BiRQA: Bidirectional Robust Quality Assessment for Images

Anonymous authors
Paper under double-blind review

Abstract

Full-Reference image quality assessment (FR IQA) is important for image compression, restoration and generative modeling, yet current neural metrics remain slow and vulnerable to adversarial perturbations. We present BiRQA, a compact FR IQA metric model that processes four fast complementary features within a bidirectional multiscale pyramid. A bottom-up attention module injects fine-scale cues into coarse levels through an uncertainty-aware gate, while a top-down cross-gating block routes semantic context back to high resolution. To enhance robustness, we introduce Anchored Adversarial Training, a theoretically grounded strategy that uses clean ”anchor” samples and a ranking loss to bound pointwise prediction error under attacks. On five public FR IQA benchmarks BiRQA outperforms or matches the previous state of the art (SOTA) while running $\sim 3\times$ faster than previous SOTA models. Under unseen white-box attacks it lifts SROCC from 0.30-0.57 to 0.60-0.84 on KADID-10k, demonstrating substantial robustness gains. To our knowledge, BiRQA is the only FR IQA model combining competitive accuracy with real-time throughput and strong adversarial resilience.

1 Introduction

Image Quality Assessment (IQA) is a fundamental problem in computer vision with applications in image restoration, compression, and generative modeling. Full-Reference (FR) IQA estimates perceived quality by comparing a distorted image with its pristine reference. While classical approaches like PSNR and SSIM are fast, they overlook many complex perceptual details, driving interest in deep learning approaches. Yet even strong neural IQA models still face two pressing issues: (i) slow inference speed that limits real-time use, and (ii) high vulnerability to adversarial perturbations, threatening reliability in safety-critical applications.

Adversarial attacks introduce imperceptible perturbations that mislead NN-based IQA models. Despite recent defenses for FR-IQA, robustness benchmarks show that many metrics remain vulnerable Gushchin et al. (2024), making them unsuitable for domains such as medical imaging, autonomous driving, and content authentication, where scores must remain trustworthy under both adversarial and benign perturbations. Moreover, these vulnerabilities allow attackers to manipulate image search results, as search engines like Bing rely on IQA metrics for ranking (Microsoft (2013)). Such attacks can also falsify public benchmark results(Huang et al. (2024); Wu et al. (2024)) by exploiting weaknesses in IQA models and artificially boosting perceived algorithm quality. For example, incorporating a vulnerable IQA metric as a loss function in image restoration can degrade actual image quality (Ding et al. (2021)) or cause visual artifacts (Kashkarov et al. (2024)). These risks highlight the urgent need for an FR IQA method that combines accuracy with adversarial robustness.

In this work, we present **BiRQA**: a precise, fast, compact, and attack-resilient FR IQA metric. The model builds a multiscale feature pyramid and injects feature maps that capture important patterns for human visual system (gradient structure, color dissimilarities, and local binary patterns) into a lightweight neural network. Information flows bidirectionally, which is generally a novel concept in IQA: a bottom-up attention lifts fine artifacts with an uncertainty-aware gate that outputs strength and confidence, reducing error propagation across scales. Then a top-down cross-gating supplies global context. This flow reduces scale-specific blind spots, yielding more precise quality scores across unseen distortions. A reliability-aware aggregation head pools each scale with GeM and combines per-scale contributions via softmax-normalized confidence weights, producing an interpretable convex combination. Reliability is strengthened through anchor-based adversarial training (AT) that fine-tunes the model to preserve the ranking of adversarial predictions with respect to clean anchors. Theoretical analysis links the anchor-based optimization objective to a maximal pointwise prediction error. Extensive experiments on standard IQA benchmarks show that BiRQA achieves superior accuracy while maintaining computational efficiency ($\sim$15 FPS on 1920×1080 images). Furthermore, our method generalizes across diverse distortion types and remains resilient to adversarial perturbations, outperforming existing methods in attack scenarios. The key contributions are:

- A novel FR IQA model architecture **BiRQA** uses bidirectional, uncertainty-aware cross-scale fusion with interpretable aggregation. The proposed CSRAM (fine→coarse) and SCGB (coarse→fine) modules exchange signals through learned gates, while a lightweight head aggregates scales with uncertainty-aware weights. The code will be publicly available.
- Theoretically grounded anchored AT uses clean anchors samples and a ranking loss to tighten a prediction error bound, boosting SROCC under attacks by up to 0.30 over the undefended model and 0.05 over the prior defenses.
- Extensive experiments on five public FR IQA benchmarks and four unseen white-box attacks show that BiRQA matches or surpasses previous SOTA metrics, runs $\sim 3\times$ faster than transformer methods, and improve integral robustness scores by up to 12%.

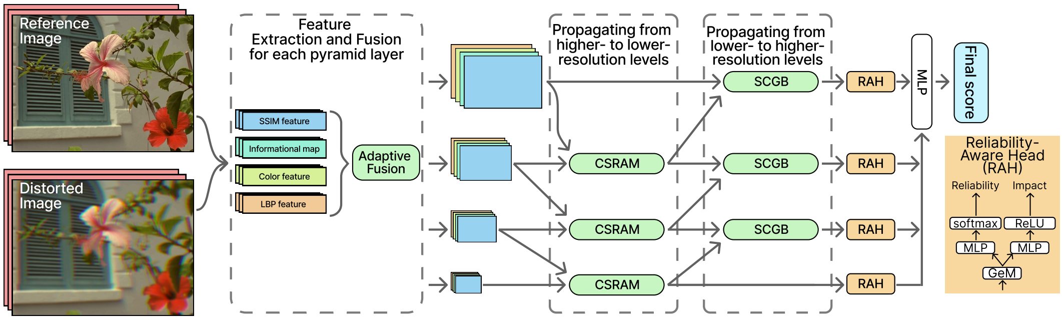
Reference Image
Distorted Image
Feature Extraction and Fusion for each pyramid layer
SSIM feature
Informational map
Color feature
LBP feature
Adaptive Fusion
Propagating from higher- to lower-resolution levels
Propagating from lower- to higher-resolution levels
CSRAM
SCGB
RAH
MLP
Final score
Reliability-Aware Head (RAH)
Reliability
Impact
softmax
ReLU
MLP
MLP
GeM

Figure 1: Overall scheme of BiRQA. A reference–distorted pair yields four feature maps per pyramid level. Cross-Scale Residual Attention Module (CSRAM) and Spatial Cross-Gating Block (SCGB) allow the model to pass information in both directions between scales. A Reliability-Aware Head (GeM + dual MLPs) estimates per-level impact and reliability.

2 Related Work

Full-Reference Image Quality Assessment. Assessing the quality of images is critical for numerous applications, such as compression, super-resolution, and other image processing techniques. FR IQA methods evaluate perceptual quality by comparing a distorted image to its pristine reference. While PSNR is fast, it correlates poorly with the Human Visual System (HVS). Metrics like SSIM (Wang et al. (2004)), MS-SSIM (Wang et al. (2003)), FSIM (Zhang et al. (2011)), SR-SIM (Zhang & Li (2012)), and VIF (Sheikh & Bovik (2006)) model structure, phase, saliency, or visibility to better match perception at low cost. These methods depend on heuristics and analytic features, ensuring computational efficiency. Deep learning further improves performance with models like LPIPS (Zhang et al. (2018)) and DISTS (Ding et al. (2020)). Transformers extend this via SwinIQA (Liu et al. (2022)), IQT (Cheon et al. (2021)), AHIQ (Lao et al. (2022)), and the current state-of-the-art (SOTA) method TOPIQ (Chen et al. (2024)), which adopts a multiscale top-down scheme with Cross-Self Attention. Several recent models further exploit multiscale attention. SwinIQA and IQT both address cross-scale information flow: SwinIQA relies on a heavy transformer pipeline, whereas proposed BiRQA model integrates lightweight CNN layers with perceptually grounded analytic features to retain speed. IQT passes representations from coarse-to-fine layers, while BiRQA's Cross-Scale Residual Attention Module (CSRAM) exchanges information bottom-up, yielding a bidirectional interaction that improves detail recovery.Beyond multiscale attention, some works explore learning to rank image quality. RankIQA (Liu et al. (2017)) trains a Siamese network to predict quality for NR IQA. To our knowledge, ranking has not yet been combined with adversarial training in FR IQA, a gap we address through the anchored ranking loss in BiRQA.

Robust IQA Methods. In FR IQA, an attack adds an imperceptible perturbation that deceives the metric. Attacks are classified as white-box, where gradients and parameters are known, or black-box, where only output scores are available (Chakraborty et al. (2021)). IQA-specific attacks have been proposed in Korhonen & You (2022); Shumitskaya (2024; 2023); Zhang et al. (2022).

Defenses divide into certified and empirical categories. While certified defenses offer provable guarantees, they remain too slow for real-time use. Empirical strategies are more practical and often revolve around adversarial training or input purification. There are some works focused on NR-IQA, including input purifications (E-LPIPS, Kettunen et al. (2019)), adversarial training (R-LPIPS (Ghazanfari et al. (2023)), AT (Chistyakova et al. (2024))) and architecture modification (Grad.Norm., Liu et al. (2024)). Although previous AT methods employ data augmentation or gradient regularization, our anchored adversarial training integrates a ranking loss to enhance robustness.

