BWAFDA: Block-wise Weighted Attention Fusion with Detail-aware for No-Reference Image Quality Assessment

Shilv Cai[1], Jian Jin[1] *, Tianang Chen[2], Zhuangzi Li[1], and Weisi Lin[1]

[1] Nanyang Technological University, 639798, Singapore
{shilyu.cai, jian.jin, zhuangzi.li, WSLin}@ntu.edu.sg
[2] Texas A&M University, College Station TX 77843, USA
tianang@tamu.edu

Abstract. We present BWAFDA, a dual-branch no-reference image quality assessment model that replaces uniform averaging with detail-aware, block-wise fusion. The context encoder branch produces multi-scale block tokens, and a lightweight detail branch extracts high-frequency cues. These streams are combined late. First, block tokens are aggregated via block-wise attention across scales, guided by a soft multi-scale prior that stabilizes cross-scale interactions and preserves localized artifacts. The resulting descriptors are then fused per scale using learned weights. Subsequently, the detail cues are injected to refine distortion-sensitive evidence. Finally, a class-token-guided head performs "where-to-trust-what" selection over blocks to produce the quality score, without the need for region labels. Comprehensive evaluations on seven public NR-IQA benchmarks show state-of-the-art (SOTA) performance. Across multiple controlled ablations, we observe consistent drops in correlation compared to the full model, indicating that region-aware masking, contextual features, and multi-scale aggregation contribute jointly rather than redundantly. Overall, detail-aware, block-wise fusion delivers SOTA accuracy while preserving data efficiency and transferability by aligning localized artifact cues with global context.

Keywords: Image quality assessment · Weighted attention · Detail aware

1 Introduction

Image Quality Assessment (IQA) aims to quantitatively evaluate images in alignment with human perceptual judgment of visual quality [25, 37, 38], serving as a crucial benchmark for applications reliant on visual fidelity and the development of image processing techniques. Objective Image Quality Assessment comprises Full-Reference (FR-IQA) [32, 42], Reduced-Reference (RR-IQA) [4, 39], and No-Reference IQA (NR-IQA) [12, 26]. NR-IQA is especially critical because real-world pipelines often lack a pristine reference image. NR-IQA models infer quality directly from the distorted content itself, learning to detect and weigh artifacts such as noise, blur, compression ringing, and exposure errors. This makes it

* Corresponding author.

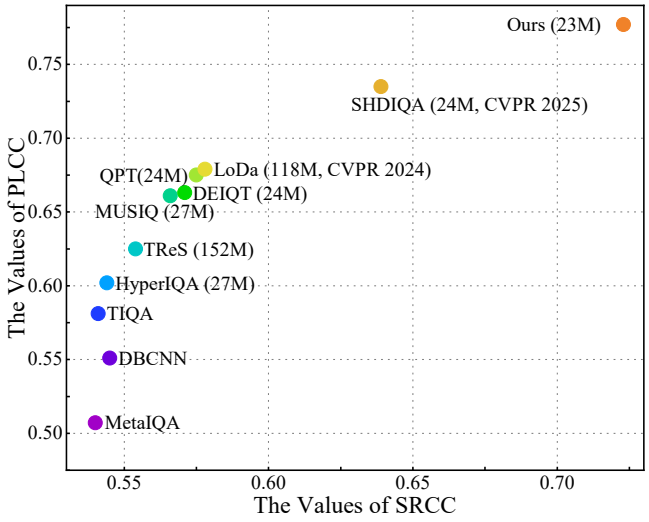

Fig. 1: Performance Comparison on the LIVEFB Dataset. The proposed BWAFDA method achieves state-of-the-art performance among all existing NR-IQA methods.

indispensable for large-scale, real-world scenarios where the diversity of content, limited bandwidth, and on-device constraints demand a fast, robust, and no-reference method of predicting quality. NR-IQA has become increasingly critical for computer vision tasks such as image restoration [2] and super-resolution [7].

There has been impressive progress in recent years thanks to advances in data-driven models, especially those based on deep neural networks [6, 12, 13, 18, 20, 23, 24, 43, 45, 46]. Among these, transformers [35, 36] have emerged as a particularly powerful architecture for this task. They are a natural fit for NR-IQA because they can capture long-range context, flexibly integrate multi-scale evidence, and support region-selective weighting. However, many transformer-based NR-IQA models [43, 45] rely on global pooling or a single CLS token for image-level summarization. While simple, this approach is suboptimal as it tends to wash out subtle, localized distortions. Small artifacts (like fine speckles in leaf textures or faint edge ringing) occupy very little area, so their signal gets drowned by large clean regions when everything is averaged together. Further downsampling weakens these fine details. Without guidance, attention often follows visually salient content instead of subtle flaws that actually affect perceived quality. These limitations motivate an architecture that retains the global reasoning of transformers while explicitly routing evidence by region and scale.

In this paper, we present BWAFDA, a dual-branch NR-IQA model that replaces uniform averaging with structured, region-aware fusion. A lightweight region-mask module predicts soft partitions over the ViT patch grid. These masks bias a block-wise multi-scale aggregator that pools evidence per region via masked cross-attention and then fuses across scales with learned per-scale weights. Finally, a class-token-guided head gates the block descriptors-implement-

ing a late "where-to-trust-what" fusion between the context encoder and the detail-aware stream, without region supervision. Across seven public NR-IQA benchmarks, BWAFDA delivers state-of-the-art PLCC and SROCC performance. The results on a representative dataset (LIVEFB) are illustrated in Fig. 1. Controlled ablations-removing the detail-aware branch, disabling mask bias, or constraining fusion to fewer scales-consistently reduce correlation, showing that region masks, detail cues, and multi-scale aggregation are complementary rather than redundant. Under 20% and 40% training-data regimes, BWAFDA maintains clear margins over strong baselines; cross-dataset tests indicate robust transfer.

Our contributions can be summarized as follows:

• We propose a dual-branch framework that integrates a transformer-based encoder with a lightweight detail-aware branch via block-wise fusion.

• We present a mask-biased, multi-scale block aggregation module with per-scale fusion that is learnt. This replaces global pooling with a mask-biased, block-wise fusion process that routes evidence by region.

• Our design is compact, data-efficient, and transferable, achieving state-of-the-art results while remaining competitive under limited training data and generalizing reliably across datasets.

