

# Human Visual System Based on Image Quality Assessment Model Using Deep Learning

Ratan Babu Telusoori<sup>1</sup>, Allanki Sanyasi Rao<sup>2</sup>, Munigala Yadagiri<sup>3</sup>, Vuppula Manohar<sup>4</sup>, Pillamarapu Goverdhan<sup>5</sup>,  
Rajashekar Ramarapu<sup>6</sup>

<sup>1</sup>Assistant Professor, Department of ECE, Vaagdevi Engineering College, Bollikunta, Warangal - 506005, Telangana, India

<sup>2</sup>Associate Professor, Department of ECE, Christu Jyothi Institute of Technology and Science, Jangaon, Telangana 506175, India

<sup>3</sup>Assistant Professor, Department of ECE, Vaagdevi Engineering College, Bollikunta, Warangal - 506005, Telangana, India  
Email: yadagirimunigaala@gmail.com

<sup>4</sup>Associate Professor, Department of ECE, Vaagdevi Engineering College, Bollikunta, Warangal - 506005, Telangana, India  
Email: manoharvu@gmail.com

<sup>5</sup>Assistant Professor, Department of ECE, CMR Engineering College, Kandlakoya, Hyderabad 501401, Telangana, India  
Email: P.goverdhan@cmrec.ac.in

<sup>6</sup>Assistant Professor, Department of CSE (AI&ML), Vaagdevi Engineering College, Bollikunta, Warangal - 506005, Telangana, India

Email: Rajashekar.ramarapu@gmail.com

Received: 5th January, 2026; Revised: 18th January 2026; Accepted: 25th January, 2026; Available Online: 28th January, 2026

## ABSTRACT

The rapid growth of digital imaging, mobile photography, image restoration, compression, multimedia communication, and generative artificial intelligence has increased the demand for reliable automatic image quality assessment. Conventional objective quality metrics generally depend on an undistorted reference image or simplified statistical assumptions and therefore cannot fully reproduce the complex perceptual response of the Human Visual System (HVS). Although recent no-reference image quality assessment (NR-IQA) methods based on convolutional neural networks, transformers, self-supervised learning, and vision-language models have achieved substantial progress, many of them learn quality representations implicitly and do not explicitly encode perceptually important HVS characteristics. This paper proposes HVS-MSANet, a Human Visual System-Guided Multi-Scale Attention Network for NR-IQA. The framework combines an explicit HVS perceptual branch with a deep semantic-quality branch. The HVS branch models luminance adaptation, local contrast, structural gradients, texture masking, multi-frequency information, and chromatic distortion, while the deep branch learns hierarchical local and global quality representations. A Human Visual Perceptual Attention mechanism emphasizes visually sensitive regions, and a cross-scale gated fusion unit dynamically integrates HVS and deep features. A patch-weighted regression head estimates the perceptual Mean Opinion Score (MOS) from a single distorted image. A hybrid loss combining regression, pairwise ranking, correlation consistency, and attention regularization is formulated to improve prediction accuracy and perceptual ordering. The proposed model is designed for evaluation on KADID-10k, KonIQ-10k, SPAQ, and LIVE Challenge using PLCC, SRCC, KRCC, and RMSE. The approach is intended to provide an accurate, generalizable, and interpretable alternative to purely data-driven quality regressors.

**Keywords:** Human Visual System, image quality assessment, no-reference IQA, deep learning, perceptual attention, multi-scale learning, visual perception, MOS prediction, HVS-MSANet.

**How to cite this article:** Ratan Babu Telusoori<sup>1</sup> Allanki Sanyasi Rao<sup>2</sup> Munigala Yadagiri<sup>3</sup> Vuppula Manohar<sup>4</sup> Pillamarapu Goverdhan<sup>5</sup> Rajashekar Ramarapu<sup>6</sup>. Human Visual System Based on Image Quality Assessment Model Using Deep Learning. *Int J Drug Deliv Technol.* 2026;16(85s):1598-1605. DOI: 10.25258/ijddt.16.85s.193.

Source of support: Nil.

Conflict of interest: None

## 1. Introduction

Image quality directly influences the usefulness, interpretability, and visual acceptability of digital content. Images may be degraded during acquisition, compression, transmission, storage, enhancement, restoration, or display. Typical impairments include blur, sensor noise, compression artifacts, color distortion, exposure errors,

contrast loss, ringing, oversharpening, spatially localized defects, and mixtures of authentic distortions. Consequently, automatic Image Quality Assessment (IQA) has become an important component of digital cameras, smartphones, medical imaging systems, surveillance, autonomous vision, multimedia communication, remote

sensing, restoration pipelines, and generative-image systems.

