

# SMART-IQA: Swin Multi-scale Attention-guided Regression Transformer for Blind Image Quality Assessment

Nuoyan Chen

School of Computer Science

Shanghai Jiao Tong University

Shanghai, China

cny123222@sjtu.edu.cn

**Abstract**—Blind image quality assessment (BIQA) for authentically distorted images presents fundamental challenges due to distortion diversity, strong content dependency, and limited annotated data. Unlike synthetic distortions with controlled characteristics, real-world images exhibit complex, non-uniform degradation patterns whose perceptual impact varies dramatically with semantic content. While HyperIQA pioneered content-adaptive assessment through a self-adaptive hyper network that separates content understanding from quality prediction, its ResNet-50 backbone struggles to capture long-range dependencies and fine-grained hierarchical features crucial for assessing diverse authentic distortions. We propose SMART-IQA, a Swin Transformer-based framework that extends the content-adaptive paradigm by integrating an Adaptive Feature Aggregation (AFA) module with dynamic attention-guided fusion. By leveraging Swin Transformer’s hierarchical window-based self-attention and preserving spatial structure through adaptive pooling, our method captures richer multi-scale representations. A lightweight channel attention mechanism enables content-aware feature weighting, allowing the model to adaptively emphasize different feature hierarchies based on image content and distortion characteristics. Extensive experiments on KonIQ-10k demonstrate that SMART-IQA achieves state-of-the-art performance with 0.9378 SRCC, outperforming the original HyperIQA by 3.18% and other competing methods. Cross-dataset evaluations further validate strong generalization capability across diverse distortion types and image domains.

**Index Terms**—Image Quality Assessment, Swin Transformer, Adaptive Feature Aggregation, Attention Mechanism, Hyper Network, Deep Learning

## I. INTRODUCTION

Blind Image Quality Assessment (BIQA) aims to automatically predict the perceptual quality of images without access to pristine references, mimicking the quality judgment process of the human visual system. While significant progress has been made for laboratory-generated synthetic distortions, assessing authentically distorted “in-the-wild” images remains a fundamental challenge. The difficulty arises from three core factors: *distortion diversity*, where multiple unknown degradations occur simultaneously in complex patterns; *content dependency*, where human quality perception is intrinsically linked to image semantics; and *data scarcity*, with limited large-scale annotated datasets for diverse authentic distortions.

The evolution of BIQA methodologies reflects a broader shift from expert-driven feature engineering to data-driven learning. Early approaches relied on hand-crafted Natural Scene Statistics (NSS) or Human Visual System (HVS)-guided features, which exhibited significant performance degradation on authentic distortions due to their content-agnostic nature. The deep learning revolution introduced CNN-based methods that learn quality features end-to-end, yet these models still applied fixed parameters uniformly across all images, failing to adapt to content-dependent quality perception.

A pivotal paradigm shift was introduced by HyperIQA [1], which pioneered *content-adaptive* assessment through a self-adaptive hyper network architecture. HyperIQA explicitly separates the IQA procedure into three stages that mirror human top-down perception: *content understanding*, *perception rule learning*, and *quality prediction*. By dynamically generating quality prediction weights  $\theta_x$  based on image semantic features, the model adapts its assessment strategy to image content, addressing the fundamental limitation that “the same distortion affects different content types differently.” This content-adaptive paradigm enables more psychologically plausible and practically effective quality assessment for diverse real-world images.

However, HyperIQA’s ResNet-50 backbone, while providing strong semantic features, has inherent limitations in capturing long-range dependencies and fine-grained hierarchical features crucial for assessing complex authentic distortions. Vision Transformers have demonstrated remarkable capabilities in modeling global context through self-attention mechanisms [2]. Swin Transformer [3] further improves efficiency and scalability through hierarchical architecture with shifted window attention, naturally producing multi-scale representations suitable for dense prediction tasks like IQA.