2 Related Works

2.1 NR-IQA with Transformers

Vision Transformer (ViT)-based NR-IQA approaches can be grouped into two types. The first type is hybrid transformers [5, 9, 19, 40, 45]. The second type is pure transformers [13, 17, 28]. Hybrid models fuse multi-scale CNN features to enhance local distortion sensitivity, but this often comes with higher architectural complexity and training overhead. Although pure transformers such as DEIQT [28] excel at forming strong global abstractions via the CLS token, a well-known limitation is their inadequate representation of fine-grained local image structure [53]. Several memory-efficient approaches employ local or block-wise self-attention [22, 29], prioritizing computational efficiency over full long-range interaction. Although sampling-based designs (*e.g.*, VT-IQA [15]) endeavor to balance global and local information through input resampling and global attention, they are not explicitly supervised to localize degradations. Consequently, their capacity to reliably capture region-specific artifacts throughout training remains limited. We introduce a Vision Transformer (ViT) framework that explicitly couples local and global information. Its design consists of four key components: a soft multi-scale prior $Pr^{(s)}$ stabilizes the cross-scale attention mechanism; a block-wise aggregation module produces distortion-aware patch descriptors; a class token-conditioned head then selectively weights these descriptors; and finally, a lightweight, detail-aware branch amplifies high-frequency cues. This approach preserves the transformer's global reasoning while providing targeted sensitivity to local defects.

2.2 NR-IQA with Detail-aware Representation

Early NR-IQA methods demonstrated the predictive power of high-frequency and edge information for perceived quality. They explicitly leveraged detail-sensitive cues, such as natural-scene statistics of local luminance or contrast (BRISQUE [25]) and joint statistics of gradients or Laplacians (GM-LOG [41]). Opinion-unaware models like IL-NIQE [47] further aggregated local statistics without MOS labels, reinforcing the value of localized detail beyond global appearance priors. With deep learning, dual-branch architectures (DBCNN [48]) combined semantic features with distortion-aware (texture or frequency) cues to better capture authentic artifacts, while remaining fully reference-free. Zhang et al. [50] enhanced feature extraction efficiency by pre-processing images into separate high-frequency and low-frequency distortion maps using the Sobel operator and Piecewise Smooth Image Approximation (PSIA). In the transformer-based methods, attention-centric NR-IQA (MANIQA [43]) assigns patch-level weights and multi-dimensional attention so that fine defects are not averaged out by global pooling, and content-adaptive frameworks (HyperIQA [34]) separate "content understanding" from "perception rule learning", effectively boosting detail cues when they matter for quality judgments. By aligning frequency- or gradient-based priors with tokenized representations and fusing them with global context, NR-IQA models can retain and emphasize fine, localized artifacts, ensuring these high-frequency cues are properly integrated into global quality inference.

2.3 NR-IQA with Multi-scale Aggregation

Multi-scale and region-aware transformers have become central to NR-IQA because they preserve localization of artifacts while reconciling evidence at the global image level. MUSIQ [13] introduces a multi-scale transformer that processes native-resolution images and explicitly fuses scales to capture quality "at different granularities", addressing the loss of fine cues under single-scale pooling. MANIQA [43] complements this with patch-weighted prediction and multi-dimensional attention, ensuring block contributions are selectively aggregated rather than uniformly averaged. Content-adaptive designs like HyperIQA [34] further show that quality estimation benefits from region-dependent rules that modulate local evidence before global scoring. However, despite these advances, SHDIQA [18] notes that sampling- or block-structured transformers often lack explicit supervision for localized degradations, making small-area artifacts hard to capture. To address this gap, we replace uniform pooling with a soft prior that regulates cross-scale attention during block-wise aggregation, preserving localized artifacts while reconciling them with global context.

3 Method

3.1 Overview

As illustrated in Fig. 2, BWAFDA takes an input RGB image I and predicts a scalar quality $q \in \mathbb{R}$. The context encoder branch $f_{\text{con}}(I)$ produces patch tokens.

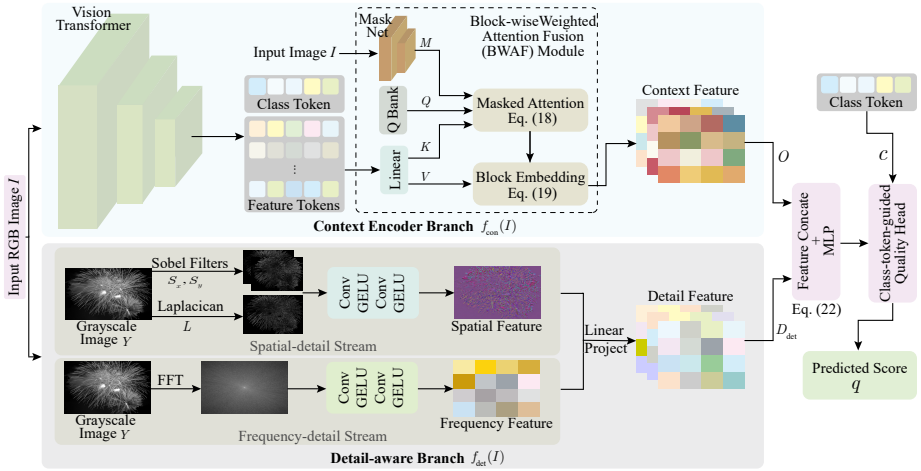

Fig. 2: Overview of the proposed BWAFDA framework. The model has two cooperating branches: (1) The context encoder branch that maps the RGB input I into feature tokens with a ViT. It also uses a mask net to produce masks M. Furthermore, it applies masked attention (Eq. (18)) over $Q/K/V$ from a learned Q-bank, and forms block embeddings (Eq. (19)) which are fused by the block-wise weighted attention fusion (BWAF) module. (2) The detail-aware branch that derives spatial-detail features from the grayscale image Y using Sobel/Laplacian filters and frequency-detail features via FFT, refined by Conv–GELU blocks. The fused context features are concatenated with detail features and passed to the class-token-guided quality head to output the predicted quality score q. Notation: I denotes the input RGB image; Y denotes the grayscale image; M denotes the learned block mask; q denotes the final predicted score.

The detail-aware branch $f_{\text{det}}(I)$ extracts high-frequency and edge-like cues and projects to the shared embedding space. The block-wise aggregator pools tokens per predicted block across multiple scales using mask-biased attention, and fuses the resulting per-block descriptors with learned per-scale weights. The quality head then applies class-token-guided gating over blocks to produce final score q.

3.2 Context Encoder and Detail-Aware Branch

Detail-aware Branch. The objective of this branch is to provide additional cues by extracting local structures and frequency-band energy that are aligned with the ViT patch grid.