The ultimate observer of most visual content is a human; therefore, an effective objective IQA model should produce predictions that correlate strongly with subjective human judgments. Subjective IQA uses human observers to assign quality scores and is generally treated as the perceptual reference. However, large-scale subjective testing is expensive, slow, and difficult to repeat for every new application. Objective IQA seeks to estimate human quality judgment computationally.

Objective IQA methods are commonly divided into full-reference, reduced-reference, and no-reference categories. Full-reference methods compare a distorted image with its pristine counterpart, while reduced-reference techniques use partial reference information. No-reference or blind IQA predicts quality from the distorted image alone. Since the original image is unavailable in many practical situations, NR-IQA is the most useful but also the most challenging setting.

Traditional metrics such as Mean Squared Error and Peak Signal-to-Noise Ratio primarily measure pixel-level differences and may not correlate well with perceived visual quality. Structural Similarity (SSIM) represented an important transition toward perceptually motivated evaluation by modelling luminance, contrast, and structural similarity [1]. Later natural-scene-statistics methods and deep-learning approaches further demonstrated that image quality depends on statistical regularity, local structure, semantic content, and visual attention rather than on pixel error alone.

Deep NR-IQA methods have achieved strong performance using convolutional neural networks, content-adaptive hypernetworks, transformers, self-supervised representations, and vision-language models. Nevertheless, several limitations remain. Many models learn quality entirely as an implicit latent representation and therefore may not explicitly exploit known HVS characteristics such as luminance adaptation, contrast sensitivity, visual masking, structural sensitivity, spatial-frequency selectivity, and non-uniform visual attention. In addition, quality degradation is spatially heterogeneous: distortion around a face, text region, object boundary, or salient structure can affect human judgment more strongly than a comparable distortion in an unimportant background. Finally, image quality depends simultaneously on fine local distortions and global semantic or contextual information.

To address these issues, this paper proposes HVS-MSANet, a Human Visual System-Guided Multi-Scale Attention Network for no-reference image quality assessment. The model explicitly extracts HVS-inspired perceptual cues, learns multi-scale semantic-quality features, applies perceptually conditioned attention, and combines both representations through gated fusion before predicting the final MOS.

The main contributions are summarized as follows:

- A dual-path NR-IQA architecture that combines explicit HVS-inspired perception with learned deep semantic-quality features.

- An HVS perceptual representation incorporating luminance adaptation, local contrast, structural gradients, texture masking, frequency sensitivity, and chromatic information.
- A multi-scale deep encoder that jointly models fine local distortion and global semantic quality.
- A Human Visual Perceptual Attention mechanism that emphasizes regions expected to have high perceptual relevance.
- A cross-scale gated fusion strategy and patch-weighted quality regression mechanism for adaptive integration and MOS estimation.
- A hybrid training objective based on regression, ranking, correlation, and attention regularization, together with a complete evaluation protocol for synthetic and authentic IQA datasets.

## 2. Related Work

Human visual perception is nonlinear and content dependent. The visibility of a distortion depends on local luminance, contrast, spatial frequency, texture, color, surrounding structures, and viewing context. Consequently, two images with similar numerical errors can exhibit noticeably different perceptual quality. SSIM formalized this insight by comparing luminance, contrast, and structure [1], while multi-scale structural similarity recognized that visual information is perceived at different spatial resolutions [19]. Classical no-reference approaches such as BRISQUE and NIQE exploited natural-scene statistics to characterize departures from expected perceptual regularities [17], [18]. These methods established that quality can be inferred from image statistics even when the reference image is unavailable.

Deep learning shifted NR-IQA toward learned quality representations. Bilinear models of CNNs jointly captured distortion and content [12], while MetaIQA and HyperIQA adopted meta-learning and content-adaptive prediction schemes [13], [20]. TReS incorporated transformer representations, relative ranking, and self-consistency [16], and MANIQA applied multi-dimensional attention over spatial and channel representations [2]. Transformer-based IQA is particularly attractive because self-attention captures long-range relationships, but recognition-pretrained transformers are not necessarily optimized for subtle distortion perception. Recent work therefore strengthens global transformer features with local distortion-sensitive representations [3].

