Electronic Letters on Computer Vision and Image Analysis 25(2):106-115, 2026

ARQUE: A Hybrid Multi-Expert Framework for No-Reference Image Quality Assessment Using Curvature Analysis

Diego Frias*, Leandro Coelho* and Vagner Fonseca*

* *Department of Earth and Exact Sciences, PIMAT Research Group, Bahia State University, Salvador, Bahia, Brazil*

Received 7 October 2025; accepted 6 March 2026

Abstract

This work introduces ARQUE (Anisotropic Richness Quality Estimation), a No-Reference Image Quality Assessment (NR-IQA) framework based on a "Mixture of Experts" (MoE) architecture. The method relies on the Anisotropic Texture Richness (ATR) metric, which derives structural integrity measures from bi-directional curvature maps. Unlike conventional approaches that frequently benefit from global statistics and content overlap, the proposed system implements a physically grounded two-stage hybrid model. A probabilistic classifier first identifies the distortion type, followed by a weighted fusion of specialist Support Vector Regressors (SVRs) operating on a 21-dimensional feature space. Experimental validation employing a rigorous 10-fold cross-validation protocol on the LIVE, CSIQ, and TID2013 datasets indicates the stability of the proposed method. In the content-independent "Standard" scenario, ARQUE yields a Pearson Linear Correlation Coefficient (PLCC) of 0.901 on LIVE and 0.868 on TID2013, surpassing the optimized BRISQUE baseline (0.708 and 0.821, respectively). Moreover, while the baseline performance declines by approximately 25% when content overlap is removed, ARQUE exhibits stability with a variation of less than 3%, suggesting a reduced dependency on data leakage compared to statistical approaches.

Key Words: Image Artifact Diagnosis, NR-IQA, Directional Curvature Analysis, Robustness Analysis.

1 Introduction

Automatic Image Quality Assessment (IQA) is a pivotal technology in modern multimedia systems. While Full-Reference (FR) metrics are mature, No-Reference (NR) assessment remains a formidable challenge. The dominant paradigm is established by Natural Scene Statistics (NSS) methods, with **BRISQUE** [1] serving as the standard baseline for deviations in statistical regularity.

While effective, these generalist models often suffer from a critical limitation: they tend to overfit to image content rather than modeling the physical distortion. Recent studies suggest that the high performance reported in legacy literature often stems from "Data Leakage"—where overlapping content between training and testing sets allows the model to memorize scenes [2]—and a heavy reliance on non-linear logistic mapping to mask prediction errors [3].

Correspondence to: leandrocoelho@uneb.br

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.2422
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

In this work, we propose ARQUE (Anisotropic Richness Quality Estimation), a framework that shifts the paradigm from Generalist Statistics to Specialist Physics. By adopting a Mixture of Experts (MoE) architecture [4], ARQUE employs specific filters calibrated to the physical curvature of artifacts.

Our contributions are threefold:

1. **Physical Robustness:** We introduce the ATR metric, a curvature-based feature that maintains performance stability even when strict content separation (Standard No-Leak) is enforced.
2. **Ablation Study:** We provide a rigorous analysis comparing Easy (Leaked) vs. Standard (No-Leak) scenarios. We demonstrate that while the **BRISQUE baseline** suffers a significant performance drop in rigorous settings, ARQUE maintains its accuracy.
3. **Efficiency:** The proposed specialist architecture operates on a compact 21-dimensional feature space, representing a **42% reduction** in dimensionality compared to the **36-dimensional BRISQUE baseline**, facilitating lighter deployment without sacrificing accuracy.
4. **Metric Linearity:** Unlike traditional approaches that rely on non-linear logistic mapping to boost correlation, ARQUE provides direct quality estimation, offering a more transparent evaluation of physical artifacts.

2 Methodology

The system processes an input image I through a pipeline of feature extraction, probabilistic classification, and weighted regression. The implementation leverages the standard 'scikit-learn' library [5] for machine learning tasks.