3 Method

FR IQA requires four properties still unmet by many deep learning models: (1) *accuracy*, (2) *low latency*, (3) *resilience to adversarial perturbations*, and (4) *sensitivity to multi-scale artifacts*. To meet these goals, we present BiRQA, a compact hybrid network that injects lightweight, human-interpretable feature maps into a bidirectional attention pyramid guided by HVS principles. Figure 1 sketches the pipeline: feature maps are arranged in a four-level pyramid, preserving fine detail at high resolution and summarizing global structure at lower resolutions. At each scale, an *Adaptive-Fusion* block reweights channels, a bottom-up *Cross-Scale Residual Attention Module* (CSRAM) lifts fine-scale cues to coarser levels, and a top-down *Spatial Cross-Gating Block* (SCGB) feeds semantic context back to higher resolutions, completing the bidirectional exchange. A Reliability-Aware Head (RAH) pools per-scale representations and aggregates them via normalized reliability weights to produce the final score. Training minimizes a PLCC-oriented regression loss together with an anchored ranking loss.

3.1 Feature Extraction

Feature selection was guided by two primary criteria: computational efficiency and the ability to detect complementary types of image degradation. To keep model fast we explored various lightweight analytic features rather than building complex image representations from raw images. We evaluated 11 candidate features, including Gabor filters, wavelet transform, entropy map, edge map, detail loss measure (Li et al. (2011)), VIF, and saliency. A total of 300+ feature combinations were tested. The following four features delivered the best accuracy–runtime trade-off; adding raw images as additional inputs to BiRQA worsened results. Full details are provided in Appendix B. The four chosen features address distinct aspects of quality degradation: (1) **SSIM map**: costs little to compute and provides a spatial implementation of the structural-similarity idea, which is consistent with how people compare distortions. (2) **Local informational content**: measures the variance of pixel intensities to estimate whether a region is highly informative. The selection of this feature was inspired by IW-SSIM (Wang & Li (2010)). (3) **YCbCr color difference map**: isolates chroma shifts and color bleeding in channels aligned with the HVS. (4) **Local Binary Patterns (LBP)**: compares each pixel with its neighboring pixels to encode local texture information into binary patterns. This method has proven effective under adversarial attacks (Asmitha et al. (2024)).

Unlike many recent IQA models that crop images to lower resolutions for faster computation and compatibility with pre-trained backbones, our approach avoids cropping to preserve global context and degradation-specific regions. Instead, we compute feature maps at the original resolution and integrate them into a pyramidal framework. This method uses four pyramid levels, each downscaled by a factor of two from the previous level, enabling the network to capture and aggregate multiscale degradation information effectively.

3.2 BiRQA Network

Following feature extraction, the BiRQA network processes these features to compute the final quality score. Each feature map at every pyramid level is first preprocessed individually to highlight significant spatial regions. These preprocessed feature tensors we denote as $\{F_i^j\}, F_i^j \in \mathbb{R}^{D_j \times H_i \times W_i}$, where H_i, W_i are feature map dimensions on i-th pyramid level, D_j denoting the number of channels for feature j.

Multi-feature Adaptive Fusion. At pyramid level i, we concatenate the four features, compute a joint attention vector $\alpha_i^j = \sigma\big(\mathrm{MLP}_i^j(\mathrm{GAP}(F_i^0 \oplus F_i^1 \oplus F_i^2 \oplus F_i^3))\big)$, and obtain the fused tensor $G_i = \phi_i(\bigoplus_{j=0}^{3} \alpha_i^j \odot F_i^j)$, where GAP is global average pooling, $\odot$ – element-wise multiplication and $\oplus$ – concatenation, ϕ_i – convolution. A Squeeze-and-Excitation block (Hu et al. (2018)) adaptively recalibrates G_i by "squeezing" spatial information into a channel descriptor via global pooling and then "exciting" (reweighting) each channel through a learned gating mechanism. This allows BiRQA to dynamically emphasize relevant feature channels and suppress noisy or redundant ones, improving robustness at negligible extra cost.

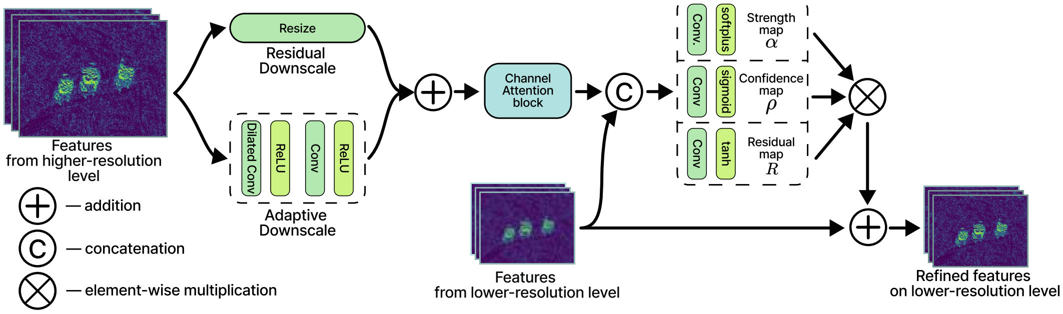

Figure 2: Scheme of the Cross-Scale Residual Attention Module (CSRAM) that lifts high-resolution cues to the next scale and injects them via uncertainty-aware gated residuals (strength α, confidence ρ, and residual R) to refine the lower-resolution features.

Naive resizing can blur high-frequency artifacts and propagate false positives; we transmit a *gated residual* whose *strength* and *confidence* are learned separately. We introduce **Cross-Scale Residual Attention Module** (**CSRAM**, Fig. 2) to interconnect scales. Fine-scale artifacts arise first; CSRAM lifts their cues upward while controlling reliability. We form a message $\hat{G}_{i+1} = \mathrm{Conv}\downarrow(G_i) + \mathrm{Resize}\downarrow(G_i)$. A channel attention block computes a spatial mask via channel pooling $M_{i+1} = \sigma\big(\mathrm{AvgPool}(\hat{G}_{i+1}) + \mathrm{MaxPool}(\hat{G}_{i+1})\big)$, and refine the message: $\hat{G}_{i+1} \leftarrow \hat{G}_{i+1} \odot M_{i+1}$. From $z_{i+1} = [G_{i+1}, M_{i+1}]$, we use three 1×1 projections ψ that yields: (i) a nonnegative injection *strength* map $\alpha_{i+1} = \mathrm{softplus}(\psi_\alpha(z_{i+1}))$, (ii) a bounded injection *confidence* map $\rho_{i+1} = \sigma(\psi_\rho(z_{i+1}))$, and (iii) feature map $R_{i+1} = \tanh(\psi_R(\hat{G}_{i+1}))$. Here, $\mathrm{softplus}$ enforces nonnegativity and $\tanh$ limits residual energy for stability. The final update is $G_{i+1} \leftarrow G_{i+1} + (\rho_{i+1} \odot \alpha_{i+1}) \odot R_{i+1}$. This uncertainty-aware gating mechanism is, to our knowledge, novel technique in FR IQA; it is parameter-light, tolerates small misalignments, and exposes interpretable reliability maps.

Spatial Cross-Gating Block (SCGB). Inspired by He et al. (2024), SCGB routes coarse context downward to suppress spurious high-resolution noise. For adjacent scales i (fine) and $i{+}1$ (coarse), we form $g_{i\leftarrow i+1} = \sigma(\mathrm{MLP}(G_{i+1}))$ and refine $G_i \leftarrow G_i + G_i \odot g_{i\leftarrow i+1}$. Together, upward CSRAM and downward SCGB provide a two-way highway that transfers only distortion evidence relevant to perceived quality.

Reliability-Aware Scale-Wise Fusion. BiRQA produces multi-scale feature tensors after SCGB, yet common fusion schemes (concatenations/MLPs or gating) hide cross-scale contributions and may be unstable across datasets. We introduce a light additive scale aggregation head that makes per-scale contributions explicit and learnable while keeping runtime overhead minimal. Post-SCGB tensors $G_i \in \mathbb{R}^{C\times H_i \times W_i}$ are fed into **Reliability-Aware Head (RAH)**: each scale is pooled to $z_i = \mathrm{GeM}(h_i(G_i)) \in \mathbb{R}^d$, where h_i is a 1×1 conv to a shared width d, and GeM is generalized-mean pooling with learnable exponent p. Two tiny MLPs produce a contribution $c_i \in \mathbb{R}$ and a reliability logit $a_i \in \mathbb{R}$. Denoting S as the number of scales, we form normalized gates $w_i = \mathrm{softmax}(a_0, \ldots, a_{S-1})$ to obtain $\hat{y} = \sum_{i=0}^{S-1} w_i c_i$, a convex, interpretable aggregation that is stable across datasets and cheap to compute.

3.3 Anchored Adversarial Training

We propose an adversarial fine-tuning strategy, AAT, that leverages adversarial examples without directly penalizing quality labels. Our approach relies on the assumption that adversarial perturbations are either imperceptible or only slightly visible, which is the most common case in real-world adversarial scenarios. A key idea is to leverage the availability of clean examples (images with reliably predicted quality) as “anchors“ for the training process.