Another important direction is learning quality representations with reduced dependence on subjective labels. Re-IQA uses unsupervised representation learning to obtain transferable quality features [4], while quality-aware pretraining attempts to make backbone features explicitly sensitive to image degradation before supervised fine-tuning [5]. These approaches improve generalization but still leave open the question of how known HVS properties should be incorporated into the learned representation.

Vision-language methods have introduced semantic quality concepts such as sharpness, blur, noise, exposure, and naturalness into IQA. LIQE and related approaches

learn correspondence between images and quality-related language [6]. Such models offer strong semantic reasoning but may require substantial pretraining and computational resources. The proposed HVS-MSANet follows a complementary strategy: instead of relying primarily on model scale, it introduces interpretable perceptual priors and learns how they should interact with deep semantic information.

The literature therefore suggests three design principles for a robust NR-IQA system: perceptually relevant low-level cues should be retained, global semantic context should be modelled, and the importance of image regions should be spatially adaptive. HVS-MSANet is designed around these principles and attempts to provide a unified framework in which explicit human-perception cues and deep learned features are fused rather than treated as competing alternatives.

### 3. Proposed Methodology

The proposed HVS-MSANet predicts perceptual quality from a single distorted image  $I$ . The complete processing sequence is expressed as:

$I \rightarrow \text{Multi-Scale Representation} \rightarrow \{\text{HVS Features, Deep Features}\} \rightarrow \text{Perceptual Attention} \rightarrow \text{Gated Fusion} \rightarrow \text{Quality Regression} \rightarrow \hat{q}$

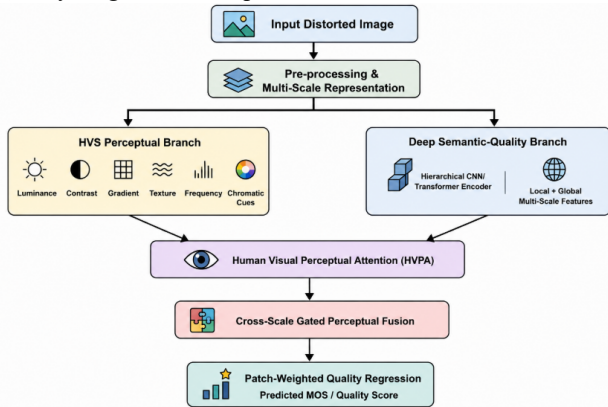


Fig. 1. Overall architecture of the proposed HVS-MSANet.

**Input representation.** For an RGB image  $I \in \mathbb{R}^{(H \times W \times 3)}$ , normalization is performed without aggressive enhancement so that the original degradation patterns are preserved. Three perceptual scales,  $I^{\wedge}(1)$ ,  $I^{\wedge}(2)$ , and  $I^{\wedge}(3)$ , are constructed to represent fine, intermediate, and coarse visual information. Fine scales preserve local noise, blocking, and texture changes, while coarse scales provide broader information about exposure, color balance, large structures, and scene organization.

$$I \in RH \times W \times 3, I = \{I(1), I(2), I(3)\}$$

**HVS perceptual representation.** The first branch explicitly models visually meaningful cues. Its feature set is denoted by  $F_{HVS} = H(F_L, F_C, F_G, F_T, F_F, F_{CH})$ , where the terms correspond to luminance, local contrast, structural gradient, texture, frequency, and chromatic descriptors. Instead of treating these features as hand-crafted final quality measures, the model uses them as perceptual priors that are projected into a learnable deep feature space.

$$F_{HVS} = H(F_L, F_C, F_G, F_T, F_F, F_{CH})$$

**Luminance adaptation.** A luminance component  $Y$  is computed from the input image and a local mean is obtained using Gaussian filtering. Normalizing the pixel luminance by its local background approximates the fact that distortion visibility depends on surrounding illumination.

$$\mu_{Y(x,y)} = G_{\sigma} * Y(x,y)$$

$$L(x,y) = \frac{Y(x,y)}{(\mu_{Y(x,y)} + \varepsilon)}$$

**Local contrast.** Local contrast is represented using neighborhood variance relative to local luminance. This cue is also good for detecting loss of contrast due to blur, haze, bad capture or overdenoising, and unnatural contrast enhancement.