(1) Spatial-detail stream. Convert the input RGB image $I \in [0,1]^{3\times H\times W}$ to grayscale image Y, then apply fixed Sobel filters S_x, S_y and a Laplacian L:

$$G_x = Y \cdot S_x, \quad G_y = Y \cdot S_y, \quad \Delta = Y \cdot L. \tag{1}$$

Stack spatial cues and project to width $\mathfrak{D}$:

$$F_{\text{sp}} = \text{Proj}_{\text{sp}}([G_x, G_y, \Delta]) \in \mathbb{R}^{\mathfrak{D}\times H\times W}. \tag{2}$$

Align to the patch grid via area pooling (same grid as tokens):

$$D_{\text{sp}} = \text{Pool}_{(p_h, p_w)}(F_{\text{sp}}) \in \mathbb{R}^{\mathfrak{D} \times P_h \times P_w}. \tag{3}$$

(2) Frequency-detail stream. Compute rFFT magnitude on grayscale Y:

$$\mathcal{M}(u, v) = |\text{rFFT2}(Y)| \in \mathbb{R}^{H \times (W/2+1)}. \tag{4}$$

Define B radial bands via masks $\{R\}_{b=1}^{B}$, where $R_b(u, v) \in \{0, 1\}$ selects frequencies with radius r in $[r_{b-1}, r_b)$. Then band energies can be described as:

$$E_b = \sum_{u,v} R_b(u, v) \cdot \mathcal{M}(u, v), b = 1, ..., B. \tag{5}$$

Tile each E_b spatially, concatenate, and project:

$$F_{\text{freq}} = \text{Proj}_{\text{freq}}([E_1, ..., E_B]) \in \mathbb{R}^{\mathfrak{D} \times H \times W}, \tag{6}$$

then pool to the patch grid:

$$D_{\text{freq}} = \text{Pool}_{(P_h, P_w)}(F_{\text{freq}}) \in \mathbb{R}^{\mathfrak{D} \times P_h \times P_w}. \tag{7}$$

(3) Detail feature fusion. Concatenate spatial and frequency features on the grid and reduce back to width $\mathfrak{D}$.

$$D_{\text{det}} = \phi([D_{\text{sp}} || D_{\text{freq}}]) \in \mathbb{R}^{\mathfrak{D} \times P_h \times P_w}, \tag{8}$$

where ϕ is a 1×1 linear projection function. Multi-scale variants $D_{\text{det}}^{(s)}$ are obtained by the same pooling schedule as the context tokens:

$$D_{\text{det}}^{(s)} = \text{Pool}_{(P_h^{(s)}, P_w^{(s)})}(D_{\text{det}}) \in \mathbb{R}^{\mathfrak{D} \times P_h^{(s)} \times P_w^{(s)}}. \tag{9}$$

Context Encoder Branch. *(1) ViT-based context encoder.* Let the input RGB image be $I \in [0, 1]^{3 \times H \times W}$. We use a ViT-style encoder as backbone (patch size is P, width is $\mathfrak{D}$). Firstly, we split I into non-overlapping patches and project:

$$z_0 = [c_0; t_1, ..., t_N] + E_{\text{pos}},\ t_n = W_e \cdot \text{vec}(I_n) \in \mathbb{R}^{\mathfrak{D}},\ N = \frac{H}{P} \cdot \frac{W}{P}, c_0 \in \mathbb{R}^{\mathfrak{D}}, \tag{10}$$

where E_{pos} is positional encoding, c_0 is the class token, and W_e is the learned patch-embedding matrix that projects the flattened patch to the model width $\mathfrak{D}$. The vec$(\cdot)$ is the vectorization operator, it flattens the patch $I_n \in \mathbb{R}^{3 \times P \times P}$ into a column vector $\text{vec}(I_n) \in \mathbb{R}^{3 \cdot P^2}$.

Then, for layer $\ell = 1, ..., L$:

$$\hat{z}_{\ell-1} = \text{LN}(z_{\ell-1}),\ z_\ell^{\text{msa}} = \hat{z}_{\ell-1} + \text{MSA}(\hat{z}_{\ell-1}),\ z_\ell = z_\ell^{\text{msa}} + \text{MLP}(\text{LN}(z_\ell^{\text{msa}})), \tag{11}$$

where LN$(\cdot)$ is layer normalization, MSA$(\cdot)$ is the multi-head self-attention module. The output can be written as:

$$z_L = [c; t_1^{(L)}, ..., t_N^{(L)}], c \in \mathbb{R}^{\mathfrak{D}}, t_n^{(L)} \in \mathbb{R}^D. \tag{12}$$

Finally, we reshape patch tokens to a grid:

$$T \in \mathbb{R}^{\mathfrak{D}\times P_h\times P_w}, \quad P_h = \frac{H}{P}, \quad P_w = \frac{W}{P}, \tag{13}$$

and form a multi-scale pyramid by area pooling:

$$T^{(s)} = \mathrm{Pool}_{(P_h^{(s)},P_w^{(s)})}(T) \in \mathbb{R}^{\mathfrak{D}\times P_h^{(s)}\times P_w^{(s)}}, s = 1, ..., S. \tag{14}$$

This encoder provides context tokens $T^{(s)}$ for the block-wise weighted attention fusion (BWAF) module.

(2) Block-wise weighted attention fusion. We propose a block-wise aggregator (BWAF) that converts multi-scale patch tokens and per-block masks into one $\mathfrak{D}$-dimensional embedding for each block in an image. At each scale, learnable block queries attend to tokens under a mask-derived probabilistic prior, and the resulting block embeddings are fused across scales via a learned convex combination.

Let B denote the batch size, $\mathfrak{D}$ denote the feature dimension, H denote the number of attention heads, and $d_h = \mathfrak{D}/H$. We consider S input scales indexed by $s \in \{1, ..., S\}$. The token grid on scale s has spatial size $P_h^{(s)} \times P_w^{(s)}$ and $N^{(s)} = P_h^{(s)} \cdot P_w^{(s)}$ tokens. The token is described as $X^{(s)} \in \mathbb{R}^{B\times N^{(s)}\times\mathfrak{D}}$ on the scale s. Each sample provides K blocks (variable per batch, but $K \leqslant K_{\max}$). During the experiment, $K_{\max}$ took the value of 4. On the scale s, the block masks $M^{(s)} \in \mathbb{R}^{B\times K\times N^{(s)}}$ reshaped from $[B, K, P_h^{(s)}, P_w^{(s)}]$; the row $M^{(s)}_{b,k,:}$ indicates the support for the tokens for block k. For query bank $Q_0 \in \mathbb{R}^{K_{\max}\times\mathfrak{D}}$, we select the first K rows at runtime. The coefficients of the linear projection are denoted as $W_Q \in \mathbb{R}^{\mathfrak{D}\times\mathfrak{D}}$, $W_K \in \mathbb{R}^{\mathfrak{D}\times\mathfrak{D}}$, and $W_v \in \mathbb{R}^{\mathfrak{D}\times\mathfrak{D}}$ respectively.