In this work, we propose SMART-IQA (Swin Multi-scale Attention-guided Regression Transformer for Image Quality Assessment), which extends the content-adaptive paradigm by integrating Swin Transformer’s hierarchical vision architecture with our proposed Adaptive Feature Aggregation (AFA) module and dynamic attention-guided fusion. Our key contributions are: (1) We replace the CNN backbone with Swin

Transformer to capture richer semantic and spatial features through hierarchical window-based self-attention, enabling superior modeling of both local distortions and global context. (2) We design the AFA module that preserves spatial structure by adaptively pooling multi-stage features to unified resolution, avoiding aggressive compression that discards distortion-relevant information. (3) We introduce a lightweight channel attention mechanism that dynamically weights different feature scales based on image content and distortion characteristics, enabling content-aware multi-scale fusion. (4) We demonstrate through comprehensive experiments that SMART-IQA achieves state-of-the-art performance on KonIQ-10k with 0.9378 SRCC, outperforming the original HyperIQA by 3.18% and other competing methods, while maintaining strong cross-dataset generalization.

## II. RELATED WORK

### A. Evolution of BIQA: From Hand-Crafted to Deep Learning

Early BIQA methodologies relied on expert-driven feature engineering, broadly categorized into Natural Scene Statistics (NSS)-based and Human Visual System (HVS)-guided approaches. NSS-based methods like BRISQUE [4] and NIQE [5] quantify quality by measuring deviations from statistical regularities of natural images in various transform domains. While computationally efficient, these hand-crafted methods are largely *content-agnostic*, applying the same quality rules regardless of image semantics, and exhibit significant performance degradation on authentic distortions whose characteristics differ fundamentally from controlled synthetic distortions.

The advent of deep learning marked a paradigm shift, enabling end-to-end learning of quality-relevant features directly from data. Early CNN-based approaches utilized patch-based inputs and relatively shallow architectures for computational efficiency. More sophisticated methods emerged: NIMA [6] predicts aesthetic and technical quality distributions using CNNs trained on large-scale datasets; DBCNN [7] employs bilinear pooling to capture interactions between distortion patterns and image content; PaQ-2-PiQ [8] learns perceptual quality representations through contrastive learning. However, these methods still suffered from *content insensitivity*—they failed to adapt their assessment based on semantic content, which is crucial since “the perceptual impact of a distortion varies dramatically with content.”

### B. Content-Adaptive Paradigm: HyperNetwork Architectures

A watershed moment in BIQA evolution came with HyperIQA [1], which introduced *content-adaptive* assessment through a self-adaptive hyper network architecture. This approach fundamentally reconceptualizes quality assessment by transitioning from static models  $\phi(x, \theta) = q$  with fixed parameters  $\theta$  to dynamic models  $\phi(x, \theta_x) = q$  where parameters are image-dependent. HyperIQA explicitly separates the IQA procedure into three stages that mirror hypothesized human top-down perception: (1) *Content Understanding*: A semantic feature extraction network (ResNet-50 pretrained on ImageNet) extracts features  $S(x)$  representing what the

image depicts. (2) *Perception Rule Learning*: A hyper network  $H(S(x), \gamma)$  dynamically generates weight parameters  $\theta_x$  based on semantic features, learning the mapping from content to quality perception rules. (3) *Quality Prediction*: A target network with dynamically generated weights predicts the final quality score using multi-scale content features.

This content-adaptive mechanism enables the model to apply different “quality perceiving rules” tailored to specific image content. For example, when assessing a clear blue sky image, the model learns to discount texture-based indicators that would incorrectly penalize intentional flat regions as blur artifacts. By judging quality based upon content understanding, the network predictions become more consistent with human perception. HyperIQA demonstrated state-of-the-art results on challenging authentic databases, validating that content adaptivity is fundamental for assessing diverse real-world images.