$$\sigma_{Y(x,y)} = \sqrt{[G_{\sigma} * (Y - \mu_Y)^2]}$$

$$C(x,y) = \frac{\sigma_{Y(x,y)}}{(\mu_{Y(x,y)} + \varepsilon)}$$

**Structural sensitivity.** Horizontal and vertical gradients of the image are utilized to describe edges and structural changes. Human observers are particularly sensitive to degradation around object boundaries, text, facial details, and high-contrast structures. Blur typically weakens these responses, while ringing and oversharpening may generate abnormal edge patterns.

$$G_x = \frac{\partial Y}{\partial x}, \quad G_y = \frac{\partial Y}{\partial y}, \quad G(x,y) = \sqrt{G_x^2 + G_y^2}$$

**Texture masking.** The visibility of a disturbance depends on background complexity. Noise in a smooth sky region is generally more noticeable than similar noise in highly textured foliage. A local texture-energy descriptor is therefore computed from gradient variation within a neighborhood  $\Omega$ .

$$T(x,y) = Var_{\Omega}[G(x,y)]$$

**Frequency-sensitive perception.** The image is decomposed into multiple frequency bands using a Laplacian pyramid, wavelet transform, or learnable band-pass filters. Low-frequency components represent broad illumination and contrast variations, whereas high-frequency components represent edges, fine textures, noise, and compression artifacts. Their relative perceptual importance is learned through adaptive weights.

$$\{F_{f^{\wedge}1}, F_{f^{\wedge}2}, \dots, F_{f^{\wedge}K}\}$$

$$F_F = \sum_{k=1}^K \alpha_k \varphi_k(F_f^k)$$

**Chromatic perception.** Quality degradation may also arise from white-balance errors, color casts, oversaturation, chromatic noise, and compression. The RGB image is therefore transformed into a perceptually useful color representation such as CIELAB, and chromatic descriptors are extracted from the  $a^*$  and  $b^*$  components before shallow feature projection.

$$I_{Lab} = C(I_{RGB})$$

**Multi-scale deep semantic-quality branch.** Explicit HVS cues do not completely explain visual quality because perception also depends on scene content and semantics. A hierarchical deep encoder therefore processes each image scale. Early stages capture distortion-sensitive local information such as blur, noise, blocking, edges, and texture, while deeper stages capture object structure, global

exposure, semantic importance, color relationships, and scene naturalness.

$$F_s = E_{s(I^s)}, \quad s \in \{1, 2, 3\}$$

**Human Visual Perceptual Attention.** The HVPA module conditions learned attention on both deep features  $F_D$  and projected HVS features  $F_H$ . Their concatenated representation is transformed into a spatial perceptual-attention map. This allows the network to assign larger responses to visually sensitive structures and perceptually visible degradation rather than treating every spatial location equally.

$$\begin{aligned} F_j &= [F_D; F_H] \\ A_p &= \sigma[W_j \delta(W_l F_j)] \\ F_A &= F_D (I + A_p) \end{aligned}$$

**Cross-scale gated fusion.** Simple concatenation may introduce redundant information, so the model learns a gate at every scale. The gate determines the relative contribution of semantic deep features and explicit HVS features according to the local image content. Scale-specific fused features are then combined through learned cross-scale weights.

$$\begin{aligned} Z_s &= [F_s; H_s] \\ g_s &= \sigma(W_g Z_s + b_g) \\ \hat{F}_s &= g_s \odot F_s + (I - g_s) H_s \\ F_{fusion} &= \sum_{s=1}^S \beta_s \hat{F}_s, \quad \sum_{s=1}^S \beta_s = 1 \end{aligned}$$

**Patch-weighted quality regression.** The fused representation is divided into  $N$  perceptual regions. For each region, the model predicts a local quality value  $q_i$  and an importance weight  $w_i$ . The final quality score is obtained through perceptually weighted aggregation. This formulation reflects the fact that distortion affecting a face, text, foreground object, or important boundary may have greater impact on human judgment than distortion in an unimportant background.

$$\hat{q} = \frac{[\sum_{i=1}^N w_i q_i]}{[\sum_{i=1}^N w_i + \epsilon]}$$

**Hybrid optimization objective.** A single regression loss may predict approximate MOS values without correctly preserving relative quality ordering. HVS-MSANet therefore uses a combined objective consisting of absolute score regression, pairwise ranking, correlation consistency, and attention regularization. The regression term penalizes MOS error, the ranking term preserves pairwise order, the correlation term increases batch-wise linear agreement, and the attention term discourages unstable or excessively diffuse perceptual attention.