2.1 The ATR Feature Extraction

The primary physical feature is the ATR, calculated from the second-order derivatives of the image intensity, consistent with the theory of edge detection [6]. Let C_h and C_v be the horizontal and vertical curvature maps obtained via Laplacian convolution. To mimic human perceptual non-linearity (Weber-Fechner Law [7]), we apply a logarithmic normalization:

$$L_d(x, y) = \ln(1 + |C_d(x, y)|), \quad d \in \{h, v\} \tag{1}$$

Crucially for a blind method, we compute the standard deviations σ_h and σ_v from the distorted image itself. These intrinsic statistics define the adaptive thresholds for the filters. A pixel (x, y) contributes to the ATR score if it satisfies specific directional constraints defined by parameters α (tolerance) and β (activation):

$$\text{Mask}_h = (L_v \geq \beta \cdot \sigma_v) \wedge (L_h < \alpha \cdot \sigma_h) \tag{2}$$

$$\text{Mask}_v = (L_h \geq \beta \cdot \sigma_h) \wedge (L_v < \alpha \cdot \sigma_v) \tag{3}$$

The ATR score is the density of pixels satisfying these conditions. Through Grid Search optimization, we identified unique (α, β) pairs that maximize correlation for specific distortions (e.g., $\alpha = 0.1, \beta = 1.8$ for JPEG in LIVE).

2.2 NSS-Augmented Classification Module

While ATR measures the intensity of a distortion, identifying the type of distortion requires structural context. We augment the ATR vector with Natural Scene Statistics (NSS) features [8]:

- Distribution Shape: Kurtosis (κ) and Skewness (s) of the curvature histograms.

- Pairwise Products: Mean and variance of products between neighboring pixels ($L(x, y) \cdot L(x, y + 1)$), capturing local correlation structures typical of blur or compression artifacts.

This combined feature vector feeds a Random Forest classifier [9], which outputs a probability distribution $P = \{p_1, p_2, ..., p_N\}$ over the N distortion classes.

2.3 Probabilistic Fusion (Soft Voting)

Instead of a "Winner-Takes-All" decision, which is prone to boundary errors, we employ a probabilistic fusion strategy. The system maintains a bank of N Specialist Support Vector Regressors (SVR) [10], denoted as $M_1, ..., M_N$. Each model M_i predicts a quality score q_i assuming the image suffers from distortion type i.

The final Differential Mean Opinion Score (DMOS) is computed as the expectation:

$$\text{DMOS}_{final} = \sum_{i=1}^{N} p_i \cdot M_i(\text{Features}_i) \tag{4}$$

where p_i is the confidence of the classifier for class i. This effectively creates a "Council of Experts", ensuring robustness even when the distortion type is ambiguous.

2.4 Feature Vector Composition and Computational Efficiency

To ensure reproducibility and clarify the computational footprint, we explicitly define the feature extraction pipeline. A key advantage of ARQUE is that the feature set required for the regression stage is a subset of the features extracted for classification, implying no additional computational cost after the diagnosis phase.

The system extracts a total of 21 unique features per image, organized as follows:

- **Common Feature Pool (21-D):**
  - **ATR Set (5 features):** The raw Anisotropic Texture Richness scores calculated for each of the five specialist filters (α_i, β_i).
  - **NSS Global (8 features):** Kurtosis, Skewness, Mean, and Standard Deviation computed for both horizontal (L_h) and vertical (L_v) curvature maps.
  - **NSS Local (8 features):** Mean and Standard Deviation of the pairwise products computed in both horizontal and vertical directions for both maps.

From this unified pool, the functional modules consume specific subsets:

1. **Classification Vector ($\mathbf{v}_{clf}$):** The Gate classifier utilizes the full **21-dimensional** vector to capture the complete structural context required for artifact diagnosis.
2. **Regression Vector ($\mathbf{v}_{reg}$):** Once the distortion type t is identified, the specific Specialist SVR receives a pruned **13-dimensional** vector, consisting of:
   - The single specific ATR score corresponding to type t.
   - The 12 most relevant NSS statistics (subset of the global and local pool).

Consequently, the total dimensionality of the ARQUE descriptor is **21**. This represents a significant reduction in complexity compared to the standard BRISQUE baseline, which requires a **36-dimensional** feature vector (18 features calculated at two scales). By operating on a lower-dimensional manifold without multi-scale downsampling, ARQUE achieves superior efficiency for real-time deployment.