Problem Formulation. Image-processing systems increasingly operate in untrusted environments where imperceptible perturbations can manipulate quality scores. We formalize an adversary that adds an ℓ_p-bounded noise δ ($\|\delta\|_p \leq \epsilon$) to the distorted image x_d of a pair (x_r, x_d). When higher scores denote better quality the attacker seeks to maximize $f_\theta(x_r, x_d + \delta)$; if lower scores indicate higher quality, the attacker aims to decrease it. The same approach could do the opposite task, as demonstrated in Antsiferova et al. (2024). This focus does not restrict the generality of our study, as the principles apply symmetrically. Formally, we define an adversarial attack as

$$\max_{\|\delta\|_p \leq \epsilon} \ f_\theta(x_r,\ x_d + \delta). \tag{1}$$

For a given model f_θ; training data D containing image pairs with associated quality label y; and a loss function $\mathcal{L}$ of the model, vanilla adversarial training is a min-max optimization problem(Chistyakova et al. (2024)):

$$\min_\theta \mathbb{E}_{(x_r, x_d, y)\sim D}\Big[\max_{\|\delta\|_p \leq \epsilon} \mathcal{L}(f_\theta(x_r, x_d + \delta), y)\Big]. \tag{2}$$

The inner maximization generates strong adversarial examples $x_d + \delta$, while the outer minimization adjusts the model parameters θ to improve robustness.

In image classification, the ground-truth label is unaffected by adversarial perturbations. Quality scores, however, *do change* slightly with perceptual content, so directly re-using equation 2 creates a label-shift. Prior work tackles this by penalizing or rescaling the label(Chistyakova et al. (2024)), but this brings two practical obstacles: (i) subjective studies follow diverse protocols, so matching MOS (Mean Opinion Score) scales must be repeated for every dataset; (ii) the metric that drives penalization can itself be vulnerable, enabling attack transfer.

3.3.1 Anchored ranking loss

To take advantage of reliable “anchors“, we define the anchored ranking loss:

$$\mathcal{L}_{\text{anchor}}(y, \tilde{y}) = \frac{1}{|S|\,N} \sum_{i\in S} \sum_{j=1}^{N} \frac{\max\{0,\ R(y,\tilde{y})_{i,j}\}}{\|R(y,\tilde{y})\|_\infty}, \tag{3}$$

where $R(y, \tilde{y})$ is an element-wise ranking matrix with elements $R(y, \tilde{y})_{i,j} = (y_i - y_j)\,\mathrm{sign}(\tilde{y}_j - \tilde{y}_i)$, N is the size of mini-batch, $|S|$ – number of anchors. Intuitively, this loss penalizes any deviation in the predicted ordering relative to the ground truth, emphasizing comparisons with clean samples. We can more effectively constrain the model's predictions on adversarial examples by anchoring the loss on a reliable subset S. Given the formulation above, we establish the following result:

Theorem 1: Pointwise Error Bound via Anchored Ranking Loss

Fix a mini-batch with ground-truth MOS $y = (y_1, \ldots, y_N)$ and predictions $\tilde{y}$. Let $\mathcal{S} \subseteq \{1, \ldots, N\}$ be the anchor indices and set

$$E = \max_i |\tilde{y}_i - y_i|, \quad R = \max_{i,j} |y_i - y_j|. \tag{4}$$

Assume constants $\lambda, \varepsilon, \Delta > 0$ exist such that: **(i) Anchor spacing**: $\Delta \leq y_{i_{k+1}} - y_{i_k} \leq \lambda$ for consecutive anchors $i_k, i_{k+1} \in \mathcal{S}$; **(ii) Anchor accuracy**: $|\tilde{y}_i - y_i| \leq \varepsilon$ for all $i \in \mathcal{S}$ Suppose that anchored ranking loss satisfies $L_{\text{anchor}} < \delta$. Then, the maximum pointwise error is bounded by

$$E \ \leq\ \varepsilon + \lambda\Big(1 + \sqrt{2\delta\, N\, R/\Delta}\Big) \ =\ \varepsilon + O\big(\lambda\sqrt{\delta}\big). \tag{5}$$

Proof Sketch. We isolate the sample with the largest prediction error E and split that error into the anchor-tolerated part $\varepsilon + \lambda$ and an overshoot τ. Because anchors are at most λ apart, inaccurate prediction for attacked sample by τ necessarily creates τ/λ rank inversions with other anchors. MOS values differ by at least Δ, so each inversion contributes a fixed positive amount to the anchored ranking loss; in total the loss grows like $(\tau/\lambda)^2$. Since the measured loss is bounded above by δ, this quadratic growth forces τ to scale no larger than $\lambda\sqrt{\delta\, N\, R/\Delta}$. Adding back the unavoidable $\varepsilon + \lambda$ part gives the claimed bound $E \le \varepsilon + \lambda\big(1 + O(\sqrt{\delta})\big)$. The full proof is provided in the Appendix A.

Numerical example

For instance, consider a dataset which MOS values lie in the interval $[0, 100]$, a nearly perfect IQA model ($\epsilon = 0.1$), and a batch size of $N = 16$ images. If, during the construction of each training batch, we do that in such a manner that the distance between neighboring anchors does not exceed $\lambda = 0.5$ ($R \le \lambda * N = 16$) and the anchored-ranking loss is $\delta = 0.01$, then the our guaranteed bound for quality score becomes less than 2% of the MOS range: $\big|\tilde{y}_j - y_j\big| \;\le\; 0.1 \;+\; 0.5 \times 3.57 \;\approx\; 1.88$

Theorem 1 requires only (i) an anchor spacing upper-bound λ — trivially satisfied by sorting a mini-batch and selecting evenly spaced samples; and (ii) an anchor-accuracy tolerance ϵ that is automatically met for a well-trained model on clean images. No dataset-specific statistics, MOS rescaling, or certified bounds are needed, so the conditions hold for any FR IQA corpus used in practice. Further details on the empirical satisfaction of these assumptions and on loss convergence are provided in Appendix C.2.

3.4 Implementation Details

For training BiRQA model, we used Adam optimizer with $lr = 10^{-4}$, batch size 32, and the following loss function:

$$\mathcal{L}(y, \hat{y}) = \alpha MSE(y, \hat{y}) - (1 - \alpha) PLCC(y, \hat{y}); \alpha = 0.7. \quad (6)$$

During adversarial training each mini-batch is drawn from a narrow MOS band, which limits both the label spread R and the anchor spacing λ. Within this band a random subset of images is left clean to serve as anchors while the remaining images are attacked. The network already predicts clean images well, so the anchor error ϵ stays negligible. Because R and λ are both bounded by the band width, the bound in Theorem 1 simplifies to a term driven mainly by λ. Narrower bands therefore tighten the guarantee and speed up convergence. Adversarial fine-tuning procedure outlined in Algorithm 1, where $\mathcal{L}_{AAT}$:

$$\mathcal{L}_{AAT} = \frac{1}{2}\mathcal{L}_{anchor} + \frac{1}{2}\mathcal{L}. \quad (7)$$

This composite loss enforces robust relative ordering under adversarial perturbations while still constraining absolute prediction error.

The complete set of parameters used for BiRQA and adversarial training procedure is listed in Appendix E.

Algorithm 1 Anchored Adversarial Fine-Tuning

```
Require: Network $f_\theta$, training set $\mathcal{D}$, mini-batch size $N$, attack routine $A$, perturbation budget $\epsilon$, desired
  anchor spacing $\lambda$
  while not converged do
      Sample mini-batch $\mathcal{B} = \{(x_r^k, x_d^k, y^k)\}_{k=1}^N$ from $\mathcal{D}$ following the procedure described in Sec. 3.4.1
      Sort $\mathcal{B}$ by $y^k$ and choose an index set $S$ s.t. $|y_{i_{k+1}} - y_{i_k}| \le \lambda$
      for $k = 1, \dots, N$ do
          $\hat{x}_d^k \leftarrow A(x_r^k, x_d^k;\, f_\theta, \epsilon)$                              ▷ solve equation 1
          $\tilde{y}^k \leftarrow f_\theta(x_r^k, \hat{x}_d^k)$
      Compute loss $\mathcal{L}_{\text{AAT}}(y^k, \tilde{y}^k)$ via equation 7
      $\theta \leftarrow \textsc{Adam}\big(\theta, \nabla_\theta \mathcal{L}_{\text{AAT}}\big)$
```

3.4.1 Mini-batch construction procedure

In all AAT-BiRQA experiments we fix the number of anchors per mini-batch to $|S| = 8$ with batch size $N = 16$. Let $MOS_r = max(MOS) - min(MOS)$ be the dataset MOS range; we set $R := \frac{MOS_r}{10} * \frac{|S|}{N} = \frac{MOS_r}{20}$. For each batch, we first sample a contiguous MOS band of width R: we draw a "low" MOS value y_{low} and form the band $[y_{low}, y_{low} + R]$. We then construct the anchors as follows: we always include the samples closest to y_{low} and $y_{low} + R$ as anchors and then select the remaining $|S| - 2$ anchors uniformly along the MOS axis inside the chosen band. By construction, the difference between any two consecutive anchors is at most λ, which satisfies the anchor-spacing assumption in Theorem 1. Then we sample the remaining $N - |S|$ non-anchor samples. This yields a approximately uniform anchor coverage along the MOS axis while satisfying the spacing condition in our theory.

4 Evaluation setup

Datasets. To compare our approach with current solutions, we provide various experiment results. We conducted intra- and cross-dataset evaluations, thorough robustness comparison, and an ablation study. As shown in Table 1, we conduct experiments on several public datasets: LIVE(Sheikh et al. (2006)), CSIQ (Larson & Chandler (2010)), TID2013 (Ponomarenko et al. (2013)), KADID-10k (Lin et al. (2019)), PIPAL (Jinjin et al. (2020)), and two-alternative forced choice (2AFC) dataset: BAPPS (Zhang et al. (2018)). When available, we used official splits for train/val/test parts and the mean value for 10 runs on random splits in 6:2:2 proportion. These splits are based on reference images to prevent content overlap.

Table 1: FR datasets used in experiments.