At scale s, for sample b, and block k, let $M^{(s)}_{b,k,:} \in \mathbb{R}^N_{\geqslant 0}$ denote the (flattened) block mask over the $N^{(s)} = P_h^{(s)} \cdot P_w^{(s)}$ tokens. We convert this mask into a token-level prior distribution $P^{(s)}_{b,k,:} \in \triangle^{N^{(s)}-1}$ via row-wise normalization:

$$S^{(s)}_{b,k} = \sum_{n=1}^{N^{(s)}} M^{(s)}_{b,k,n}, \quad P^{(s)}_{b,k.n} = \begin{cases} \frac{M^{(s)}_{b,k,n}}{S^{(s)}_{b,k}+\varepsilon}, & S^{(s)}_{b,k} > 0, \\ \frac{1}{N^{(s)}}, & S^{(s)}_{b,k} = 0, \end{cases} \tag{15}$$

with a small constant $\varepsilon > 0$ for numerical stability. Here $b \in [1, B]$ indexes batch samples (images), $k \in [1, K]$ indexes blocks for that sample ($K \leqslant K_{\max}$), and $s \in [1, S]$ indexes scales; tokens within a scale are indexed by $n \in [1, N^{(s)}]$. $\triangle^{N^{(s)}-1} = \{p \in \mathbb{R}^{N^{(s)}} : p_n \geqslant 0 \text{ for all } n, \sum_{n=1}^{N^{(s)}} p_n = 1\}$ means the $(N^{(s)} - 1)$-dimensional probability simplex. We inject $\log P^{(s)}$ additively into attention logits as a soft prior that encourages, but does not force, focus on mask-selected tokens.

We form block queries from the bank, broadcast to the batch, project queries, keys, and values:

$$Q^{(s)} = \underbrace{Q_0[1:K,:]}_{\in\mathbb{R}^{K\times\mathfrak{D}}} \cdot W_Q, \quad K^{(s)} = X^{(s)} \cdot W_K, \quad V^{(s)} = X^{(s)} \cdot W_V. \tag{16}$$

The output is then split into H attention heads. Here, H refers to the number of heads, and the superscripts s and b are omitted for brevity. For the h-th head,

$$Q^{(h)} \in \mathbb{R}^{B\times K\times d_h}, \quad K^{(h)},\ V^{(h)} \in \mathbb{R}^{B\times N^{(s)}\times d_h},\ \text{for } h = 1, ..., H. \tag{17}$$

Let the log-prior as $Pr^{(s)} = \log\ P^{(s)} \in \mathbb{R}^{B\times K\times N^{(s)}}$. The masked attention can be described as

$$A^{(h,s)} = \text{softmax}(\frac{Q^{(h)}(K^{(h)})^T}{\sqrt{d_h}} + Pr^{(s)}) \in \mathbb{R}^{B\times K\times N^{(s)}}, \tag{18}$$

and the per-scale block embedding is

$$\begin{aligned} O^{(h,s)} &= A^{(h,s)} \cdot V^{(h)} \in \mathbb{R}^{B\times K\times d_h}, \\ O^{(s)} &= \text{Concat}_h O^{(h,s)} \in \mathbb{R}^{B\times K\times \mathfrak{D}}. \end{aligned} \tag{19}$$

The prior biases attention toward desired regions while the learned query-key similarity determines fine-grained token weights, yielding robust block descriptors.

We maintain a global learnable vector of fusion logits $w \in \mathbb{R}^S$. To activate S scales, we compute as follows:

$$\alpha_s = \frac{\exp(w_s)}{\sum_{t=1}^{S}\exp(w_t)}, s = 1, ..., S, \tag{20}$$

and fuse per-scale outputs into the final block representation

$$O = \sum_{s=1}^{S} \alpha_s \cdot O^{(s)} \in \mathbb{R}^{B\times K\times \mathfrak{D}}. \tag{21}$$

Thus, the module produces one $\mathfrak{D}$-dimensional embedding per block after an adaptive, convex multi-scale aggregation. In our experiments, the max-scales is set to 3. The aggregator learns per-block queries that perform mask-guided cross-attention over multi-scale tokens and combines scales through a learned convex fusion, producing compact and discriminative block embeddings suitable for downstream quality assessment.

3.3 Class-token-guided quality head

Let $c \in \mathbb{R}^{\mathfrak{D}}$ denote global class token from context encoder branch, and let $D_{\text{det}} = \{D^k_{\text{det}}\}_{k=1}^K$, be the block-aligned detail descriptors produced by the detail-aware branch. For each block k, we fuse the contextual and detail information by concatenating the BWAF output with the aligned detail descriptor and projecting back to width $\mathfrak{D}$:

$$D_k = \psi([O^k||D^k_{\text{det}}]), \tag{22}$$

where O^k is block-wise context embedding from BWAF, $\psi(\cdot)$ is linear projection.

For each block k, we concatenate the global context with the block descriptor and map it through a lightweight MLP to obtain a gate logit q_k:

$$\begin{aligned} u_k &= [c||D_k] \in \mathbb{R}^{2\cdot\mathfrak{D}}, \\ q_k &= w_2 \cdot \sigma(w_1 \cdot u_k + b_1) + b_2, \end{aligned} \tag{23}$$

where $\sigma(\cdot)$ is a pointwise nonlinearity (*e.g.*, GELU). In the gating MLP, $w_1 \in \mathbb{R}^{H\times 2\cdot\mathfrak{D}}$ and $b_1 \in \mathbb{R}^H$ are the hidden-layer weight matrix and bias that map the concatenated input $[c||D_k]$ to a hidden feature, while $w_2 \in \mathbb{R}^H$ and $b_2 \in \mathbb{R}$ are the output-layer weight vector and bias that convert this hidden feature into the scalar gate logit q_k. A softmax over blocks yields normalized importance weights:

$$\begin{aligned} g_k &= \frac{\exp(q_k)}{\sum_{j=1}^{K}\exp(q_j)}, \\ g &= [g_1, ..., g_K] \in \triangle^{K-1}, \end{aligned} \tag{24}$$

where $\Delta^{K-1} = \{g \in \mathbb{R}^K : g_i \geq 0 \text{ for all } i, \sum_{i=1}^{K} g_i = 1\}$ denotes the $(K-1)$-dimensional probability simplex. Thus, the allocation of importance across blocks is conditioned on global context c as well as the feature evidence D_k.

The head forms a gated summary feature by combining block descriptors:

$$\widetilde{D} = \sum_{k=1}^{K} g_k \cdot D_k \in \mathbb{R}^{\mathfrak{D}}. \tag{25}$$

The image-level quality score q is obtained via a regressor:

$$q = w_{\text{reg}} \cdot \phi(\widetilde{D}) + b_{\text{reg}}. \tag{26}$$

In this case, the function denoted by the symbol $\phi(\cdot)$ can be an optional hidden layer (MLP) with nonlinearity. The learnable weight vector of the final regression layer is represented by w_{reg}, and the corresponding bias is represented by b_{reg}.

3.4 Training Objective

Given a supervised mini-batch with image-level Mean Opinion Score (MOS) targets q^{gt}, model makes the prediction q. The loss combines a regression term, a pairwise ranking term, and two regularizers on the soft region masks:

$$\begin{aligned} \mathcal{L} =& \mathcal{L}_{\text{reg}}(q, q^{\text{gt}}) + \mathcal{L}_{\text{rank}}(q, q^{\text{gt}}) \\ &+ s(t) \cdot (\lambda_{\text{ent}} \cdot \mathcal{H}(M) + \lambda_{\text{tv}} \cdot \text{TV}(M)), \end{aligned} \tag{27}$$

where $\mathcal{L}_{\text{reg}} = \text{SmoothL1}(q, q^{\text{gt}})$ penalizes absolute error in the predicted quality. $\mathcal{L}_{\text{rank}} = \text{RankHinge}(q, q^{\text{gt}})$ [1] enforces correct relative ordering of image qualities in the batch.