3 Experimental Setup

3.1 Datasets and Protocol

We evaluated the framework on three benchmarks: LIVE [3], CSIQ [11], and TID2013 [12]. To ensure rigor, we adopted a 10-Fold Cross-Validation protocol with an 80%/20% training/testing split. Crucially, we defined three evaluation scenarios to test robustness with increasing levels of difficulty:

- **S1 (Easy):** Standard random image split with logistic mapping (contains data leakage).
- **S2 (Standard):** Strict random scene separation with logistic mapping (scientific standard).
- **S3 (Hardcore):** Strict random scene separation without logistic mapping (raw physical linearity).

Table 1 details the structure of the datasets used. To ensure a consistent evaluation across benchmarks, we focused on the five distortion types common to most multimedia applications: JPEG, JPEG2000 (JP2K), White Noise (WN), Gaussian Blur (GBLUR), and Fast Fading (FF).

Table 1: Structure of the Image Quality Datasets used in the evaluation.

Dataset	Ref. Images	Distortions	Levels	Total Images
LIVE	29	5	5–6	779
CSIQ	30	6 (4 used)	4–5	866
TID2013	25	24 (4 used)	5	3000 (Subset)

Unless otherwise stated, reported results refer to the rigorous **Scenario S2 (Standard)**.

3.2 Baselines and Metrics

We compare ARQUE against the optimized BRISQUE [1], whose implementation details and training protocol are described in Appendix A. Performance is measured by Pearson Linear Correlation Coefficient (PLCC) for linearity, Spearman Rank Order Correlation Coefficient (SROCC) for monotonicity, and Root Mean Squared Error (RMSE) for accuracy.

4 Results and Discussion

4.1 Gate Performance: Classification Accuracy

The effectiveness of the Mixture of Experts architecture relies on the accurate identification of distortion types. Table 2 presents the classification accuracy of the Random Forest module in the Standard (No-Leak) scenario.

Table 2: Mean Classification Accuracy (F1-Score) of the distortion identification module across 10 folds.

Distortion	LIVE	CSIQ	TID2013
FASTFADING	52.3%	-	-
GBLUR	67.1%	76.2%	81.1%
JP2K	68.1%	77.6%	78.2%
JPEG	70.1%	76.7%	79.0%
WN	82.5%	94.6%	96.9%
Overall	**68.0%**	**81.3%**	**83.8%**

The results indicate a progressive improvement in discrimination capability on larger and more diverse datasets, peaking at 83.8% on TID2013. The classification module proves exceptionally effective at identifying unstructured degradations, with White Noise (WN) detection rates exceeding 94% on CSIQ and TID2013. This confirms the sensitivity of the ATR feature to high-frequency stochastic disruptions.

The lower aggregate performance observed on the LIVE dataset (68.0%) is primarily driven by the "Fast Fading" artifact (52.3%), a complex distortion combining packet loss and error concealment that exhibits spectral signatures similar to compression artifacts (JPEG/JP2K) in the curvature domain. However, within the ARQUE framework, perfect classification is not a strict prerequisite. The probabilistic soft-voting mechanism mitigates these ambiguities by allowing a weighted contribution from multiple experts, ensuring that the final quality score remains accurate even when the distortion type is not uniquely identified.

4.2 Stratified Performance Analysis

To investigate the specialist capabilities, Table 3 compares the PLCC of ARQUE and BRISQUE stratified by distortion type.

Table 3: Stratified PLCC Comparison (Scenario S2: Standard). Bold indicates the best method.

Type	**LIVE**		**CSIQ**		**TID2013**	
	BRISQUE	ARQUE	BRISQUE	ARQUE	BRISQUE	ARQUE
FF	0.631	**0.839**	-	-	-	-
GBLUR	0.723	**0.903**	0.692	**0.780**	0.798	**0.835**
JP2K	0.544	**0.904**	0.638	**0.750**	0.837	**0.848**
JPEG	0.528	**0.880**	0.697	**0.779**	0.855	**0.905**
WN	0.633	**0.971**	0.385	**0.776**	0.709	**0.823**

The data indicates that ARQUE yields consistently higher correlation coefficients than the baseline across all evaluated distortions and datasets. The most substantial performance differential is observed in high-frequency artifacts, specifically White Noise (WN). On the CSIQ dataset, the proposed method achieves a correlation of 0.776, nearly doubling the baseline performance of 0.385, while reaching a near-optimal PLCC of 0.971 on LIVE.