### C. Transformer-Based Architectures for IQA

The subsequent emergence of transformer-based architectures represents another significant advancement, leveraging self-attention mechanisms to model complex long-range dependencies across entire feature maps. Unlike HyperNetworks that adapt parameters to content, transformers directly model the relational structure within content itself through attention operations  $\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k})V$ .

MANIQA [9] exemplifies this evolution with its Multi-dimension Attention Network, which integrates several innovations: Vision Transformer (ViT) for multi-scale feature extraction, Transposed Attention Blocks (TAB) applying self-attention across channel dimensions to encode global context, Scale Swin Transformer Blocks (SSTB) enhancing local patch interactions, and a dual-branch structure for patch-weighted quality prediction. This channel-wise attention complements traditional spatial attention, enabling the model to understand which feature channels are most relevant for quality assessment. MUSIQ [10] introduces multi-scale transformers processing images at multiple resolutions. TReS [11] combines transformers with relative ranking loss for improved generalization.

Empirical evaluations demonstrate that transformer-based models achieve superior performance compared to HyperNetwork approaches on both synthetic and authentic datasets, with performance advantages stemming from: (1) *Global Context Modeling*: Self-attention directly relates any two image regions, capturing long-range dependencies affecting holistic quality perception. (2) *Multi-Dimensional Feature Interaction*: Both spatial and channel-wise attention model the complex interplay between quality-degrading factors. (3) *Shared Attention Mechanisms*: Unlike HyperNetworks generating unique parameters per image, transformers apply shared attention that can improve generalization across diverse content.

### D. Vision-Language Models and Multimodal Integration

Recent work has begun exploring Vision-Language Models (VLMs) like CLIP for BIQA, leveraging rich semantic

understanding from large-scale pre-training. LIQE [12] formulates BIQA as multitask learning over quality, scene, and distortion through vision-language correspondence. CLIP-IQA [13] demonstrates zero-shot quality assessment through text-image similarity. While promising for zero-shot generalization, these methods face challenges in computational efficiency and reasoning reliability for deployment.

### E. Our Approach

Building upon the content-adaptive paradigm established by HyperIQA, we propose SMART-IQA which integrates the global modeling capabilities of transformer architectures with dynamic multi-scale feature fusion. By replacing the ResNet-50 backbone with Swin Transformer and introducing Adaptive Feature Aggregation with attention-guided fusion, our method captures both the content-aware adaptivity of HyperNetworks and the superior representational power of transformers for assessing complex authentic distortions.

## III. METHOD

### A. Overview

Following the content-adaptive paradigm of HyperIQA [1], we formulate BIQA as  $\phi(x, \theta_x) = q$  where  $\theta_x = H(S(x), \gamma)$  are image-dependent parameters generated by a HyperNetwork  $H$  based on semantic features  $S(x)$ . Our SMART-IQA extends this paradigm with three key innovations: (1) *Swin Transformer backbone* for hierarchical multi-scale feature extraction with global context modeling, (2) *Adaptive Feature Aggregation (AFA)* module that preserves spatial structure while unifying multi-scale features, and (3) *channel attention mechanism* for content-aware dynamic weighting of different feature scales. Figure 1 illustrates the complete architecture.

### B. Hierarchical Feature Extraction via Swin Transformer

Traditional CNN backbones in IQA models, such as ResNet-50 in HyperIQA, extract features through convolutions with limited receptive fields, struggling to capture long-range dependencies crucial for holistic quality perception. In contrast, Vision Transformers leverage self-attention mechanisms to model global relationships, but standard ViT architectures lack the hierarchical multi-scale representations essential for capturing both fine-grained distortion patterns and high-level semantic content.