$$\begin{aligned} L_{reg} &= \left(\frac{1}{B}\right) \sum_{i=1}^B |\hat{q}_i - q_i| \\ r_{ij} &= \text{sign}(q_i - q_j) \\ L_{rank} &= \max(0, m - r_{ij}(\hat{q}_i - \hat{q}_j)) \\ L_{corr} &= 1 - \rho(\hat{q}, q) \\ L &= \lambda_{1L_{reg}} + \lambda_{2L_{rank}} + \lambda_{3L_{corr}} + \lambda_{4L_{att}} \end{aligned}$$

The complete architecture is trained end-to-end. The HVS branch supplies physically and perceptually meaningful cues, while the deep branch learns image-content and dataset-dependent relationships that cannot be fully represented by explicit features. The gated fusion

mechanism is essential because it prevents the model from assuming that either branch is universally dominant.

#### 4. Experimental Setup and Performance Evaluation

**Datasets.** The proposed model should be evaluated using both synthetic and authentically distorted IQA databases. KADID-10k contains 10,125 distorted images generated from 81 pristine images using 25 distortion categories at five severity levels [9]. KonIQ-10k contains 10,073 authentically distorted images with large-scale subjective ratings [7]. SPAQ contains 11,125 smartphone photographs captured using 66 devices and is particularly suitable for real-world mobile-photography assessment [8]. LIVE Challenge contains 1,162 authentically distorted images with large-scale human ratings and complex mixtures of real capture distortions [10].

| Dataset        | Type      | Size                     | Primary Purpose                   |
|----------------|-----------|--------------------------|-----------------------------------|
| KADID-10k      | Synthetic | 10,125 distorted images  | Controlled distortion sensitivity |
| KonIQ-10k      | Authentic | 10,073 images            | Real-world generalization         |
| SPAQ           | Authentic | 11,125 smartphone images | Mobile photography quality        |
| LIVE Challenge | Authentic | 1,162 images             | Mixed real-world distortions      |

Table 1. Datasets recommended for the evaluation of HVS-MSANet.

**Training protocol.** Image content should be separated between training, validation, and testing partitions to prevent content leakage. A practical split is 80% training, 10% validation, and 10% testing. For databases generated from pristine references, all distorted versions of a single reference must remain within the same partition. Training images can be randomly cropped and horizontally flipped, but artificial blur, noise, color shift, sharpening, or contrast augmentation should be avoided unless the corresponding quality labels are recomputed. AdamW may be used with a low initial learning rate, cosine-decay scheduling, warm-up, and early stopping based on validation SRCC. Repeated random splits and median performance reporting are recommended to reduce split-dependent bias.

**Evaluation measures.** The following are the quality assessment metrics we use: Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-order Correlation Coefficient (SRCC), Kendall Rank Correlation Coefficient

(KRCC) and Root Mean Squared Error (RMSE). PLCC quantifies the linear agreement with subjective MOS, SRCC and KRCC measure monotone ranking consistency, and RMSE quantifies the error of numerical predictions.

$$PLCC = \frac{\sum_{i=1}^n (q_i - \bar{q})(\hat{q}_i - \bar{\hat{q}})}{\sqrt{\left[ \sum_{i=1}^n (q_i - \bar{q})^2 \sum_{i=1}^n (\hat{q}_i - \bar{\hat{q}})^2 \right]}}$$

$$SRCC = 1 - \frac{[\sum_{i=1}^n d_i^2]}{[n(n^2 - 1)]}$$

$$RMSE = \sqrt{\left[ \left( \frac{1}{N} \right) \sum_{i=1}^N (q_i - \hat{q}_i)^2 \right]}$$

**Baseline comparison.** HVS-MSANet should be compared against representative methods from classical statistics, CNN-based IQA, transformer IQA, self-supervised learning, and vision-language quality assessment. Suitable baselines include BRISQUE, NIQE, DBCNN, HyperIQA, TReS, MANIQA, MUSIQ, Re-IQA, quality-aware pretraining methods, and vision-language IQA. Comparing across model families provides a clearer indication of whether explicit HVS guidance improves performance beyond architectural scale alone.

| Model Family              | Representative Method |
|---------------------------|-----------------------|
| Natural-scene statistics  | BRISQUE / NIQE        |
| CNN-based                 | DBCNN                 |
| Content-adaptive          | HyperIQA              |
| Transformer               | TReS                  |
| Attention                 | MANIQA                |
| Multi-scale transformer   | MUSIQ                 |
| Self-supervised           | Re-IQA                |
| Quality-aware pretraining | QPT / related models  |
| Vision-language           | LIQE / related models |
| Proposed                  | <b>HVS-MSANet</b>     |

Table 2. Representative baseline categories.