Significant improvements are also evident in compression-based distortions within the LIVE dataset. While the generalist baseline exhibits a marked performance drop for JP2K (0.544) and JPEG (0.528) under the rigorous No-Leak scenario, ARQUE maintains high linearity with scores of 0.904 and 0.880, respectively. These margins ($\Delta > 0.35$) suggest that the baseline relies heavily on content overlap to model compression artifacts, whereas the proposed specialist architecture effectively captures the distortion manifold independent of scene content.

4.3 Quantitative Analysis and Robustness

Table 4 presents the comprehensive results of the 10-fold cross-validation. The inclusion of the improvement metric (Δ) and the significance test (p-value) allows for a nuanced analysis of the performance across the three defined levels of difficulty.

4.3.1 The Gap between Easy and Standard Scenarios

The comparison between S1 (Easy) and S2 (Standard) reveals the extent of overfitting in generalist models. In the LIVE dataset, when shifting from the Easy setup (with leakage) to the Standard setup (no leakage), BRISQUE's performance drops significantly (PLCC $0.85 \rightarrow 0.70$). ARQUE, however, maintains robustness (PLCC $0.92 \rightarrow 0.90$), achieving a net improvement of +27.11% over the baseline. This suggests that a portion

Table 4: Statistical & Stability Analysis. Comparison of Mean PLCC ± Standard Deviation across 10 folds. The Δ column shows the relative improvement of ARQUE over BRISQUE. (*) denotes statistical significance ($p < 0.05$).

Dataset	Scenario	BRISQUE	ARQUE	Δ Improv. (%)	p-value
LIVE	S1 (Easy)	0.8554 ± 0.0155	**0.9256 ± 0.0148**	**+8.20%**	**<0.001***
	S2 (Standard)	0.7084 ± 0.0949	**0.9005 ± 0.0450**	**+27.11%**	**<0.001***
	S3 (Hardcore)	0.6039 ± 0.0943	**0.8939 ± 0.0469**	**+48.01%**	**<0.001***
TID2013	S1 (Easy)	0.8777 ± 0.0427	**0.8916 ± 0.0275**	**+1.58%**	0.3541
	S2 (Standard)	0.8206 ± 0.0561	**0.8666 ± 0.0347**	**+5.61%**	0.0807
	S3 (Hardcore)	0.8007 ± 0.0634	**0.8496 ± 0.0492**	**+6.11%**	**0.0451***
CSIQ	S1 (Easy)	0.7499 ± 0.0470	**0.7929 ± 0.0483**	**+5.73%**	0.1102
	S2 (Standard)	0.6219 ± 0.1076	**0.7690 ± 0.0484**	**+23.66%**	**0.0091***
	S3 (Hardcore)	0.5637 ± 0.1401	**0.7514 ± 0.0518**	**+33.31%**	**0.0038***

of the baseline's performance in legacy benchmarks may be attributed to data leakage, whereas ARQUE relies on physical feature extraction.

4.3.2 Linearity in the Hardcore Scenario

The S3 (Hardcore) scenario, which removes the aid of logistic mapping, serves as a test for linearity. On the challenging TID2013 dataset, while the improvement in the Standard scenario (S2) was marginal ($p = 0.08$), the gap widens in the Hardcore scenario, where ARQUE becomes statistically superior ($p = 0.045$). This result indicates that the raw output of ARQUE is more linear with human perception than the raw output of NSS-based methods.

4.4 Stability Analysis (Slope Chart)

To visualize the degradation of performance across the three levels of difficulty, we present the Stability Slope analysis in Fig. 1 and Table 5.