We adopt Swin Transformer [3] as our backbone network, which combines the benefits of hierarchical feature extraction with efficient window-based self-attention. Given an input image  $x \in \mathbb{R}^{H \times W \times 3}$ , the Swin Transformer processes it through  $K$  hierarchical stages (in our implementation,  $K = 4$ ), producing multi-scale feature maps:

$$\{F^1, F^2, \dots, F^K\} = \text{SwinTransformer}(x) \quad (1)$$

where  $F^i \in \mathbb{R}^{H_i \times W_i \times C_i}$  denotes the feature map from stage  $i$ , with spatial dimensions  $(H_i, W_i)$  and channel dimension  $C_i$ .

The hierarchical structure follows a pyramid design with progressively decreasing spatial resolutions and increasing

channel dimensions. Specifically, for stage  $i > 1$ , the spatial-channel relationship is governed by:

$$H_i = \frac{H_{i-1}}{s_i}, \quad W_i = \frac{W_{i-1}}{s_i}, \quad C_i = r_i \cdot C_{i-1} \quad (2)$$

where  $s_i$  is the spatial downsampling factor (typically 2) and  $r_i$  is the channel expansion ratio (typically 2). This pyramid structure enables the model to capture information at multiple scales:  $(H_1, W_1, C_1) \rightarrow (H_2, W_2, C_2) \rightarrow \dots \rightarrow (H_K, W_K, C_K)$ .

Within each stage, the core computation is the shifted window-based multi-head self-attention (SW-MSA), which operates on non-overlapping windows to achieve linear complexity with respect to image size. For a feature map partitioned into  $M \times M$  windows, the attention operation within each window  $\mathcal{W}$  is:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} \quad (3)$$

where  $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{M^2 \times d}$  are the query, key, and value matrices projected from features within window  $\mathcal{W}$ , and  $d$  is the head dimension. By alternating regular window partitioning and shifted window partitioning across consecutive blocks, the model establishes connections between neighboring windows, enabling information flow across the entire feature map while maintaining computational efficiency.

This hierarchical design naturally captures both low-level texture patterns in early stages ( $F^1, F^2$ ) crucial for detecting distortion artifacts such as blur, noise, and compression artifacts, and high-level semantic features in later stages ( $F^{K-1}, F^K$ ) necessary for content understanding and context-aware quality assessment.

### C. Adaptive Feature Aggregation (AFA) Module

A fundamental challenge in multi-scale BIQA is how to effectively aggregate features from different hierarchical levels. Direct concatenation of multi-scale features is infeasible due to mismatched spatial dimensions, while naive global average pooling completely discards spatial structure that may be crucial for localizing non-uniform distortions. We address this challenge through our proposed Adaptive Feature Aggregation (AFA) module, which unifies features from different stages to a common spatial resolution while preserving spatial structure.

**Motivation and Design Rationale.** The key insight is that different stages capture complementary quality-relevant information: early stages ( $F^1, F^2$ ) with high spatial resolution are sensitive to local distortions and fine-grained texture degradation, while later stages ( $F^{K-1}, F^K$ ) with rich semantic information capture global structure and content-dependent quality attributes. However, these features reside in different spatial-channel spaces. Our AFA module bridges this gap through a two-step transformation: spatial alignment via adaptive pooling and channel alignment via learned projections.

**Spatial Alignment via Adaptive Pooling.** For each stage  $i \in \{1, 2, \dots, K-1\}$ , we first apply adaptive average pooling to unify the spatial resolution to a target size  $(H_{\text{target}}, W_{\text{target}})$ :

$$\tilde{F}^i = \text{AdaptiveAvgPool}(F^i, H_{\text{target}} \times W_{\text{target}}) \quad (4)$$