Let $M \in \mathbb{R}^{B\times K\times P_h\times P_w}$ denote the soft region masks, where a softmax is applied along the block dimension K at every spatial location (i, j), *i.e.*,

$\sum_{k=1}^{K} M_{b,k,i,j} = 1$. These masks are predicted by a lightweight convolutional mask net attached to the context encoder branch. The mask net is a shallow, fully convolutional module that converts hand-crafted SNR, edge, and residual-noise cues into K soft masks on the patch grid. These masks guide block-wise aggregation while remaining lightweight and independent of the ViT backbone. We refer readers to the supplementary material for architectural details. Importantly, the mask net is trained end-to-end with the total loss $\mathcal{L}$. Define the spatial index set $\Omega = \{1, \ldots, P_h\} \times \{1, \ldots, P_w\}$, $|\Omega| = P_h \cdot P_w$. $\mathcal{H}(M)$ encourages confident (low-entropy) soft masks [10]:

$$\mathcal{H}(M) = -\frac{1}{B \cdot K \cdot |\Omega|} \sum_{b=1}^{B} \sum_{k=1}^{K} \sum_{(i,j)\in\Omega} M_{b,k,i,j} \cdot \log M_{b,k,i,j}. \tag{28}$$

$\mathrm{TV}(M) = \mathrm{mean}(|\bigtriangledown_x M| + |\bigtriangledown_y M|)$ encourages spatial smoothness [30]. The weights λ_{ent} and λ_{tv} are user-set hyper-parameters with a default value of 10^{-2} and 2×10^{-2}, respectively. The regularizers are modulated by: $s(t) = 0.25 + 0.75 \cdot \frac{t}{T}$, where t is the current epoch and T is the total number of epochs, which down-weights mask penalties as training progresses.

4 Experiments

4.1 Experimental Settings

Datasets. Experiments are conducted on 7 typical NR-IQA datasets. These include 4 synthetic datasets. These are LIVE [33], CSIQ [14], TID2013 [27], and KADID [21]. The other 3 are authentic datasets. These are LIVEC [8], KonIQ [11], and LIVEFB [44]. For authentic datasets, LIVEC (LIVE Challenge) contains 1,162 authentically distorted images captured by diverse mobile devices and photographers. KonIQ includes 10,073 images collected from public sources. LIVEFB is a large real-world picture quality dataset with 39,810 images. For the remaining synthetic datasets, LIVE has 779 distorted images with 5 distortion types, CSIQ includes 866 distorted images built from 30 references with 6 distortion types, and KADID contains 10,125 distorted images produced from 81 references across 25 distortion types and 5 levels. TID2013 is widely used as a synthetic-distortion benchmark and comprises 3,000 distorted images generated from 25 reference images under 24 distortion types at 5 degradation levels.

Evaluation Metrics. Spearman's Rank-Order Correlation Coefficient (SRCC) and Pearson's Linear Correlation Coefficient (PLCC) are reported to measure prediction monotonicity and accuracy, respectively. Values ranging from -1 to 1 are used to indicate performance. Higher values indicate better performance.

Training Settings. Our proposed BWAFDA IQA network is trained end-to-end with SGD [3] optimizer and mixed precision (AMP). The backbone of the

Table 1: Performance comparison measured by the average of SRCC and PLCC. Best results are in bold, second-best are underlined.

	LIVE		TID2013		CSIQ		KADID		LIVEC		KonIQ		LIVEFB	
Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
ILNIQE [47]	0.906	0.902	0.648	0.521	0.865	0.822	0.558	0.534	0.508	0.508	0.537	0.523	0.332	0.294
BRISQUE [25]	0.944	0.929	-	-	0.748	0.812	0.567	0.528	0.629	0.629	0.685	0.681	0.341	0.303
DBCNN [48]	0.971	0.968	0.865	0.816	0.959	0.946	0.856	0.851	0.869	0.851	0.884	0.875	0.551	0.545
TIQA [45]	0.965	0.949	0.858	0.846	0.838	0.825	0.855	0.850	0.861	0.845	0.903	0.892	0.581	0.541
MetaIQA [52]	0.959	0.960	0.868	0.856	0.908	0.899	0.775	0.762	0.802	0.835	0.856	0.887	0.507	0.540
HyperIQA (27M) [34]	0.966	0.962	0.858	0.840	0.942	0.923	0.845	0.852	0.882	0.859	0.917	0.906	0.602	0.544
TReS (152M) [9]	0.968	0.969	0.883	0.863	0.942	0.922	0.858	0.859	0.877	0.846	0.928	0.915	0.625	0.554
MUSIQ (27M) [13]	0.911	0.940	0.815	0.773	0.893	0.871	0.872	0.875	0.746	0.702	0.928	0.916	0.661	0.566
Re-IQA (48M) [31]	0.971	0.970	0.861	0.804	0.960	0.947	0.885	0.872	0.854	0.840	0.923	0.914	-	-
DEIQT (24M) [28]	0.982	0.980	0.908	0.892	0.963	0.946	0.887	0.889	0.894	0.875	0.934	0.921	0.663	0.571
QPT (24M) [51]	-	-	-	-	-	-	-	-	0.914	0.895	0.941	0.927	0.675	0.575
LIQE (151M) [49]	0.951	0.970	-	-	0.939	0.936	0.931	0.930	0.910	0.904	0.908	0.919	-	-
QFM-IQM (24M) [16]	0.983	0.981	0.932	0.916	0.965	0.954	0.906	0.906	0.913	0.891	0.936	0.922	0.667	0.567
LoDa (118M) [40]	0.979	0.975	0.901	0.869	-	-	0.936	0.931	0.899	0.876	0.944	0.932	0.679	0.578
SHDIQA (24M) [18]	0.984	0.982	-	-	0.972	0.962	0.940	0.937	0.922	0.909	0.948	0.937	0.735	0.639
BWAFDA (23M)	**0.985**	**0.983**	**0.975**	**0.972**	**0.988**	**0.983**	**0.972**	**0.970**	**0.928**	**0.917**	**0.949**	**0.940**	**0.777**	**0.723**
Rel. gain vs 2nd-best	+0.10%	+0.10%	+4.61%	+6.11%	+1.65%	+2.18%	+3.40%	+3.52%	+0.65%	+0.88%	+0.11%	+0.32%	+5.71%	+13.15%

context encoder branch is ViT-S [35] (DeiT-III). The model is trained for 40 epochs with warm-up, then cosine annealing. We use a cosine learning rate schedule with 10% warm-up, employing a rate of 5×10^{-5} for the context encoder backbone and 5×10^{-4} for all remaining parameters. During training, we apply random cropping with crop sizes that are multiples of 16. At test time, we evaluate on the full image after padding it to the nearest multiple of 16 to satisfy ViT-S's input constraints. We divide the data into two parts: 80% for training and 20% for testing. We then repeat the training process 10 times, using different random numbers for each repetition. Finally, we report the mean SRCC and PLCC to reduce the effect of sampling bias. All experiments are conducted on four NVIDIA RTX 4090 GPUs or single NVIDIA RTX 5090 GPU.