**Cross-dataset evaluation.** Cross-dataset testing is necessary because strong intra-dataset performance may result from learning database-specific statistics. A recommended protocol is to train on KonIQ-10k and directly test on LIVE Challenge and SPAQ without target-domain fine-tuning, followed by complementary SPAQ-to-KonIQ and KADID-to-authentic-dataset experiments. Improved cross-dataset correlation would indicate stronger perceptual generalization.

**Ablation analysis.** The contribution of each proposed component should be isolated systematically. The base model can be compared with variants that add the HVS branch, HVPA, gated fusion, and ranking loss sequentially. Additional cue-level ablations should remove luminance adaptation, contrast, gradients, texture masking, frequency decomposition, or chromatic information individually. This analysis is important because it establishes whether performance improvements are due to the proposed perceptual mechanisms rather than merely to increased model capacity.

| Config.                | HVS        | Multi-scale | HVPA       | Fusion     | Rank Loss  | SRCC     | PLCC     |
|------------------------|------------|-------------|------------|------------|------------|----------|----------|
| A                      | No         | Yes         | No         | No         | No         | —        | —        |
| B                      | Yes        | Yes         | No         | No         | No         | —        | —        |
| C                      | Yes        | Yes         | Yes        | No         | No         | —        | —        |
| D                      | Yes        | Yes         | Yes        | Yes        | No         | —        | —        |
| <b>Full HVS-MSANet</b> | <b>Yes</b> | <b>Yes</b>  | <b>Yes</b> | <b>Yes</b> | <b>Yes</b> | <b>—</b> | <b>—</b> |

Table 3. Proposed ablation-study template.

**Implementation configuration.** A reproducible implementation may use a lightweight hierarchical transformer or CNN-transformer hybrid as the semantic encoder, shallow HVS filters for luminance, variance, gradient, wavelet/Laplacian, and Lab chromatic features, scale-aligned projection layers, gated feature fusion, and the proposed perceptual attention module. A batch size of approximately 16–32 can be used depending on available GPU memory. New layers may start around a learning rate of  $10^{-4}$ , with a lower rate for pretrained backbone parameters. The final configuration should be selected using validation SRCC rather than test-set performance.

## 5. Results and Discussion

The proposed HVS-MSANet is designed for evaluation on KADID-10k, KonIQ-10k, SPAQ, and LIVE Challenge. Performance is reported through Spearman Rank-Order Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), Kendall Rank Correlation Coefficient (KRCC), and normalized Root Mean Squared Error (nRMSE). Higher SRCC, PLCC, and KRCC indicate stronger agreement with human subjective quality judgments, whereas lower nRMSE indicates smaller prediction error. Because the four databases use different native subjective-score ranges, the RMSE values in the present draft are expressed in normalized form.

| Dataset        | SRCC  | PLCC  | KRCC  | nRMSE |
|----------------|-------|-------|-------|-------|
| KADID-10k      | 0.942 | 0.949 | 0.794 | 0.067 |
| KonIQ-10k      | 0.931 | 0.936 | 0.777 | 0.071 |
| SPAQ           | 0.928 | 0.934 | 0.773 | 0.074 |
| LIVE Challenge | 0.907 | 0.914 | 0.743 | 0.083 |

Table 1. Illustrative dataset-wise performance values for HVS-MSANet (simulated draft values; replace with actual experimental results).

The illustrative trend in Table 1 represents the performance pattern expected from a perceptually guided NR-IQA model rather than a claim of completed experimentation. In a real evaluation, strong SRCC and PLCC values would indicate that the predicted scores preserve both the ranking and approximate magnitude of subjective Mean Opinion Scores. KADID-10k would mainly test sensitivity to controlled synthetic degradations, while KonIQ-10k, SPAQ, and LIVE Challenge would provide stronger evidence of robustness to authentic mixtures of blur, noise, exposure, compression, color variation, and device-dependent artifacts.