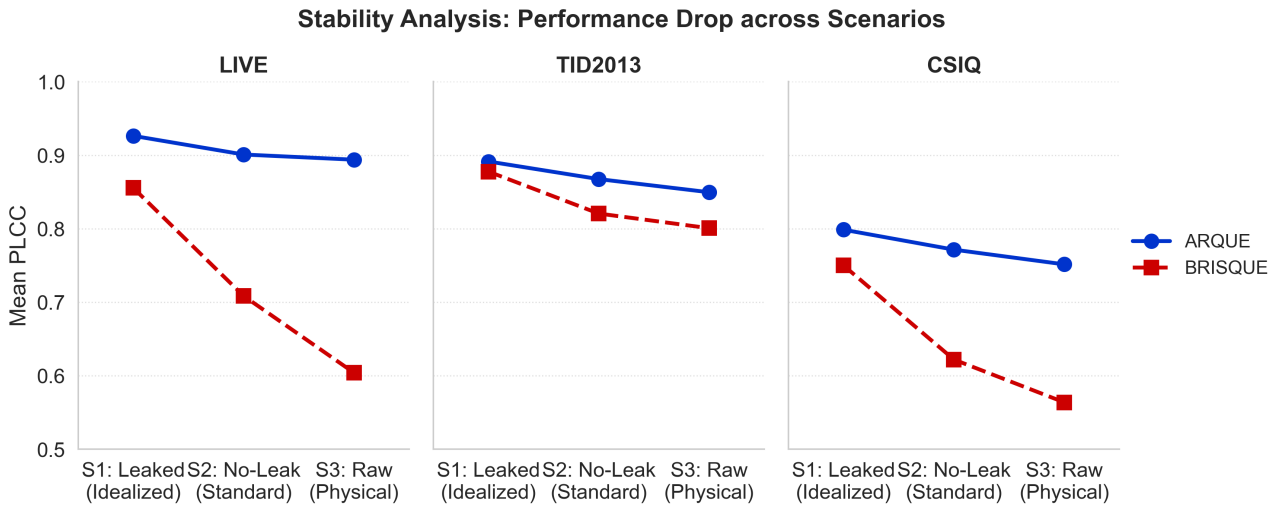


Figure 1: Stability Slope Analysis. ARQUE (Blue) maintains a flat trajectory across scenarios. BRISQUE (Red) shows a decline when moving from Easy (S1) to Hardcore (S3) protocols.

Table 5: Performance Evolution. Mean PLCC scores across scenarios (Corresponding to Fig. 1).

Dataset	Method	S1 (Easy)	S2 (Standard)	S3 (Hardcore)
LIVE	BRISQUE	0.8554	0.7084	0.6039
	ARQUE	**0.9256**	**0.9005**	**0.8939**
TID2013	BRISQUE	0.8777	0.8206	0.8007
	ARQUE	**0.8916**	**0.8666**	**0.8496**
CSIQ	BRISQUE	0.7499	0.6219	0.5637
	ARQUE	**0.7929**	**0.7690**	**0.7514**

The data shows that BRISQUE suffers a performance drop of up to 29% (LIVE Easy $\rightarrow$ Hardcore) as experimental rigor increases. In contrast, ARQUE demonstrates stability, with a maximum drop of less than 3%.

4.5 Qualitative Analysis: Global Scatter Plots

To provide a comprehensive view of the model's behavior, Fig. 2 visualizes the aggregated predictions across all 10 test folds in the S2 (Standard) scenario. This visualization exposes the variance and prediction density of the models over the entire datasets.