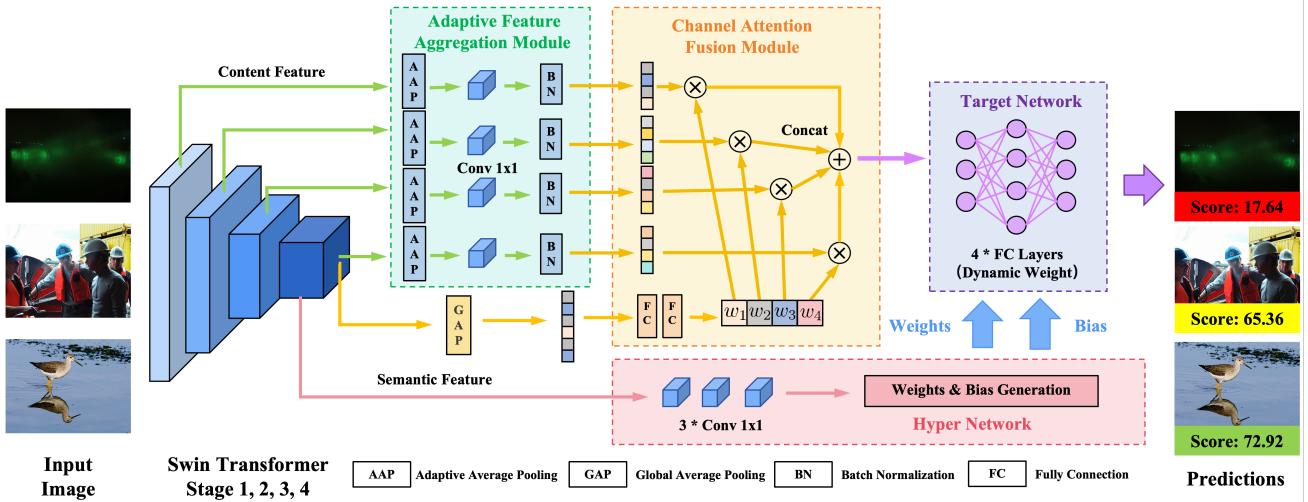


Fig. 1. Architecture of SMART-IQA. The pipeline consists of: (1) Swin Transformer backbone with  $K$  hierarchical stages extracting multi-scale features  $\{F^1, F^2, \dots, F^K\}$  with progressively decreasing spatial resolutions and increasing channel dimensions, (2) Adaptive Feature Aggregation (AFA) module that unifies spatial dimensions through adaptive pooling and  $1 \times 1$  convolution, producing aligned features  $\{F_{\text{pool}}^1, F_{\text{pool}}^2, \dots, F_{\text{pool}}^{K-1}\}$  at target resolution  $H_{\text{target}} \times W_{\text{target}}$ , (3) Channel Attention Fusion module that uses the deepest feature  $F^K$  to generate attention weights  $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_K\}$  via global pooling and FC layers, dynamically weighting multi-scale features based on content, (4) HyperNet that generates dynamic parameters  $\theta_x = \{W_1, b_1, W_2, b_2\}$  for the TargetNet based on  $F^K$ , and (5) TargetNet that predicts the final quality score  $q$  using the attention-weighted feature vector  $v_x$  and content-adaptive parameters  $\theta_x$ . The orange-highlighted attention module enables content-aware feature fusion, while the red dashed arrows indicate dynamic weight generation. In our implementation, we use  $K = 4$  stages (see Section 3.3 for details).

where  $\tilde{F}^i \in \mathbb{R}^{H_{\text{target}} \times W_{\text{target}} \times C_i}$ . The adaptive pooling operation partitions each spatial location  $(h, w)$  of the input feature map into a receptive field and computes the average:

$$\tilde{F}^i[h, w, :] = \frac{1}{|R_{h,w}|} \sum_{(p,q) \in R_{h,w}} F^i[p, q, :] \quad (5)$$

where  $R_{h,w}$  is the receptive field corresponding to output position  $(h, w)$ , with size determined by:

$$|R_{h,w}| = k_h \times k_w, \quad k_h = \left\lceil \frac{H_i}{H_{\text{target}}} \right\rceil, \quad k_w = \left\lceil \frac{W_i}{W_{\text{target}}} \right\rceil \quad (6)$$

This adaptive pooling strategy has a critical advantage over fixed-kernel pooling: it automatically adjusts the receptive field size based on the input-output resolution ratio, ensuring each output spatial location aggregates information from an appropriately sized region. For higher-resolution early-stage features, larger receptive fields aggregate more local information, while for lower-resolution later-stage features, smaller receptive fields preserve more spatial detail.