Dataset	Resolution	Ref. Images	Dist. Images	Ratings
LIVE	768 × 512	29	779	25k
CSIQ	512 × 512	30	866	5k
TID2013	512 × 384	25	3,000	524k
KADID-10k	512 × 384	81	10,125	30.4k
PIPAL	288 × 288	250	23,200	1.13M
PieAPP	256 × 256	200	20,280	1M+

Evaluation Metrics. We evaluate performance using two widely accepted correlation metrics for datasets with MOS values: Pearson's Linear Correlation Coefficient (PLCC) and Spearman's Rank-Order Correlation Coefficient (SROCC). Both metrics are in the range [-1, 1], with a positive value meaning a positive correlation. A larger SROCC indicates a more accurate ranking ability of the model, while a larger PLCC indicates a more accurate fitting ability of the model. We also use a paired bootstrap test (1k resamples) to assess if differences in SROCC are statistically significant.

Adversarial Robustness Comparison Methodology. We compare the effectiveness of the proposed adversarial training method for the BiRQA proposed model and for LPIPS models to keep consistency with previous works. KADID-10k train and test parts were used for adversarial training and testing, utilizing the Projected Gradient Descent with 10 iterations (PGD-10, Madry et al. (2017)) as attack method. The attack budget was $\epsilon = 8/255$. We compare different adversarial training techniques by SROCC and Integral Robustness Score (IR-Score, Chistyakova et al. (2024)) on clean and attacked data.

The IR-Score assesses the model's ability to withstand perturbations of varying strengths, as recommended in Carlini et al. (2019). Adversarial examples were generated with perturbation magnitudes $\epsilon \in \mathcal{E} = \{2, 4, 8, 10\}/255$. Scores were normalized using min-max scaling and mapped to a unified domain via neural optimal transport to account for distributional differences. The IR-Score is calculated as follows:

$$\text{IRscore} = R_f - R_{\hat{f}}, \text{ where } R_g = \frac{1}{N|\mathcal{E}|} \sum_{i=1}^{N} \sum_{\epsilon \in \mathcal{E}} \Big(g(x_r^{(i)}, x_d^{(i)}) - g(x_r^{(i)}, x_d^{(i)} + \delta_{g,\epsilon}^{(i)}) \Big). \tag{8}$$

Here $(x_r^{(i)}, x_d^{(i)})$ are reference/distorted images, $\delta_{g,\epsilon}^{(i)}$ is a perturbation at budget ϵ crafted against model $g \in \{f, \hat{f}\}$, $\hat{f}$ is the adversarially trained variant, and N is the number of images. Larger values indicate greater average reduction in attack-induced drop versus the baseline.

Table 2: Cross-dataset performance on benchmarks (PLCC and SROCC are reported in corresponding columns for each benchmark). The best values are bolded, and the second best are underlined.

Method	Trained on **KADID-10k**						Trained on **PIPAL**					
	LIVE		CSIQ		TID2013		LIVE		CSIQ		TID2013	
WaDIQaM-FR Bosse et al. (2017)	0.940	0.947	0.901	0.909	0.834	0.831	0.895	0.899	0.834	0.822	0.786	0.739
PieAPP Prashnani et al. (2018)	0.908	0.919	0.877	0.892	0.859	0.876	—	—	—	—	—	—
LPIPS-VGG Zhang et al. (2018)	0.934	0.932	0.896	0.876	0.749	0.670	0.901	0.893	0.857	0.858	0.790	0.760
DISTS Ding et al. (2020)	0.954	0.954	0.928	0.929	0.855	0.830	0.906	0.915	0.862	0.859	0.803	0.765
AHIQ Lao et al. (2022)	0.952	0.970	0.955	0.951	0.889	0.885	0.903	0.920	0.861	0.865	0.804	0.763
TOPIQ Chen et al. (2024)	0.957	0.974	0.963	**0.969**	0.916	0.915	**0.913**	**0.939**	0.908	0.908	0.846	0.816
BiRQA (ours)	**0.967**	**0.977**	**0.966**	0.967	**0.925**	**0.921**	0.911	0.933	**0.913**	**0.912**	**0.855**	**0.824**
AAT-BiRQA (ours)	0.962	0.970	0.961	0.962	0.918	0.917	0.909	0.925	0.909	0.910	0.850	0.817

Table 3: Robustness of adversarially trained IQA models on KADID-10k. SROCC is evaluated at $\epsilon = 8/255$; IR-Score uses $\epsilon \in \{2, 4, 8, 10\}/255$. Bold numbers denote the best result for each model. All adversarial variants were trained only with PGD-10.

Model	SROCC↑					IR-Score↑				Train time
	Clean	FGSM	C&W	AutoAttack	FACPA	FGSM	C&W	AutoAttack	FACPA	(min)
base LPIPS	**0.893**	0.542	0.260	0.239	0.496	—	—	—	—	**46**
R-LPIPS	0.858	0.570	0.327	0.266	0.515	0.541	0.403	0.385	0.507	123
AT-LPIPS	0.852	0.730	0.523	0.481	0.753	0.722	0.596	0.510	0.613	101
AAT-LPIPS (ours)	0.865	**0.753**	**0.589**	**0.536**	**0.810**	**0.783**	**0.652**	**0.582**	**0.630**	345
base TOPIQ	**0.938**	0.524	0.284	0.269	0.512	—	—	—	—	**352**
R-TOPIQ	0.879	0.533	0.305	0.314	0.509	0.572	0.431	0.456	0.520	437
AT-TOPIQ	0.892	0.839	0.513	0.552	0.760	0.763	0.615	0.550	0.611	514
AAT-TOPIQ (ours)	0.917	**0.844**	**0.593**	**0.580**	**0.811**	**0.802**	**0.691**	**0.601**	**0.645**	558
base BiRQA	**0.954**	0.568	0.295	0.350	0.503	—	—	—	—	**105**
R-BiRQA	0.902	0.571	0.291	0.357	0.502	0.563	0.427	0.414	0.525	218
AT-BiRQA	0.907	0.801	0.573	0.560	0.788	0.769	0.638	0.551	0.620	205
AAT-BiRQA (ours)	0.943	**0.836**	**0.614**	**0.602**	**0.819**	**0.811**	**0.690**	**0.614**	**0.657**	267

5 Results

Full Reference Benchmarks. Across standard FR-IQA benchmarks (LIVE, CSIQ, TID2013, PieAPP, PIPAL), BiRQA matches or surpasses prior SOTA in both PLCC and SROCC (e.g., LIVE 0.989/0.988, CSIQ 0.981/0.979; see Tab. 5 in Appendix due to limited space). On PieAPP, correlation is slightly lower, likely due to its broader distortion coverage. Per-distortion analysis shows SROCC ≈0.90-0.95 for most categories, with reduced effectiveness on some specific types such as radial geometric transforms. To assess the generalization capabilities of the proposed model, we performed cross-dataset evaluations. The model was trained on large KADID-10k and PIPAL datasets and tested on LIVE, CSIQ, and TID2013 datasets. The results are presented in Table 2. The proposed base BiRQA performs comparably to the TOPIQ model and exceeds it in 9 out of 12 experiments, highlighting its robust generalization across diverse datasets. AAT-BiRQA performs slightly worse than vanilla BiRQA, but, nonetheless, keeps up with the SOTA performance. This demonstrates that our adversarial training achieves robustness with negligible cost to normal-case performance – a key advantage over many defenses.

Adversarial Robustness Comparison. We benchmark four variants of three full-reference IQA (LPIPS, TOPIQ and the proposed BiRQA) against common ℓ_∞ white-box attacks. Compared methods are: (1) **base**: no adversarial training; (2) **R-**: vanilla adversarial training; (3) **AT-**: adversarial training with label smoothing (Chistyakova et al. (2024)); (4) **AAT-** (ours): Anchored Adversarial Training. All AT and AAT models were optimized with a PGD-10 attack of budget $\epsilon = 8/255$. At test time we evaluate four unseen attacks: FGSM(Goodfellow et al. (2014)), C&W(Carlini & Wagner (2017)), AutoAttack (AA, Croce & Hein (2020)) and the perceptual FACPA attack (Shumitskaya (2023)). Robustness is measured by SROCC and IR-Score on the KADID-10k dataset.

Table 3 presents the results, which shows that AAT achieves state-of-the-art robustness and outperforms other approaches by 0.02-0.06 SROCC points and by similar margins on IR-Score. AAT also provides the best SROCC on unperturbed test set compared to other defense methods. Even without defense, BiRQA possesses more robustness compared to both LPIPS and TOPIQ, surpassing them by $0.02 - 0.03$ in terms of IR-Score. Although AAT requires more time during adversarial fine-tuning phase, it provides the best overall results. More experiments can be found in the Appendix C.3-C.4, including experiments on 2AFC datasets under White-Box and Black-Box attacks.

Theoretical bounds in practice. Our adversarial training uses an anchored ranking loss that ties each perturbed sample to a small set of clean anchor images within the mini-batch. Theorem 1 establishes that as this loss approaches zero, the maximum prediction error on any adversarial example is provably bounded by a small constant. In practice, the empirical errors respect this bound, as shown in Fig. 6b in Appendix. Moreover, the objective drops below 10^{-2} within 500 iterations, indicating stable, fast optimization (see Fig. 6a in Appendix).

Statistical significance was tested on the PIPAL by a paired bootstrap of SROCC differences (1k resamples). As Table 7 in Appendix shows, BiRQA exceeds every previous FR IQA metric, gaining up to 0.57 SROCC over PSNR and a positive but very small ∼0.003 over the strongest baselines AHIQ and TOPIQ (statistically significant on PIPAL due to the large sample size, though the margin is small in absolute terms), while the robust variant AAT-BiRQA sacrifices only 0.007.