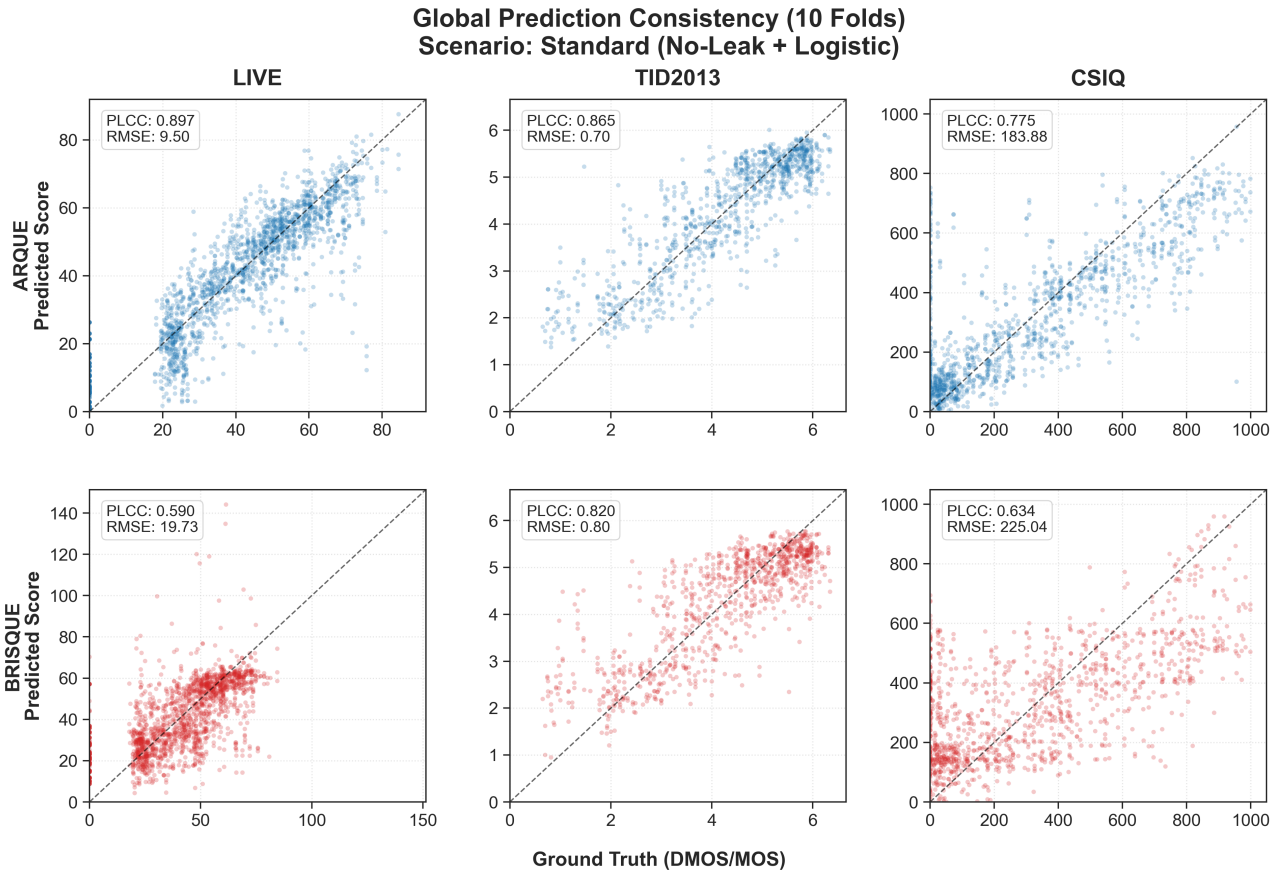


Figure 2: Global Prediction Consistency (S2: Standard). Comparison of ARQUE (Top Row) versus BRISQUE (Bottom Row) across LIVE, TID2013, and CSIQ datasets. The aggregated data from 10 folds reveals that ARQUE maintains a tight clustering along the ideal diagonal (dashed line). In contrast, BRISQUE exhibits dispersion, particularly visible in the CSIQ dataset, illustrating the limitations of relying on generalist statistics without scene context.

Visual inspection of the global data reinforces the quantitative findings. ARQUE (Top Row) maintains a cohesive structural relationship with human perception, regardless of the training fold, producing a highly linear distribution. Conversely, the dispersion observed in BRISQUE (Bottom Row) visualizes the uncertainty cost when the model is forced to generalize without data leakage.

5 Discussion and Conclusion

The experimental results support the hypothesis that distortion-specific decomposition improves estimation accuracy. While generalist models such as BRISQUE are required to approximate a complex, non-linear mapping across heterogeneous physical phenomena, the proposed architecture enables specialist SVRs to model the manifold of individual distortion types. This decomposition renders the regression task more tractable, resulting in improved precision while utilizing a reduced feature space.

Methodologically, the ATR metric serves as a physically interpretable measure of degradation intensity, complementing the NSS features employed for artifact classification. The subsequent probabilistic fusion stage effectively smooths predictions, mitigating potential discontinuities at the decision boundaries between distortion categories.

In conclusion, the ARQUE framework offers a physically grounded approach to No-Reference IQA. The analysis indicates that minimizing content dependency through physical modeling enhances robustness compared to purely statistical baselines. Future work will investigate cross-dataset generalization and the optimization of the proposed architecture for real-time video quality assessment.