4.2 Experimental Results

Evaluation on Individual Database. Tab. 1 shows how the proposed BWAFDA compares with other (SOTA) NR-IQA methods. These include ViT-based IQA methods such as SHDIQA [18] and LoDa [40], as well as degradation modelling-based methods such as Re-IQA [31] and QPT [51]. The superior performance of the proposed BWAFDA over all other methods across all seven datasets is clearly evident. It is worth noting that BWAFDA achieves a significant performance enhancement on the most challenging LIVEFB dataset. We primarily attribute the +5.71% PLCC and +13.15% SRCC gains to block-wise weighted fusion with detail-aware branches. This provides robust local defect sensing while maintaining global contextual information, which is particularly effective for the noisy, real-world LIVEFB dataset.

Evaluation of Data-efficient Learning. We demonstrate the data efficiency of our proposed approach in Tab. 2. We train models using 20% and 40% of the available images, following the efficient validation protocol outlined in SHDIQA [18], LoDa [40], and DEIQT [28]. This process was repeated 10 times

with different seeds, and the average SRCC and PLCC scores are reported. Under the 20% and 40% low-data regimes on LIVEC and KonIQ datasets, our method achieves comparable or better performance than prior methods. We attribute these gains to our integrated approach of block-wise weighted fusion and detail-aware branches. This combination not only enhances local defect sensing but also effectively suppresses content bias and label noise, leading to improved rank consistency and linear correlation.

Table 2: Validation of data-efficient learning with a training set containing 20% and 40% of the images.

Mode	Methods	LIVEC		KonIQ	
		PLCC	SRCC	PLCC	SRCC
20%	DEIQT [28]	0.822	0.792	0.908	0.888
	LoDa [40]	0.854	0.815	0.923	0.907
	SHDIQA [18]	0.866	0.840	0.928	0.915
	BWAFDA	**0.904**	**0.884**	**0.929**	**0.919**
40%	DEIQT [28]	0.855	0.838	0.922	0.903
	LoDa [40]	0.879	0.849	0.935	0.922
	SHDIQA [18]	0.889	0.869	**0.942**	0.929
	BWAFDA	**0.913**	**0.896**	**0.942**	**0.932**

Table 3: SRCC on the cross datasets validation. The best results are highlighted in bold, second-best are underlined.

Training	LIVEFB		LIVEC	KonIQ	LIVE	CSIQ
Testing	KonIQ	LIVEC	KonIQ	LIVEC	CSIQ	LIVE
DBCNN [48]	0.716	0.724	0.754	0.755	0.758	0.877
P2P-BM [44]	0.755	0.738	0.740	0.770	0.712	-
TReS [9]	0.713	0.740	0.733	0.786	0.761	-
DEIQT [28]	0.733	0.781	0.744	0.794	0.781	0.932
LoDa [40]	<u>0.763</u>	**0.805**	0.745	0.811	-	-
SHDIQA [18]	0.723	0.791	<u>0.782</u>	<u>0.819</u>	<u>0.789</u>	<u>0.943</u>
BWAFDA	**0.805**	<u>0.803</u>	**0.828**	**0.846**	**0.810**	**0.949**

Cross-Database Evaluation. Cross-dataset experiments are used to evaluate BWAFDA's ability for generalization, with the NR-IQA model being trained on one dataset and tested on others. The SRCC averages for five datasets are shown in Tab. 3. BWAFDA achieves the best performance in five out of six cross-dataset evaluations. BWAFDA attains the top SRCC on almost all train-test pairs. Compared with strong baselines (*e.g.*, LoDa [40], DEIQT [28], TReS [9]), our method achieves higher correlations consistently, often by 0.02-0.04 SRCC, and never collapses in any transfer direction. This indicates that our proposed model captures distortion cues that are less dependent on the specific content distribution of a dataset.

FLOPs/Inf. Time Comparison. To evaluate computational efficiency, we compare model complexity under a fixed 224×224 input resolution (see Tab. 4). Inference latency is measured on an NVIDIA RTX 5090 GPU (Windows 11) by averaging 100 forward passes after 30 warm-up iterations. Our BWAFDA contains only 23M parameters, achieving 14.9 GFLOPs and 6.6 ms per image. These results confirm a favorable accuracy–efficiency trade-off, underscoring the practical advantage of our approach.

4.3 Ablation Study

Ablation on Masked Prior. In the "Without Prior" variant in Tab. 5, we ablate the masked prior $Pr^{(s)} = \log P^{(s)}$ from the attention logits (see Eq. (18)).

Table 4: Complexity comp.

Method	Param. (M)	GFLOPs	Inf. Time (ms)
HyperIQA [34]	27	8.650	4.465
TReS [9]	152	39.910	10.276
MUSIQ [13]	27	43.469	16.857
DEIQT [28]	24	9.340	3.686
LIQE [49] W/O CLIP	151	8.817	3.082
LIQE [49] W CLIP	151	2958.530	40.907
LoDa [40]	118	47.269	8.696
Ours	23	14.885	6.595

This ablation leads to consistent performance drops compared with our full model. On the LIVEFB database, the PLCC score fell from 0.777 to 0.754, a decrease of 3.0%, while the SRCC score declined from 0.723 to 0.677, experiencing a 6.8% drop. Similarly, on the LIVEC database, the PLCC decreased from 0.928 to 0.908 (-2.2%), and the SRCC dropped from 0.917 to 0.893 (-2.6%). From a mechanical perspective, adding $\log P^{(s)}$ is equivalent to reweighting attention using a probability mask specific to the scale, which suppresses implausible token interactions and stabilizes rankings. These effects are particularly beneficial for the noisier, in-the-wild LIVEFB data. Overall, the masked prior acts as a lightweight, structure-aware regularizer that improves both absolute agreement (PLCC) and ordinal consistency (SRCC), and is therefore an important contributor to our model's robustness and SOTA performance.