**Channel Alignment via Learned Projections.** After spatial unification, features from different stages still have heterogeneous channel dimensions  $C_i$ . We employ  $1 \times 1$  convolutions to project all features to a unified channel dimension  $C_{\text{unified}}$ :

$$F_{\text{pool}}^i = \text{ReLU}(\text{Conv}_{1 \times 1}(\tilde{F}^i; \mathbf{W}^i)) \in \mathbb{R}^{H_{\text{target}} \times W_{\text{target}} \times C_{\text{unified}}} \quad (7)$$

where  $\mathbf{W}^i \in \mathbb{R}^{C_{\text{unified}} \times C_i \times 1 \times 1}$  are learnable projection weights specific to stage  $i$ . The  $1 \times 1$  convolution serves two purposes: (1) channel dimension standardization for subsequent concatenation, and (2) learning stage-specific feature transformations that emphasize quality-relevant patterns at each scale. The ReLU activation introduces non-linearity, enabling

the projection to learn complex transformations beyond linear combinations.

**Multi-Scale Feature Unification.** After processing all  $K-1$  stages through spatial and channel alignment, we obtain a set of unified feature maps:

$$\mathcal{F}_{\text{AFA}} = \{F_{\text{pool}}^1, F_{\text{pool}}^2, \dots, F_{\text{pool}}^{K-1}\} \quad (8)$$

where each  $F_{\text{pool}}^i \in \mathbb{R}^{H_{\text{target}} \times W_{\text{target}} \times C_{\text{unified}}}$  shares the same spatial and channel dimensions. For the deepest stage  $F^K$ , which typically already has spatial resolution  $H_K = H_{\text{target}}$ , we optionally apply a similar  $1 \times 1$  convolution for channel alignment if  $C_K \neq C_{\text{unified}}$ . These unified features form the foundation for subsequent content-aware weighting via the channel attention mechanism.

#### D. Channel Attention for Content-Aware Feature Weighting

While the AFA module unifies multi-scale features into a common representation space, a critical question remains: *how should features from different scales be weighted for quality prediction?* Fixed equal weighting ignores the fundamental insight that different quality levels and distortion types may require emphasizing different feature hierarchies. High-quality images can often be assessed primarily through high-level semantic features, while low-quality images with visible distortions necessitate careful examination of low-level texture patterns. To address this, we introduce a channel attention mechanism that dynamically determines the importance of each scale based on image content.

**Global Semantic Descriptor Extraction.** We leverage the deepest stage feature  $F^K \in \mathbb{R}^{H_K \times W_K \times C_K}$ , which encodes the most abstract semantic representation of image content, to

guide the attention weight generation. A global descriptor is extracted via global average pooling:

$$\mathbf{g} = \text{GAP}(F^K) = \frac{1}{H_K \cdot W_K} \sum_{h=1}^{H_K} \sum_{w=1}^{W_K} F^K[h, w, :] \in \mathbb{R}^{C_K} \quad (9)$$

This operation compresses the spatial dimensions while preserving the channel-wise statistics, yielding a compact representation of the global semantic content. The choice of  $F^K$  for attention generation is motivated by the hypothesis that high-level semantic understanding (e.g., recognizing whether an image depicts a natural scene, portrait, or architectural structure) should inform which feature scales are most relevant for quality assessment.