Computational Complexity. We assessed the computational efficiency of each IQA model by executing 100 forward passes on 100 random images from the PDAP-HDDS(Liu et al. (2018)) dataset and averaging the resulting runtimes. Experiments were performed on an NVIDIA A100 GPU (80 GB). Measurements are end-to-end: every operation required by model is considered, including the feature calculation for BiRQA.

Figure 3 summarizes the trade-off between accuracy and efficiency. Our method achieves performance on par with TOPIQ while running substantially faster than all competing approaches. Figure 3 also reports the number of parameters for each model. BiRQA has 5.5M parameters, which is smaller than most of its counterparts, including LPIPS and TOPIQ.

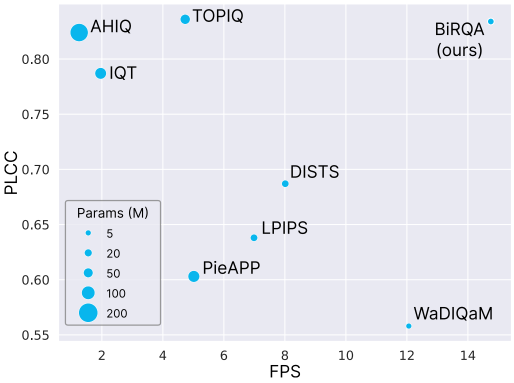

Figure 3: Computational efficiency (FPS) vs. Performance (PLCC) comparison on PDAP-HDDS dataset with image size of 1920 × 1080 pixels.

Ablation Study and Feature Selection. We exhaustively trained 231 candidate models covering every combination of 1-, 2-, and 3-feature sets drawn from an 11-feature pool on KADID-10k. Each candidate was scored by a Pareto trade-off between SROCC and inference speed. We dropped the three weakest features and evaluated all 70 four-feature sets from the remaining eight. Validation SROCC flattened at four features; adding a fifth would increase inference cost without accuracy gains. The final combination (SSIM, Informational Map, Color Difference, and LBP) shows the best SROCC while remains relatively fast and captures complementary structure, information content, chromatic shifts, and fine texture cues. A model that uses only raw image pairs outperforms any single analytic feature map, but adding raw image input to the chosen four lowers SROCC, indicating the Color Difference map already embeds the raw signal. Complete ablations can be found in Appendix B.

Figure 4: Ablation study for BiRQA model. The CSRAM and SCGB modules were replaced with cross-attention layers and element-wise multiplication. The Reliability-Aware Head (RAH) module was replaced with pooling and MLP.

CSRAM	SCGB	RAH	PLCC	SROCC
✗	✗	✗	0.801	0.813
✓	✗	✗	0.907	0.911
✓	✓	✗	0.925	0.928
✓	✓	✓	0.938	0.942

Table 4 presents results on the KADID-10k dataset for the variation of the BiRQA model. Enabling CSRAM lifts correlations markedly, highlighting that cross-scale interchange with uncertainty-aware gating is crucial for capturing fine artifacts without losing global context. Adding SCGB then activates fully bidirectional information flow between scales and yields a further, consistent improvement, while the reliability-aware head consolidates these signals and sharpens calibration.

6 Conclusion

In this work, we introduce BiRQA, a novel Full-Reference IQA metric that balances state-of-the-art performance while maintaining computational efficiency. It combines analytic feature maps with a lightweight neural network architecture, that incorporates a multi-scale pyramid framework, adaptive fusion and cross-scale attention mechanisms. This design captures multi-scale perceptual differences, achieving an inference speed of 15 FPS on 1920×1080 images, surpassing many counterparts. Extensive evaluations confirm that BiRQA matches or exceeds leading methods.

To address the critical challenge of adversarial robustness, we propose Anchored Adversarial Training (AAT) with an anchored-ranking loss and prove a mini-batch pointwise-error bound under mild assumptions. Empirically, the anchored loss drops quickly and remains low, and prediction errors stay within the theoretical bound. AAT delivers clear robustness gains over other defenses across four unseen attacks, with only a small drop on clean data. The trends persist when summarized by IR-Score across budgets. Limitations include weaker performance on some distortions (e.g., radial geometry). We hope these findings encourage further work into robust, efficient IQA models.

7 Reproducibility statement

We will release the full training/evaluation code, pre-trained checkpoints, and an environment file (Conda environment.yml) in the supplementary and on a public repository upon acceptance. Our experiments use standard FR-IQA datasets—LIVE, CSIQ, TID2013, KADID-10k, PIPAL (and BAPPS for 2AFC)—with official splits when available; otherwise we report the mean over 10 runs using 6:2:2 reference-image-based splits to avoid content leakage.

We provide:

- Code & config: scripts for clean training and Anchored Adversarial Training (AAT), evaluation on clean and attacked sets, IR-score computation.
- Hyperparameters: all values needed to reproduce results (optimizer, batch size, learning rate, loss weights, AAT mixing, PGD settings, ϵ budgets).
- Checkpoints & logs: trained weights and per-split predictions.

References

Anastasia Antsiferova, Khaled Abud, Aleksandr Gushchin, Ekaterina Shumitskaya, Sergey Lavrushkin, and Dmitriy Vatolin. Comparing the robustness of modern no-reference image-and video-quality metrics to adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 700–708, 2024.

KA Asmitha, Vinod Puthuvath, KA Rafidha Rehiman, and SL Ananth. Deep learning vs. adversarial noise: a battle in malware image analysis. *Cluster Computing*, 27(7):9191–9220, 2024.

Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45, 2021.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 2024.

Manri Cheon, Sung-Jun Yoon, Byungyeon Kang, and Junwoo Lee. Perceptual image quality assessment with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 433–442, 2021.

Anna Chistyakova, Anastasia Antsiferova, Maksim Khrebtov, Sergey Lavrushkin, Konstantin Arkhipenko, Dmitriy Vatolin, and Denis Turdakov. Increasing the robustness of image quality assessment models through adversarial training. *Technologies*, 12:220, 11 2024. doi: 10.3390/technologies12110220.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, 129(4):1258–1281, 2021.

Sara Ghazanfari, Siddharth Garg, Prashanth Krishnamurthy, Farshad Khorrami, and Alexandre Araujo. R-lpips: An adversarially robust perceptual similarity metric. *arXiv preprint arXiv:2307.15157*, 2023.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Alexander Gushchin, Khaled Abud, Georgii Bychkov, Ekaterina Shumitskaya, Anna Chistyakova, Sergey Lavrushkin, Bader Rasheed, Kirill Malyshev, Dmitriy Vatolin, and Anastasia Antsiferova. Guardians of image quality: Benchmarking defenses against adversarial attacks on image quality metrics, 2024. URL https://arxiv.org/abs/2408.01541.

Chenlong He, Qi Zheng, Ruoxi Zhu, Xiaoyang Zeng, Yibo Fan, and Zhengzhong Tu. Cover: A comprehensive video quality evaluator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5799–5809, 2024.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 633–651. Springer, 2020.

Egor Kashkarov, Egor Chistov, Ivan Molodetskikh, and Dmitriy Vatolin. Can no-reference quality-assessment methods serve as perceptual losses for super-resolution?, 2024.

Markus Kettunen, Erik Härkönen, and Jaakko Lehtinen. E-lpips: robust perceptual image similarity via random transformation ensembles. *arXiv preprint arXiv:1906.03973*, 2019.

Jari Korhonen and Junyong You. Adversarial attacks against blind image quality assessment models. In *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, pp. 3–11, 2022.

Shanshan Lao, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang. Attentions help cnns see better: Attention-based hybrid image quality assessment network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1140–1149, 2022.

Eric Larson and Damon Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *J. Electronic Imaging*, 19:011006, 01 2010. doi: 10.1117/1.3267105.

Songnan Li, Fan Zhang, Lin Ma, and King Ngi Ngan. Image quality assessment by separately evaluating detail losses and additive impairments. *IEEE Transactions on Multimedia*, 13(5):935–949, 2011. doi: 10.1109/TMM.2011.2152382.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3, 2019. doi: 10.1109/QoMEX.2019.8743252.

Jianzhao Liu, Xin Li, Yanding Peng, Tao Yu, and Zhibo Chen. Swiniqa: Learned swin distance for compressed image quality assessment. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 1795–1799, 2022.

Tsung-Jung Liu, Hsin-Hua Liu, Soo-Chang Pei, and Kuan-Hsien Liu. A high-definition diversity-scene database for image quality assessment. *IEEE Access*, 6:45427–45438, 2018.

Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*, pp. 1040–1049, 2017.

Yujia Liu, Chenxi Yang, Dingquan Li, Jianhao Ding, and Tingting Jiang. Defense against adversarial attacks on no-reference image quality models with gradient norm regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25554–25563, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Microsoft. A behind the scenes look at how bing is improving image search quality. https://blogs.bing.com/search-quality-insights/2013/08/23/a-behind-the-scenes-look-at-how-bing-is-improving-image-search-quality, 2013.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

Nikolay Ponomarenko, O. Ieremeiev, Vladimir Lukin, Karen Egiazarian, Lina Jin, J. Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Color image database tid2013: Peculiarities and preliminary results. In *2013 4th European Workshop on Visual Information Processing, EUVIP 2013*, pp. 106–111, 06 2013.

Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817, 2018.