Ablation on Multi-scale. The adoption of a three-scale architecture in our final model is justified by the ablation results (Tab. 5), where the single-scale ("Multiscale One") and two-scale ("Multiscale Two") variants lag behind the three-scale design. At the same time, the step from one to two scales offers negligible improvement (*e.g.*, SRCC: 0.677 $\rightarrow$ 0.678 on LIVEFB). By contrast, introducing the third scale brings pronounced gains: PLCC +0.022 (+2.91%) and SRCC +0.045 (+6.64%) on LIVEFB; PLCC +0.018 (+1.98%) and SRCC +0.022 (+2.46%) on LIVEC. This performance leap stems from (i) the multi-scale coverage of fine-to-coarse visual cues and (ii) our detail-aware, block-wise weighted fusion, which dynamically selects informative scales per region instead of relying on naive averaging. Consequently, the three-scale design materially strengthens both PLCC and SRCC.

Ablation on Detail-aware Branch. As shown in Tab. 5, removing the detail-aware branch from our full model leads to clear performance drops. The degradation is more pronounced on the in-the-wild LIVEC dataset. For example, PLCC decreases by 0.017 and SRCC by 0.022. The drop on LIVEFB is smaller but still consistent. These results suggest that modelling details, such as edges, textures, and local artifacts, are particularly important for real-world distortions. In such cases, global semantic cues alone are often ambiguous. Mechanistically,

Table 5: Ablation results on LIVEFB and LIVEC datasets.

Mode	LIVEFB		LIVEC	
	PLCC	SRCC	PLCC	SRCC
Without Detail-aware Branch	0.774	0.720	0.911	0.895
Without Prior	0.754	0.677	0.908	0.893
Multiscale One	0.753	0.677	0.908	0.892
Multiscale Two	0.755	0.678	0.910	0.895
Ours (BWAFDA)	**0.777**	**0.723**	**0.928**	**0.917**

the detail-aware branch provides high-frequency evidence. BWAFDA uses this evidence to emphasize the most informative scales dynamically. Therefore, the detail-aware branch complements contextual features effectively. It helps stabilize rank-order consistency and improves absolute score agreement.

4.4 Qualitative Analysis

We provide qualitative visualizations of the detail-aware branch feature maps (as shown in Fig. 3), where high responses concentrate on edges/high-frequency structures that typically correlate with perceptual distortions. In addition, Tab. 5 confirms its contribution quantitatively: removing the detail-aware branch leads to consistent drops in PLCC/SRCC, indicating it indeed contributes to the final distortion-aware prediction.

In Fig. 4, we show KonIQ examples with their block-wise importance maps. For well-predicted images (top row), the model assigns high importance to regions that are perceptually informative-such as textured areas, object boundaries, or local artifacts-while largely down-weighting flat backgrounds, leading to predictions that closely match the ground-truth MOS. For poorly-predicted images, importance sometimes becomes too diffuse or biased toward salient objects, so subtle or highly localized distortions are underweighted. Overall, the block-wise fusion is effective, and remaining mismatches mostly occur for visually complex or ambiguous images where quality assessment is inherently difficult.

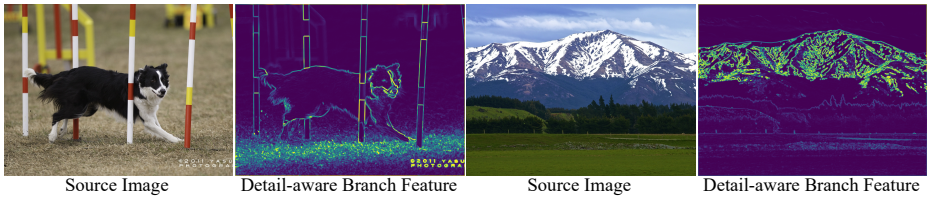

Fig. 3: Visualization of the detail-aware branch features. Compared with the source images, the extracted feature maps highlight fine-grained local structures, including edges, contours, textures, and other high-frequency details.

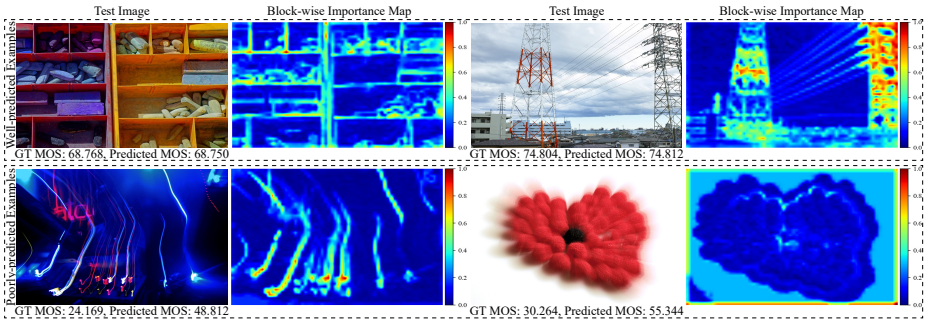

Fig. 4: Qualitative examples on KonIQ showing test images and corresponding block-wise importance maps. The top row shows well-predicted samples, where the predicted MOS is close to the ground truth, while the bottom row shows poorly predicted samples. Warmer colors indicate blocks that are more influential for quality prediction.

5 Conclusion

We introduced BWAFDA, an NR-IQA model that keeps the transformer's global view but adds two missing pieces: (1) a multi-scale, block-wise fusion that aggregates evidence where artifacts actually appear, and (2) a lightweight detail-aware branch that preserves high-frequency cues. A soft prior, $Pr^{(s)}$, guides cross-scale attention, and a class token head converts the fused block features into a single quality score, without the need for region labels. Experiments show three key findings. Firstly, local detail, multi-scale, and global context work together; removing any of these factors reduces performance. Secondly, our block-wise selective fusion compares favorably against global pooling baselines, particularly when the image contains small, scattered artifacts. Thirdly, the method remains data-efficient and can be transferred easily across datasets. These observations suggest two next steps: learning stronger priors from self-supervision and extending BWAFDA to video IQA and explainable region attribution. Overall, these results confirm the effectiveness of BWAFDA and its clear advantage over SOTA NR-IQA models.

Acknowledgements

The authors would like to thank NTU Smart Lab for providing the essential research infrastructure, computational resources, and a stimulating collaborative environment that greatly facilitated this work. We also appreciate the lab members for their valuable discussions and technical support.