**Scale Importance Prediction via Gating Network.** The global descriptor  $\mathbf{g}$  is fed through a lightweight two-layer gating network to predict the importance of each hierarchical stage:

$$\mathbf{z} = \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{g} + \mathbf{b}_1) \quad (10)$$

$$\alpha = \sigma(\mathbf{W}_2 \cdot \mathbf{z} + \mathbf{b}_2) \quad (11)$$

where  $\mathbf{W}_1 \in \mathbb{R}^{d_{\text{hidden}} \times C_K}$  and  $\mathbf{b}_1 \in \mathbb{R}^{d_{\text{hidden}}}$  define the first fully connected layer with hidden dimension  $d_{\text{hidden}}$ ,  $\mathbf{z} \in \mathbb{R}^{d_{\text{hidden}}}$  is the intermediate representation,  $\mathbf{W}_2 \in \mathbb{R}^{K \times d_{\text{hidden}}}$  and  $\mathbf{b}_2 \in \mathbb{R}^K$  define the second layer that outputs  $K$  attention logits, and  $\sigma(\cdot)$  is the element-wise sigmoid function. The resulting attention vector  $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_K]^T \in (0, 1)^K$  represents the learned importance weights for all  $K$  stages.

The two-layer design with bottleneck dimension  $d_{\text{hidden}} < C_K$  serves two purposes: (1) it reduces the number of parameters for computational efficiency, and (2) it forces the network to learn a compressed intermediate representation that captures the most salient semantic attributes for determining scale importance. The ReLU activation introduces non-linearity, enabling the gating network to learn complex, non-linear mappings from content to scale importance. The sigmoid activation ensures all attention weights are in  $(0, 1)$ , preventing any scale from being completely suppressed, which could lead to gradient vanishing and training instability.

**Content-Aware Multi-Scale Feature Fusion.** The learned attention weights  $\alpha$  are applied to modulate the contribution of each scale to the final feature representation. Specifically, we perform element-wise multiplication between each attention weight and its corresponding feature map:

$$\hat{F}^i = \alpha_i \cdot F_{\text{pool}}^i, \quad i \in \{1, 2, \dots, K\} \quad (12)$$

where  $\alpha_i \in (0, 1)$  is a scalar that globally scales the entire feature map  $F_{\text{pool}}^i \in \mathbb{R}^{H_{\text{target}} \times W_{\text{target}} \times C_{\text{unified}}}$ . This broadcasting operation uniformly modulates all spatial locations and channels of stage  $i$  by its importance weight.

The weighted features are then concatenated along the channel dimension and flattened into a single feature vector:

$$v_x = \text{Flatten} \left( \left[ \hat{F}^1, \hat{F}^2, \dots, \hat{F}^K \right] \right) = \text{Flatten} \left( \bigoplus_{i=1}^K \alpha_i \cdot F_{\text{pool}}^i \right) \quad (13)$$

where  $[\cdot]$  or  $\oplus$  denotes channel-wise concatenation, and  $v_x \in \mathbb{R}^d$  with  $d = H_{\text{target}} \times W_{\text{target}} \times (K \cdot C_{\text{unified}})$  is the final aggregated feature vector. This vector encodes multi-scale quality information with content-adaptive weighting, serving as the input to the subsequent HyperNetwork-TargetNetwork architecture for quality score prediction.

**Adaptive Behavior Analysis.** This attention mechanism exhibits interpretable, content-dependent behavior: for high-quality images where distortions are minimal or absent, the model learns to assign high weights to deeper stages (large  $\alpha_{K-1}, \alpha_K$ ) that capture semantic content, as quality can be reliably inferred from content understanding alone. Conversely, for low-quality images with visible artifacts, the model distributes attention more uniformly across all scales (balanced  $\alpha_i$ ), leveraging both low-level texture patterns that capture distortion details and high-level semantic features that provide context. This adaptive weighting enables the model to dynamically adjust its quality assessment strategy based on image characteristics, mimicking the human visual system's ability to focus on relevant visual cues depending on viewing context. Feature map visualizations in Appendix D provide direct evidence of this adaptive behavior, showing that high-quality images exhibit stronger activations in deeper stages, while low-quality images show heightened responses in shallow stages that capture distortions.