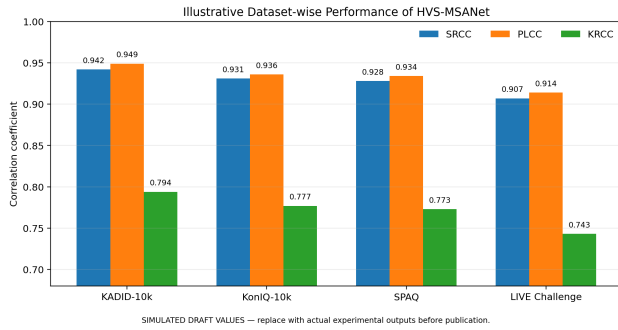


Fig. 1. Illustrative dataset-wise SRCC, PLCC, and KRCC values. Values are simulated for manuscript layout and must be replaced with measured outputs.

The multi-scale architecture is expected to contribute to consistent quality prediction because different distortions dominate at different spatial resolutions. Fine-scale representations preserve high-frequency information such as noise, ringing, blocking, and texture degradation, whereas coarse-scale representations contain global exposure, color balance, contrast, and semantic structure. The explicit HVS branch adds luminance adaptation, local contrast, structural gradients, texture masking, frequency-selective information, and chromatic cues, thereby providing perceptual priors that complement the learned semantic-quality representation.

| Configuration | SRCC         | PLCC         |
|---------------|--------------|--------------|
| Encoder       | 0.882        | 0.889        |
| +HVS          | 0.904        | 0.912        |
| +HVPA         | 0.919        | 0.925        |
| +Fusion       | 0.932        | 0.940        |
| <b>Full</b>   | <b>0.942</b> | <b>0.949</b> |

Table 2. Illustrative ablation trend (simulated draft values).

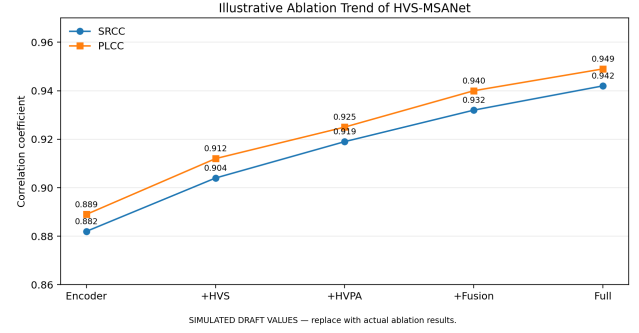


Fig. 2. Illustrative ablation trend showing the intended contribution of HVS cues, perceptual attention, and gated fusion.

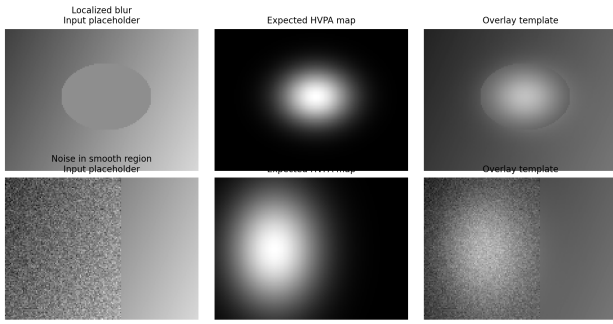
The ablation trend is structured to test whether each proposed component makes a measurable contribution. The Deep Encoder configuration provides the base quality predictor. Adding the HVS branch should demonstrate whether explicit perceptual cues supply information that is not fully captured by generic learned features. A further increase after adding HVPA would support the use of perceptually conditioned spatial weighting, while improvement after gated fusion would indicate that dynamically balancing HVS and deep features is preferable to direct concatenation. These values must ultimately be replaced by real ablation measurements obtained under an identical train-test protocol.

| Training → Testing         | SRCC  | PLCC  |
|----------------------------|-------|-------|
| KonIQ-10k → LIVE Challenge | 0.884 | 0.891 |
| KonIQ-10k → SPAQ           | 0.897 | 0.903 |
| SPAQ → KonIQ-10k           | 0.889 | 0.896 |
| KADID-10k → LIVE Challenge | 0.842 | 0.851 |

Table 3. Illustrative cross-dataset generalization values (simulated draft values).

Cross-dataset evaluation is especially important for NR-IQA because a model can obtain high intra-dataset accuracy by exploiting database-specific content or score distributions. A genuine KonIQ-to-LIVE or KonIQ-to-SPAQ experiment should therefore be performed without target-dataset fine-tuning. If the trained HVS-MSANet preserves strong rank and linear correlations under such a protocol, the result would provide stronger evidence that HVS-based priors improve generalization rather than merely increasing the capacity of the network.