Hamid Sheikh, Muhammad Sabir, and Alan Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 15:3440–51, 12 2006. doi: 10.1109/TIP.2006.881959.

Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.

Dmitriy Vatolin Shumitskaya, Anastasia Antsiferova. Fast adversarial cnn-based perturbation attack on no-reference image- and video-quality metrics. In *The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023, Kigali, Rwanda, May 5, 2023*. OpenReview.net, 2023.

Dmitriy Vatolin Shumitskaya, Anastasia Antsiferova. Towards adversarial robustness verification of no-reference image-and video-quality metrics. *Computer Vision and Image Understanding*, 240: 103913, 2024.

Zhou Wang and Qiang Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on image processing*, 20(5):1185–1198, 2010.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25456–25467, 2024.

Lin Zhang and Hongyu Li. Sr-sim: A fast and high performance iqa index based on spectral residual. In *2012 19th IEEE international conference on image processing*, pp. 1473–1476. IEEE, 2012.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Weixia Zhang, Dingquan Li, Xiongkuo Min, Guangtao Zhai, Guodong Guo, Xiaokang Yang, and Kede Ma. Perceptual attacks of no-reference image quality models with human-in-the-loop. *Advances in Neural Information Processing Systems*, 35:2916–2929, 2022.

APPENDIX

A PROOF OF THEOREM 1

Notation. Let $y = (y_1, \dots, y_N) \in \mathbb{R}^N$ be ground-truth mean opinion scores (MOS) and $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_N)$ be the model predictions on the current mini-batch. Let $\mathcal{S} \subseteq \{1, \dots, N\}$ be the *anchor* indices, which denote clean samples in mini-batch. The remaining samples are attacked.

Denote the element-wise ranking matrix $R_{i,j}(y, \tilde{y}) = (y_i - y_j)\,\mathrm{sign}(\tilde{y}_j - \tilde{y}_i)$, and write $\|R\|_\infty := \max_{i,j} |R_{i,j}|$. We define the anchored ranking loss as

$$L_{\text{anchor}}(y, \tilde{y}) = \frac{1}{|\mathcal{S}|\,N} \sum_{i \in \mathcal{S}} \sum_{j=1}^{N} \frac{\max\{0, R_{i,j}(y, \tilde{y})\}}{\|R\|_\infty}. \tag{9}$$

Let

$$E := \max_{1 \le j \le N} \big|\tilde{y}_j - y_j\big| \quad \text{and} \quad R := \max_{i,j} \big|y_i - y_j\big|.$$

Assumption. There exist a mini-batch such that for some constants $\lambda, \varepsilon, \Delta > 0$:

1. Anchor density. The anchor MOS values are sorted and successive anchors differ by at most λ:

$$\Delta < y_{i_{k+1}} - y_{i_k} \le \lambda \quad \forall\, i_k \in \mathcal{S}.$$

2. Anchor accuracy. $\big|\tilde{y}_i - y_i\big| \le \varepsilon \quad \forall i \in \mathcal{S}$.
3. MOS resolution. $\big|y_i - y_j\big| \ge \Delta \quad \forall i \ne j$.
4. Lowest-MOS and highest-MOS samples in the mini-batch are included in the anchor set $\mathcal{S}$

Theorem 1 (Pointwise Error Bound via Anchored Ranking Loss). If the anchored loss satisfies $L_{\text{anchor}}(y, \tilde{y}) < \delta$, then

$$E \le \varepsilon + \lambda\Big(1 + \sqrt{\tfrac{2\delta\, N\, R}{\Delta}}\Big). \tag{10}$$

Proof. **1. Pick the worst sample.** Choose $j^\star = \arg\max_j |\tilde{y}_j - y_j|$ and set $e := \tilde{y}_{j^\star} - y_{j^\star}$, so $|e| = E$. The case $e < 0$ is symmetric; assume $e > 0$ henceforth.

2. Define the overshoot. Put

$$\tau := e - \varepsilon - \lambda.$$

If $e \le \varepsilon + \lambda$ then equation 10 holds trivially, so assume $\tau > 0$.

3. Count the mis-ordered anchors. Because consecutive anchor MOS are at most λ apart, every sample is within $\lambda/2$ of some anchor. Shifting the prediction of $j^\star$ by an additional $\tau + \lambda/2 > \lambda/2$ therefore flips its order with at least

$$m := \Big\lceil \tfrac{\tau}{\lambda} \Big\rceil \ge \tfrac{\tau}{\lambda}$$

anchors whose ground-truth MOS are strictly larger than $y_{j^\star}$. There is always at least one anchor because of Assumption 4.

4. Lower-bound the loss contribution. For each such anchor i we have $y_i > y_{j^\star}$ but $\tilde{y}_{j^\star} > \tilde{y}_i$, hence $R_{i,j^\star}(y, \tilde{y}) \ge y_i - y_{j^\star}$. By the MOS-resolution assumption every violated pair contributes at least $\Delta, 2\Delta, \dots, m\Delta$, so

$$\sum_{k=1}^{m} (y_{j^\star} + k\Delta - y_{j^\star}) = \Delta \frac{m(m+1)}{2} \ge \frac{\Delta\, m^2}{2}.$$

5. Relate to L_{anchor}. Only one column (that of $j^\star$) is used, therefore

$$L_{\text{anchor}} \ge \frac{\frac{\Delta\, m^2}{2}}{N\, R} = \frac{\Delta\, m^2}{2\, N\, R}.$$

6. Solve for τ. The assumption $L_{\text{anchor}} < \delta$ implies $\frac{\Delta\, m^2}{2\, N\, R} < \delta$. Substituting $m \geq \tau/\lambda$ gives $\tau^2 < \frac{2\delta\, N\, R}{\Delta}\, \lambda^2$.

7. Express e and E. Using $\tau = e - \varepsilon - \lambda$ and $E = |e|$ yields $e \leq \varepsilon + \lambda + \lambda\sqrt{\frac{2\delta\, N\, R}{\Delta}}$, which is exactly equation 10.

□

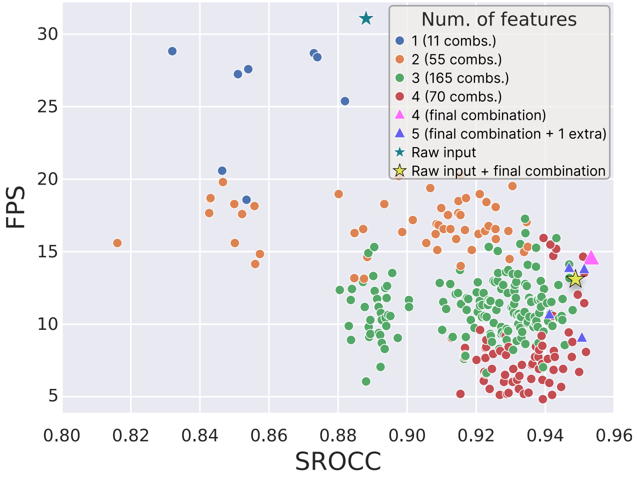

Figure 5: SROCC and inference FPS on KADID-10k for different feature sets. The chosen quartet lies on the Pareto frontier, offering the best accuracy. FPS was measured on images with 1920×1080 resolution.

B Choice of Features

We carefully designed a list of possible features for our model, including SSIM, Informational Map, Color Difference, Local Binary Pattern (LBP), Gabor Filters, Wavelet Transform, Entropy Map, Edge Map, Detail Loss Measure (DLM), Visual Information Fidelity (VIF) and Saliency Map. Short descriptions for these features are provided in Table 4.

We began with these 11 candidate analytic feature maps and exhaustively evaluated all 1-, 2-, and 3-feature combinations (11, 55, and 165 models, respectively) on KADID-10k (60/20/20 train/val/test split), measuring SROCC and inference speed (FPS). FPS was measured end-to-end (feature computation + the neural network) at 1920 × 1080 resolution on a single NVIDIA A100 80 GB GPU. Each model was trained from scratch under the same architecture and hyperparameters; ≈2 GPU-hours per run, totaling ≈600 GPU-hours. Based on joint accuracy–speed, the three weakest features (Saliency, Wavelet Transform, and Entropy Map) were removed. From the remaining eight features, we trained all $\binom{8}{4} = 70$ four-feature combinations. We did not explore all five-feature sets: their cost is prohibitive for real-time deployment and validation SROCC already saturates at four features (adding any additional feature to the best four-feature set yields no improvement). As a check, we added each of the remaining features to the best four-feature combination and tested them. The resulting SROCC values decreased upon adding a fifth feature, likely due to redundancy. In these experiments we varied only the input features to BiRQA, keeping the architecture fixed.

Figure 5 presents all 300+ evaluated feature combinations. The selected features SSIM, Informational Map (IM), Color Difference (CD), and LBP are complementary: SSIM captures structural deviations, IM reflects local information content, CD measures chromatic/luminance discrepancies in a perceptually motivated space, and LBP encodes fine-scale texture. A model using only raw image pairs outperforms any single analytic feature. However, concatenating raw pixels with the four selected features reduces SROCC. This likely stems from redundancy, because CD already encodes pixel level differences in an alternative color space, resulting in increased input dimensionality under a fixed capacity predictor, which can hurt generalization.