Code Availability

The complete source code for the ARQUE framework, including the feature extraction modules, trained models, and reproduction scripts for the benchmarks presented in this paper, is available at:
`https://github.com/diego-frias/ARQUE`

References

[1] A. Mittal, A. K. Moorthy, A. C. Bovik, "No-reference image quality assessment in the spatial domain", *IEEE Transactions on Image Processing* 21(12):4695–4708, 2012. doi:10.1109/TIP.2012.2214050

[2] F. Sesguen, C. B. Akgul, T. Y. Ozyer, "Evaluation of No-Reference Image Quality Assessment Algorithms on Different Databases using Various Validation Methods", *IEEE Access* 7:124699–124713, 2019. doi:10.1109/ACCESS.2019.2938459

[3] H. R. Sheikh, M. F. Sabir, A. C. Bovik, "A Statistical Evaluation of Recent Full Reference Quality Assessment Algorithms", *IEEE Transactions on Image Processing* 15(11):3440–3451, 2006. doi:10.1109/TIP.2006.881959

[4] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, "Adaptive mixtures of local experts", *Neural Computation* 3(1):79–87, 1991. doi:10.1162/neco.1991.3.1.79

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research* 12:2825–2830, 2011.

[6] D. Marr, E. Hildreth, "Theory of edge detection", *Proceedings of the Royal Society of London B* 207(1167):187–217, 1980. doi:10.1098/rspb.1980.0020

[7] G. T. Fechner, *Elemente der Psychophysik*, Breitkopf u. Härtel, 1860.

[8] D. L. Ruderman, "Statistics of natural images", *Network: Computation in Neural Systems* 5(4):517–548, 1994. doi:10.1088/0954-898X_5_4_006

[9] L. Breiman, "Random forests", *Machine Learning* 45:5–32, 2001. doi:10.1023/A:1010933404324

[10] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, V. Vapnik, "Support vector regression machines", *Advances in Neural Information Processing Systems* 9, 1997.

[11] E. C. Larson, D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy", *Journal of Electronic Imaging* 19(1):011006, 2010. doi:10.1117/1.3267105

[12] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C. C. J. Kuo, "Image database TID2013: Peculiarities, results and perspectives", *Signal Processing: Image Communication* 30:57–77, 2015. doi:10.1016/j.image.2014.10.009

A Baseline Implementation Details

To ensure a fair comparison and reproducibility, we re-implemented the BRISQUE [1] baseline in Python using the `OpenCV` and `scikit-learn` libraries. The implementation adheres strictly to the algorithm described in the original literature, adapted for the same experimental protocol used for ARQUE.

A.1 Feature Extraction Pipeline

The extraction process follows three steps:

1. **Color Conversion:** Images are converted to grayscale using the ITU-R 601-2 luma transform ($Y = 0.299R + 0.587G + 0.114B$).
2. **MSCN Normalization:** Mean Subtracted Contrast Normalized (MSCN) coefficients are computed to capture deviations from natural statistics:

$$\hat{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C} \tag{5}$$

   where $\mu(i,j)$ and $\sigma(i,j)$ are local mean and variance computed with a 7×7 Gaussian kernel.
3. **Statistical Fitting:** An Generalized Gaussian Distribution (GGD) is fitted to the MSCN coefficients, and an Asymmetric GGD (AGGD) is fitted to the pairwise products of neighboring coefficients (horizontal, vertical, main-diagonal, secondary-diagonal).

This process is repeated at two scales (original and downsampled by a factor of 2), resulting in a final feature vector of dimension $18 \times 2 = 36$.

A.2 Optimization and Fairness Protocol

Unlike legacy benchmarks that often report results based on the parameters provided by the original authors (optimized for LIVE), we performed a specific optimization for each dataset and each training fold.

We employed Support Vector Regression (SVR) with an RBF kernel. To guarantee that the baseline achieved its theoretical upper bound, we conducted an exhaustive Grid Search within the training set of each fold (Nested Cross-Validation). The search space was defined as:

- **Penalty (C):** $[10, 100, 1000, 10000]$
- **Kernel Coefficient (γ):** $[0.001, 0.01, 0.1, 1.0]$
- **Epsilon (ϵ):** $[0.01, 0.1, 0.5]$