References

1. Agarwal, S., Niyogi, P.: Stability and generalization of bipartite ranking algorithms. In: Proceedings of the International Conference on Computational Learning Theory. pp. 32–47 (2005)

2. Banham, M.R., Katsaggelos, A.K.: Digital image restoration. IEEE Signal Processing Magazine **14**(2), 24–41 (1997)
3. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of the 19th International Conference on Computational StatisticsParis. pp. 177–186 (2010)
4. Carnec, M., Le Callet, P., Barba, D.: Objective quality assessment of color images based on a generic perceptual reduced reference. Signal Processing: Image Communication **23**(4), 239–256 (2008)
5. Chen, C., Mo, J., Hou, J., Wu, H., Liao, L., Sun, W., Yan, Q., Lin, W.: Topiq: A top-down approach from semantics to distortions for image quality assessment. IEEE Transactions on Image Processing **33**, 2404–2418 (2024)
6. Chen, T., Jin, J., Cai, S., Li, Z., Lin, W.: Mugsqa: Novel multi-uncertainty-based gaussian splatting quality assessment method, dataset, and benchmarks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 11737–11741 (2026)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(2), 295–307 (2016)
8. Ghadiyaram, D., Bovik, A.C.: Massive online crowdsourced study of subjective and objective picture quality. IEEE Transactions on Image Processing **25**(1), 372–387 (2016)
9. Golestaneh, S.A., Dadsetan, S., Kitani, K.M.: No-reference image quality assessment via transformers, relative ranking, and self-consistency. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3989–3999 (2022)
10. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: Proceedings of the Advances in Neural Information Processing Systems. vol. 17, pp. 529–536 (2004)
11. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing **29**, 4041–4056 (2020)
12. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1733–1740 (2014)
13. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
14. Larson, E.C., Chandler, D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. Journal of Electronic Imaging **19**(1), 11006–11006 (2010)
15. Li, L., Song, T., Wu, J., Dong, W., Qian, J., Shi, G.: Blind image quality index for authentic distortions with local and global deep feature aggregation. IEEE Transactions on Circuits and Systems for Video Technology **32**(12), 8512–8523 (2022)
16. Li, X., Gao, T., Hu, R., Zhang, Y., Zhang, S., Zheng, X., Zheng, J., Shen, Y., Li, K., Liu, Y., et al.: Adaptive feature selection for no-reference image quality assessment by mitigating semantic noise sensitivity. In: Proceedings of the International Conference on Machine Learning. pp. 27808–27821 (2024)
17. Li, X., Hu, R., Zheng, J., Zhang, Y., Zhang, S., Zheng, X., Li, K., Shen, Y., Liu, Y., Dai, P., et al.: Integrating global context contrast and local sensitivity for blind

image quality assessment. In: Proceedings of the 41st International Conference on Machine Learning. pp. 27920–27941 (2024)
18. Li, X., Nie, W., Zhang, Y., Hu, R., Li, K., Zheng, X., Cao, L.: Distilling spatially-heterogeneous distortion perception for blind image quality assessment. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 2344–2354 (2025)
19. Li, X., Zhang, Y., Shen, Y., Li, K., Hu, R., Zheng, X., Zhao, S.: Feature denoising diffusion model for blind image quality assessment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5004–5012 (2025)
20. Li, Z., Jin, J., Cai, S., Lin, W.: R4-cgqa: Retrieval-based vision language models for computer graphics image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9468–9477 (2026)
21. Lin, H., Hosu, V., Saupe, D.: Kadid-10k: A large-scale artificially distorted iqa database. In: Proceedings of the Eleventh International Conference on Quality of Multimedia Experience. pp. 1–3 (2019)
22. Liu, H., Abbeel, P.: Blockwise parallel transformers for large context models. In: Proceedings of the Advances in Neural Information Processing Systems. vol. 36, pp. 8828–8844 (2023)
23. Liu, J., Li, X., Peng, Y., Yu, T., Chen, Z.: Swiniqa: Learned swin distance for compressed image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1795–1799 (2022)
24. Meng, L., Jin, J., Wang, Y., Wang, M., Lin, G., Liang, C., Sun, J., Zhang, H., Lin, W.: Boosting the no-reference image quality assessment via low-quality pseudo references. IEEE Transactions on Circuits and Systems for Video Technology (2026)
25. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012)
26. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2013)
27. Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al.: Image database tid2013: Peculiarities, results and perspectives. Signal Processing: Image Communication **30**, 57–77 (2015)
28. Qin, G., Hu, R., Liu, Y., Zheng, X., Liu, H., Li, X., Zhang, Y.: Data-efficient image quality assessment with attention-panel decoder. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2091–2100 (2023)
29. Qiu, J., Ma, H., Levy, O., Yih, W.t., Wang, S., Tang, J.: Blockwise self-attention for long document understanding. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP. pp. 2555–2565 (2020)
30. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena **60**(1-4), 259–268 (1992)
31. Saha, A., Mishra, S., Bovik, A.C.: Re-iqa: Unsupervised learning for image quality assessment in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5846–5855 (2023)
32. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Transactions on Image Processing **15**(2), 430–444 (2006)
33. Sheikh, H.R., Sabir, M.F., Bovik, A.C.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing **15**(11), 3440–3451 (2006)

34. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3664–3673 (2020)
35. Touvron, H., Cord, M., Jégou, H.: Deit iii: Revenge of the vit. In: Proceedings of the European Conference on Computer Vision. pp. 516–533 (2022)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems. vol. 30 (2017)
37. Wang, Z., Bovik, A.C.: A universal image quality index. IEEE Signal Processing Letters **9**(3), 81–84 (2002)
38. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
39. Wang, Z., Simoncelli, E.P.: Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In: Proceedings of the Human Vision and Electronic Imaging X. vol. 5666, pp. 149–159 (2005)
40. Xu, K., Liao, L., Xiao, J., Chen, C., Wu, H., Yan, Q., Lin, W.: Boosting image quality assessment through efficient transformer adaptation with local feature enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2662–2672 (2024)
41. Xue, W., Mou, X., Zhang, L., Bovik, A.C., Feng, X.: Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. IEEE Transactions on Image Processing **23**(11), 4850–4862 (2014)
42. Xue, W., Zhang, L., Mou, X., Bovik, A.C.: Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. IEEE Transactions on Image Processing **23**(2), 684–695 (2014)
43. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1190–1199 (2022)
44. Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., Bovik, A.: From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3575–3585 (2020)
45. You, J., Korhonen, J.: Transformer for image quality assessment. In: Proceedings of the IEEE International Conference on Image Processing. pp. 1389–1393 (2021)
46. Zhang, G., Jin, J., Liu, M., Yao, C., Lin, W.: Qd-pcqa: Quality-aware domain adaptation for point cloud quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17143–17152 (2026)
47. Zhang, L., Zhang, L., Bovik, A.C.: A feature-enriched completely blind image quality evaluator. IEEE Transactions on Image Processing **24**(8), 2579–2591 (2015)
48. Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology **30**(1), 36–47 (2020)
49. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14071–14081 (2023)

50. Zhang, Z., Sun, W., Min, X., Zhu, W., Wang, T., Lu, W., Zhai, G.: A no-reference deep learning quality assessment method for super-resolution images based on frequency maps. In: Proceedings of the IEEE International Symposium on Circuits and Systems. pp. 3170–3174 (2022)
51. Zhao, K., Yuan, K., Sun, M., Li, M., Wen, X.: Quality-aware pre-trained models for blind image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22302–22313 (2023)
52. Zhu, H., Li, L., Wu, J., Dong, W., Shi, G.: Metaiqa: Deep meta-learning for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14143–14152 (2020)
53. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proceedings of the IEEE **109**(1), 43–76 (2021)