Table 4: Description of evaluated IQA features

Feature	**Description / Key Benefit**
SSIM	Structural Similarity Index compares luminance, contrast and structural components between a reference and a test image, yielding a single score that tracks perceived quality with high correlation to the human visual system (HVS).
Informational Map	Generates a spatial weighting map based on local information content (gradient magnitude), giving more influence to perceptually important, detail-rich regions.
Color Difference	Computes perceptual color difference in YCbCr color space, making the metric sensitive to chromatic distortions that luminance-only measures may miss.
LBP	Local Binary Patterns encode micro-texture by thresholding each neighborhood against its centre pixel; the histogram is gray-scale and rotation robust, providing a compact descriptor of fine texture changes.
Gabor Filters	A bank of Gabor kernels isolates edge and texture energy in specific frequency-orientation bands, capturing blur, ringing and other anisotropic artifacts.
Wavelet Transform	Discrete wavelet decomposition splits the image into multi-resolution subbands; analyzing coefficients across scales localizes blur or compression artifacts while preserving both spatial and frequency information.
Entropy Map	Computes local Shannon entropy inside sliding windows; high-entropy areas correspond to regions with greater visual information, enabling quality scores that prioritize complex, information-dense regions.
Edge Map	Gradient-based edge extraction detects intensity discontinuities and object boundaries; comparing edge strength between reference and distorted images is effective at spotting blur or sharpening artifacts.
Detail Loss Measure (DLM)	Measures the dissimilarity of high-frequency gradients between image pairs, providing a direct quantification of lost fine-scale detail (e.g., due to denoising, compression or over-smoothing).
VIF	Visual Information Fidelity models images as Gaussian Scale Mixtures and computes the mutual information lost in the distorted version, grounding the metric in natural-scene statistics and information theory.
Saliency	Uses the saliency neural network model to predict likely gaze locations, helping the final metric align with where observers are most likely to look.

C Additional Analysis and Results

C.1 More Results on FR Benchmarks

Table 5 presents a comprehensive comparison across widely used FR IQA benchmarks. The proposed BiRQA model achieves state-of-the-art or competitive results on most datasets, including LIVE, CSIQ, PieAPP, and the large-scale PIPAL dataset, for both PLCC and SROCC metrics. Notably, BiRQA reaches or surpasses a correlation of 0.98 on LIVE and CSIQ, and outperforms recent transformer-based methods on the more challenging PieAPP and PIPAL datasets while maintaining high efficiency.

AAT-BiRQA, which incorporates our proposed anchored adversarial training scheme, offers slightly lower correlations on clean data due to the regularization effect of adversarial robustness, but still maintains strong overall performance. This makes it preferable in safety-critical or attack-prone environments.

BiRQA shows slightly lower scores on TID2013, where it ranks just below the best-performing model (AHIQ). This can be attributed to the peculiar distortion types present in TID2013 such as chromatic aberration, mean shift, and severe radial distortions, that are underrepresented in modern training sets. Additionally, some distortions in TID2013 are known to interact poorly with analytic features (e.g., SSIM and LBP), which may limit BiRQA's ability to fully capture perceptual degradation in those cases. We hypothesize that deeper fine-tuning or explicit modeling of these artifact types may further improve performance on such legacy datasets.

Overall, BiRQA and its adversarially trained variant AAT-BiRQA demonstrate strong generalization across datasets and distortion types, validating the effectiveness of bidirectional multiscale fusion and anchor-based adversarial training.

Table 5: Quantitative comparison with related works on public FR benchmarks, including the traditional LIVE, CSIQ, TID2013 with MOS labels, and recent large-scale datasets PieAPP, PIPAL with 2AFC labels. The best and second results are bold and underlined, respectively, and "—" indicates the score is not available or not applicable.

Method	LIVE		CSIQ		TID2013		PieAPP		PIPAL	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
PSNR	0.865	0.873	0.819	0.810	0.677	0.687	0.135	0.219	0.277	0.249
SSIM (2004)	0.937	0.948	0.852	0.865	0.777	0.727	0.245	0.316	0.391	0.361
MS-SSIM (2003)	0.940	0.951	0.889	0.906	0.830	0.786	0.051	0.321	0.163	0.369
VIF (2006)	0.960	0.964	0.913	0.911	0.771	0.677	0.250	0.212	0.479	0.397
MAD (2010)	0.968	0.967	0.950	0.947	0.827	0.781	0.231	0.304	0.580	0.543
VSI (2014)	0.948	0.952	0.928	0.942	0.900	0.897	0.364	0.361	0.517	0.458
DeepQA (2017)	0.982	0.981	0.965	0.961	0.947	0.939	0.172	0.252	—	—
WaDIQaM (2017)	0.980	0.970	—	—	0.946	0.940	0.439	0.352	0.548	0.553
PieAPP (2018)	0.986	0.977	0.975	0.973	0.946	0.945	0.842	0.831	0.597	0.607
LPIPS-VGG (2018)	0.978	0.972	0.970	0.967	0.944	0.936	0.654	0.641	0.633	0.595
DISTS (2020)	0.980	0.975	0.973	0.965	0.947	0.943	0.725	0.693	0.687	0.655
JND-SalCAR (2020)	0.987	0.984	0.977	0.976	0.956	0.949	—	—	—	—
AHIQ (2022)	**0.989**	0.984	0.978	0.975	**0.968**	**0.962**	0.840	0.838	0.823	0.813
TOPIQ (2024)	0.984	0.984	0.980	0.978	0.958	0.954	0.849	0.841	0.830	0.813
BiRQA (ours)	**0.989**	**0.988**	**0.981**	**0.979**	0.964	0.959	**0.852**	**0.845**	**0.837**	**0.822**
AAT-BiRQA (ours)	0.984	0.980	0.980	0.975	0.958	0.952	0.840	0.830	0.831	0.811

C.2 Anchor-loss Convergence and Theoretical Bounds

Figure 6 (a) traces the anchored-ranking loss $\mathcal{L}_{anchor}$ over the first 1,000 optimization steps (mini-batches). Two curves are shown: the raw batch-wise loss (blue) and an exponential-moving-average with a window of 20 iterations (orange). By iteration ∼650 the smoothed loss drops below 10^{-3} and remains there for the rest of training, with only small mini-batch jitter ($< 10\%$ relative amplitude). This meets the target $\delta = 10^{-3}$ used in Theorem 1, so the theoretical bound on pointwise prediction error is already guaranteed after less than two epochs. The plot confirms that anchored adversarial training converges quickly and stably, delivering the tight loss levels required for the robustness guarantee without extended hyper-parameter tuning.

Figure 6 (b) compares the theoretical bound of Theorem 1 with the observed maximum pointwise error during AAT fine-tuning on KADID-10k. The empirical curve never exceeds the bound and decays at the same exponential rate, giving concrete evidence that the bound is valid.

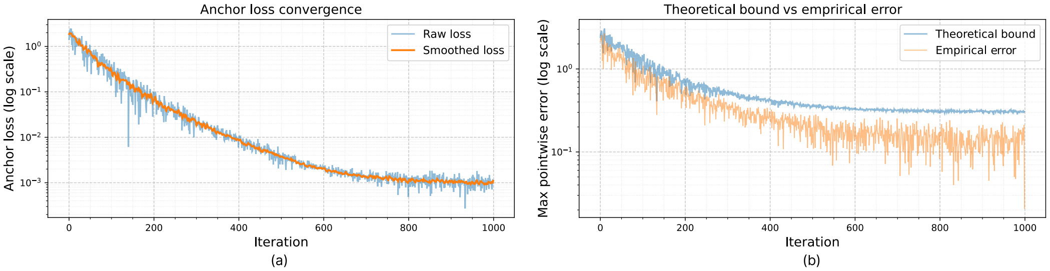

Figure 6: (a): Convergence of the Anchored Ranking Loss over 1,000 iterations. (b): Comparison of bounds, provided by Theorem 1 with empirical values of maximum pointwise errors.

C.3 Robustness Under Different Attack Strengths

Figure 7 shows how Spearman rank-order correlation varies with FGSM attack strength on the KADID-10k test set. We evaluate BiRQA, LPIPS and TOPIQ together with their anchored adversarial training versions under five ℓ_∞ budgets $\epsilon = \{0, 2, 4, 8, 10\}/255$, where $\epsilon = 0$ corresponds to clean images. All models are trained on the KADID-10k training set. BiRQA consistently maintains higher correlation than LPIPS and TOPIQ as perturbation strength increases. Anchored adversarial training markedly reduces the performance drop for each metric, with the anchored BiRQA variant retaining the highest SROCC across all budgets.

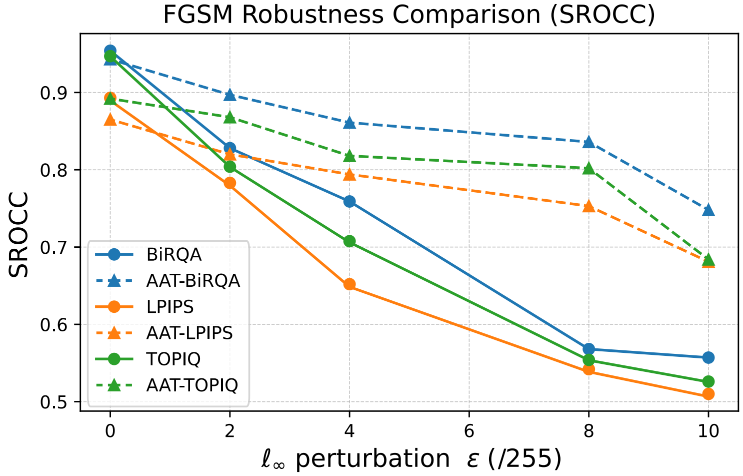

Figure 7: Robustness of FR IQA metrics to FGSM attack with different perturbation budgets evaluated on KADID-10k dataset. Solid lines show SROCC of vanilla models without adversarial training, dashed lines show SROCC of models trained with AAT.