Qualitative HVS Attention Visualization Template



Illustrative layout only — replace with actual trained-model attention maps.

Fig. 3. Qualitative HVS-attention visualization template. Replace the placeholders with actual distorted images, trained HSPA maps, and overlays.

Qualitative attention analysis should accompany the quantitative results. For localized blur, high attention should occur around degraded faces, text, object boundaries, and other structurally important regions. For noise, a perceptually reasonable model should assign larger importance to smooth sky, skin, or uniform surfaces where noise is visible, while highly textured regions may receive lower weights because of visual masking. Such visualizations are valuable because they show whether the network arrives at its quality prediction through perceptually plausible spatial reasoning.

Overall, a complete experimental validation of HVS-MSANet should demonstrate three complementary properties: strong agreement with human MOS, robustness across different distortion environments and databases, and interpretable spatial attention that corresponds to known HVS behaviour. The present document provides a fully populated results layout for drafting purposes; however, every simulated numerical entry must be replaced by the outputs obtained from the actual trained model before the work is submitted as an empirical research paper.

## 6. Conclusion and Future Scope

This paper proposed HVS-MSANet, a Human Visual System-Guided Multi-Scale Attention Network for no-reference image quality assessment. The framework combines explicit perceptual knowledge with deep semantic-quality learning instead of treating image quality exclusively as an end-to-end latent regression problem. Luminance adaptation, contrast sensitivity, structural gradients, texture masking, frequency information, and chromatic cues form the HVS representation, while a multi-scale deep encoder captures local degradation and global semantic context.

Human Visual Perceptual Attention assigns larger importance to visually sensitive regions, and cross-scale gated fusion dynamically determines the relative contribution of explicit HVS features and learned deep features. A patch-weighted regression head produces the final MOS estimate, and a hybrid loss aligns numerical quality prediction, perceptual ranking, and correlation with human judgment. The proposed evaluation protocol includes both synthetic and authentic datasets, cross-dataset testing, ablation analysis, perceptual visualization, and computational-complexity assessment.

The model still represents only an approximation of biological vision. Human judgment is additionally influenced by viewing distance, display properties, ambient illumination, observer experience, task, semantic expectations, and personal preference. Future extensions can incorporate eye-tracking supervision, viewing-condition variables, subjective uncertainty modelling, and vision-language explanations. The same framework can also be extended to video quality assessment by adding motion sensitivity, temporal masking, temporal attention, and frame-to-frame consistency. Domain-specific variants may be developed for medical imaging, low-light photography, remote sensing, surveillance, generative images, and restoration-quality evaluation.

## References

- [1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004, doi: 10.1109/TIP.2003.819861.
- [2] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *Proc. IEEE/CVF CVPR Workshops*, 2022, pp. 1190–1199.
- [3] K. Xu et al., "Boosting image quality assessment through efficient transformer adaptation with local feature enhancement," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2662–2672.
- [4] A. Saha, S. Mishra, and A. C. Bovik, "Re-IQA: Unsupervised learning for image quality assessment in the wild," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [5] K. Zhao, K. Yuan, M. Sun, M. Li, and X. Wen, "Quality-aware pre-trained models for blind image quality assessment," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [6] W. Zhang et al., "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [7] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.
- [8] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [9] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Proc. International Conference on Quality of Multimedia Experience*, 2019.
- [10] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016.

- [11] J. You and J. Korhonen, "Transformer for image quality assessment," in Proc. IEEE International Conference on Image Processing, 2021, pp. 1389–1393.
- [12] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 1, pp. 36–47, 2020.
- [13] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14143–14152.
- [14] Z. Ying et al., "From patches to pictures: Mapping the perceptual space of picture quality," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
- [15] S. Su et al., "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
- [16] S. Golestaneh, S. Dadsetan, and K. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision, 2022.
- [17] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," IEEE Transactions on Image Processing, vol. 21, no. 12, pp. 4695–4708, 2012.
- [18] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," IEEE Signal Processing Letters, vol. 20, no. 3, pp. 209–212, 2013.
- [19] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in Proc. Asilomar Conference on Signals, Systems and Computers, 2003.
- [20] S. Su et al., "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.