradigm of data sampling for image and video quality assessment," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 4, pp. 3792–3801.
- [52] Y. Li, Y. Mi, P. Zhou, Y. Wu, X. Wang, and S. Liu, "Unleashing vision transformer potential in image quality assessment via global-local adaptive interaction," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2026, pp. 10567–10571.

- [53] M. Oszust, "No-reference quality assessment of noisy images with local features and visual saliency models," *Inf. Sci.*, vol. 482, pp. 334–349, May 2019.
- [54] R. Verma and J. Ali, "A comparative study of various types of image noise and efficient noise removal techniques," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 10, pp. 617–622, 2013.
- [55] ITU-R, *Methodologies for the Subjective Assessment of the Quality of Television Images*, document Recommendation ITU-R BT.500-15, May 2023. [Online]. Available: <https://www.itu.int/rec/R-REC-BT.500-15-202305-I/en>
- [56] L. Azzari and A. Foi, "Indirect estimation of signal-dependent noise with nonadaptive heterogeneous samples," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3459–3467, Aug. 2014.
- [57] C. Liu, R. Szeliski, S. Bing Kang, C. L. Zitnick, and W. T. Freeman, "Automatic estimation and removal of noise from a single image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 299–314, Feb. 2008.
- [58] D. Levický and P. Foriš, "Human visual system models in digital image watermarking," *Radioengineering*, vol. 13, no. 4, pp. 38–43, 2004.
- [59] S. Majumder, T. S. Das, and S. K. Sarkar, "DWT and SVD based image watermarking scheme using noise visibility and contrast sensitivity," in *Proc. Int. Conf. Recent Trends Inf. Technol. (ICRTIT)*, Jun. 2011, pp. 938–942.
- [60] J. Mannos and D. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 4, pp. 525–536, Jul. 1974.
- [61] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
- [62] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K pretraining for the masses," 2021, *arXiv:2104.10972*.
- [63] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1692–1700.
- [64] P. Sedgwick, "Spearman's rank correlation coefficient," *BMJ*, vol. 349, Nov. 2014, Art. no. g7327.
- [65] H. Abdi, "The Kendall rank correlation coefficient," *Encyclopedia Meas. Statist.*, vol. 2, pp. 508–510, Jun. 2007.
- [66] P. Sedgwick, "Pearson's correlation coefficient," *BMJ*, vol. 345, Jul. 2012, Art. no. e4483.
- [67] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?—arguments against avoiding RMSE in the literature," *Geosci. Model Develop.*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014.
- [68] H. Fu, Y. Wang, W. Yang, A. C. Kot, and B. Wen, "DP-IQA: Utilizing diffusion prior for blind image quality assessment in the wild," 2024, *arXiv:2405.19996*.
- [69] Y. Huang, Z. Duan, P. Chen, L. Cai, L. Li, and W. Lin, "Learning scene-invariant distribution for generalizable blind image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 36, no. 1, pp. 609–621, Jan. 2026.
- [70] J. Wang, K. C. Chan, and C. C. Loy, "Exploring CLIP for assessing the look and feel of images," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, 2023, pp. 2555–2563.
- [71] L. Agnolucci, L. Galteri, and M. Bertini, "Quality-aware image-text alignment for opinion-unaware image quality assessment," 2024, *arXiv:2403.11176*.
- [72] M. Elad, "On the origin of the bilateral filter and ways to improve it," *IEEE Trans. Image Process.*, vol. 11, no. 10, pp. 1141–1151, Oct. 2002.
- [73] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 60–65.
- [74] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.
- [75] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.
- [76] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [77] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Comput. Vis. Image Understand.*, vol. 158, pp. 1–16, May 2017.
- [78] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.
- [79] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1712–1722.
- [80] C. Ren, X. He, C. Wang, and Z. Zhao, "Adaptive consistency prior based deep network for image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8596–8606.
- [81] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.
- [82] J. Zhang, Y. Zhang, J. Gu, J. Dong, L. Kong, and X. Yang, "Xformer: Hybrid X-shaped transformer for image denoising," 2023, *arXiv:2303.06440*.
- [83] B. Li, H. Zhao, W. Wang, P. Hu, Y. Gou, and X. Peng, "MaIR: A locality- and continuity-preserving mamba for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 7491–7501.
- [84] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [85] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.