C.4 2AFC Datasets Robustness Comparison

We evaluated four variations of three FR IQA models (LPIPS, TOPIQ and BiRQA) on Two Alternative Forced Choice (2AFC) dataset BAPPS. We compare **AAT** to (1) **base**: no adversarial training; (2) **R-**: vanilla adversarial training; (3) **AT-**: adversarial training with label smoothing. All AT and AAT models were optimized with a PGD-10 attack of budget $\epsilon = 8/255$. These models were attacked by seven unseen methods: FGSM, C&W, AutoAttack (AA), FACPA, the perceptual attack of Zhang et al., SquareAttack and Parsimonious, including 5 White Box and 2 Black Box methods. All attacks during testing were evaluated with a budget of $\epsilon = 8/255$. We report accuracy on clean and attacked versions of the BAPPS test set.

Table 6 presents the results, which shows that AAT achieves state-of-the-art robustness and outperforms other approaches by 0.02-0.1 accuracy points. AAT also provides the best accuracy on unperturbed test set compared to other defense methods. Even without defense, BiRQA possesses more robustness compared to LPIPS and TOPIQ, surpassing them by $0.03 - 0.06$ in terms of accuracy.

Table 6: Accuracy on 2AFC BAPPS test set of different adversarial training techniques, which were applied to LPIPS, TOPIQ and BiRQA models. The best results for each model are bolded. PGD-10 with $\epsilon = 8/255$ was used during training.

Model			White-Box Attacks				Black-Box Attacks	
	Clean	FGSM	C&W	AutoAttack	FACPA	Zhang et al.	SquareAttack	Parsimonious
base LPIPS	**0.742**	0.260	0.102	0.135	0.415	0.301	0.511	0.513
R-LPIPS	0.728	0.487	0.306	0.298	0.471	0.375	0.567	0.533
AT-LPIPS	0.729	0.495	0.440	0.412	0.589	0.436	0.615	0.586
AAT-LPIPS (ours)	0.736	**0.520**	**0.486**	**0.459**	**0.632**	**0.494**	**0.643**	**0.618**
base TOPIQ	**0.784**	0.358	0.320	0.341	0.496	0.416	0.542	0.552
R-TOPIQ	0.762	0.420	0.391	0.350	0.523	0.459	0.589	0.581
AT-TOPIQ	0.768	0.493	0.467	0.479	0.584	0.510	0.634	0.614
AAT-TOPIQ (ours)	0.775	**0.526**	**0.544**	**0.552**	**0.612**	**0.546**	**0.658**	**0.653**
base BiRQA	**0.794**	0.405	0.350	0.376	0.521	0.462	0.571	0.574
R-BiRQA	0.771	0.491	0.382	0.417	0.545	0.526	0.614	0.587
AT-BiRQA	0.771	0.573	0.473	0.485	0.570	0.581	0.662	0.672
AAT-BiRQA (ours)	0.782	**0.594**	**0.561**	**0.580**	**0.637**	**0.610**	**0.690**	**0.703**

C.5 Statistical Significance

We measure practical improvements by the difference in Spearman rank-order correlation (SROCC):

$$\Delta\rho_{i,j} = SROCC(y^{MOS}, y_i) - SROCC(y^{MOS}, y_j),$$

where y^{MOS} is the vector of mean-opinion scores, i and j index two IQA metrics, and y_i, y_j are their predictions. For every pair of IQA metrics, we estimate $\Delta\rho$ and its 95% confidence interval (CI) via a paired non-parametric bootstrap: We used the PIPAL dataset with a size of $N = 23,200$ distorted images for this experiment. From the test set of N image pairs we drew $R = 1{,}000$ bootstrap samples of size N with replacement, keeping the MOS vector and the two prediction vectors aligned. On each resample $b \in [1, ..., R]$ we computed $\Delta\rho^{(b)}$. The median of $\{\Delta\rho^{(b)}\}$ is reported as the effect size. The 2.5th and 97.5th percentiles form the two-sided 95% CI.

An improvement is considered significant when the lower CI bound is positive and $\Delta\rho \geq 0.01$, consistent with recent FR IQA benchmarks. Table 7 shows the results. The bootstrap makes no distributional assumptions, accounts for dependence between predictions, and remains valid even when error variance varies across the MOS range.

Table 7 shows that BiRQA outperforms every prior FR IQA metric on PIPAL. The gain is dramatic against classical metrics (e.g., +0.567 SROCC over PSNR) and even if small remains statistically significant against the strongest modern baselines (AHIQ/TOPIQ). The robustness-enhanced variant (AAT-BiRQA) sacrifices no more than 0.007 SROCC, confirming that anchored adversarial fine-tuning adds security almost for free.

Table 7: Pairwise ΔSROCC (±95% CI) on PIPAL (N = 23,000 distorted images) for 1,000 paired nonparametric bootstrap resamples (image pairs drawn with replacement; MOS and predictions kept aligned). Positive values favor the row metric; negative values favor the column metric. Only the lower triangle is shown; "—" indicates the symmetric counterpart.

	PSNR	SSIM	MAD	WaDIQaM	LPIPS	DISTS	AHIQ	TOPIQ	BiRQA (ours)
PSNR	—	—	—	—	—	—	—	—	—
SSIM	0.103 ±0.003	—	—	—	—	—	—	—	—
MAD	0.281 ±0.005	0.178 ±0.001	—	—	—	—	—	—	—
WaDIQaM	0.292 ±0.007	0.189 ±0.001	0.011 ±0.001	—	—	—	—	—	—
LPIPS	0.336 ±0.007	0.232 ±0.001	0.054 ±0.003	0.044 ±0.001	—	—	—	—	—
DISTS	0.399 ±0.006	0.295 ±0.001	0.117 ±0.006	0.107 ±0.004	0.063 ±0.002	—	—	—	—
AHIQ	0.564 ±0.017	0.460 ±0.014	0.282 ±0.014	0.272 ±0.012	0.228 ±0.009	0.165 ±0.004	—	—	—
TOPIQ	0.565 ±0.016	0.461 ±0.014	0.283 ±0.015	0.273 ±0.013	0.229 ±0.010	0.166 ±0.004	0.001 ±0.000	—	—
BiRQA (ours)	0.567 ±0.014	0.463 ±0.014	0.285 ±0.013	0.275 ±0.013	0.231 ±0.009	0.168 ±0.004	0.003 ±0.000	0.002 ±0.000	—
AAT-BiRQA (ours)	0.560 ±0.016	0.457 ±0.013	0.278 ±0.013	0.268 ±0.011	0.224 ±0.008	0.161 ±0.003	−0.004 ±0.001	−0.005 ±0.001	−0.007 ±0.001

D Comparison with Purification Defenses

Table 8 compares the proposed AAT technique with adversarial purification methods. These methods do not modify the IQA model itself. Instead they are preprocess input images. We compare AAT-BiRQA against basic preprocessing techniques such as Random Flip, Random Rotate and the specialized DiffPure defense method (Nie et al. (2022)). Results show that AAT-BiRQA outperforms all other purification methods, except DiffPure on SquareAttack. This likely reflects that SquareAttack has the least similar perturbations to PGD-10 used during adversarial training.

Table 8: Accuracy on 2AFC BAPPS test set for AAT-BiRQA compared to some adveersarial purification methods. The best results are bolded. PGD-10 with $\epsilon = 8/255$ was used during training of AAT-BiRQA.

Model	White-Box Attacks						Black-Box Attacks	
	Clean	FGSM	C&W	AutoAttack	FACPA	Zhang et al.	SquareAttack	Parsimonious
base BiRQA	**0.794**	0.405	0.350	0.376	0.521	0.462	0.571	0.574
base BIRQA + Random Flip	0.751	0.471	0.467	0.422	0.569	0.570	0.606	0.608
base BIRQA + Random Rotate	0.720	0.446	0.431	0.390	0.534	0.512	0.541	0.553
base BIRQA + DiffPure	0.741	0.522	0.498	0.452	0.619	0.603	**0.712**	0.698
AAT-BiRQA (ours)	0.782	**0.594**	**0.561**	**0.580**	**0.637**	**0.610**	0.690	**0.703**

E List of Parameters

The complete set of hyperparameters for both clean and adversarial training is provided in Table 9. For standard training, we use a regression-based objective that balances MSE and PLCC. In the adversarial setting, our anchored fine-tuning strategy integrates PGD-based attacks into the training loop and jointly optimizes clean and ranking losses. For the adversarial attacks and defenses, we have always used the default parameters except for ϵ, whose value is explicitly stated.

Table 9: List of hyper-parameters and description of experimental set-up for BiRQA.

Parameter	**Value**	**Parameter**	**Value**
Vanilla (clean) training		*Anchored Adversarial Training (AAT)*	
Optimizer	Adam	Inner attacker	PGD-10
Adam (β_1, β_2)	0.9/0.999	PGD step size	$2/255$
Batch size	32	PGD norm type	ℓ_∞
Epochs	2500	Perturbation budget ϵ	train: $8/255$; test: $\{2, 4, 8, 10\}/255$
Learning rate	10^{-4}	Anchor spacing λ	0.5
Loss function	$\mathcal{L}_{clean} = \alpha_1\text{MSE} - (1-\alpha_1)\text{PLCC}$ $\alpha_1 = 0.7$	AAT loss	$\alpha_2\mathcal{L}_{anchor} + (1-\alpha_2)\mathcal{L}_{clean}$ $\alpha_2 = 0.5$