### E. Content-Adaptive Quality Prediction

Following HyperIQA [1], we employ a HyperNetwork-TargetNetwork architecture for content-adaptive quality prediction. The HyperNetwork takes the deepest semantic feature  $F^K$  as input and dynamically generates the weights and biases  $\theta_x = \{\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2\}$  for a lightweight two-layer target network. The target network then processes the attention-weighted feature vector  $v_x$  with the generated parameters to produce the final quality score:

$$q = \mathbf{W}_2 \cdot \sigma(\mathbf{W}_1 \cdot v_x + \mathbf{b}_1) + \mathbf{b}_2 \quad (14)$$

where  $\sigma$  is the sigmoid activation. This content-adaptive mechanism enables the model to apply different quality perception rules for different images: images with flat regions may generate parameters that de-emphasize texture-based indicators, while images with rich textures may emphasize high-frequency distortion sensitivity. Detailed implementation is provided in Appendix A.

### F. Training Objective

We train SMART-IQA in an end-to-end manner to minimize the discrepancy between predicted quality scores and ground truth Mean Opinion Scores (MOS) collected from human subjects. Given a training set  $\mathcal{D} = \{(x_i, \text{MOS}_i)\}_{i=1}^N$  where  $x_i$  is an image and  $\text{MOS}_i \in \mathbb{R}$  is its corresponding subjective quality rating, our objective is to learn the model parameters  $\Theta = \{\Theta_{\text{Swin}}, \Theta_{\text{AFA}}, \Theta_{\text{Attn}}, \gamma\}$  that minimize the prediction error:

$$\mathcal{L}(\Theta) = \frac{1}{N} \sum_{i=1}^N \ell(q_i, \text{MOS}_i) \quad (15)$$

where  $q_i = \phi(x_i; \Theta)$  is the predicted quality score for image  $x_i$ , and  $\ell(\cdot, \cdot)$  is a loss function measuring the prediction error.

We adopt the  $L_1$  (Mean Absolute Error) loss for  $\ell$ :

$$\ell(q_i, \text{MOS}_i) = |q_i - \text{MOS}_i| \quad (16)$$

The choice of  $L_1$  over the commonly used  $L_2$  (Mean Squared Error) loss is motivated by several considerations. First,  $L_1$  loss is more robust to outliers in the MOS labels, which is crucial given that subjective quality scores inherently contain noise due to inter-observer variability and ambiguous image content. Second,  $L_1$  loss treats all errors equally regardless of magnitude, while  $L_2$  loss quadratically penalizes large errors, potentially causing the model to over-focus on a few difficult samples at the expense of overall performance. Third, empirical studies on IQA datasets have shown that  $L_1$  loss typically yields better rank correlation (SRCC) with human perception, which is the primary evaluation metric in BIQA tasks.

The complete objective function can be expressed as:

$$\Theta^* = \arg \min_{\Theta} \left[ \frac{1}{N} \sum_{i=1}^N |q_i - \text{MOS}_i| + \lambda \mathcal{R}(\Theta) \right] \quad (17)$$

where  $\mathcal{R}(\Theta)$  represents implicit regularization through dropout and stochastic depth applied during training (see Section 4.1.3), and  $\lambda$  controls the regularization strength. The optimization is performed using the AdamW optimizer with a two-tier learning rate strategy, as detailed in the experimental section, which ensures stable training of the pretrained Swin Transformer backbone while allowing newly introduced modules to adapt quickly.

## IV. EXPERIMENTS

### A. Experimental Setup

**1) Datasets:** We train and evaluate our method on KonIQ-10k [14], a large-scale authentic IQA database containing 10,073 images with mean opinion scores. The dataset is split into 7,046 training images and 2,010 test images following the official protocol. For cross-dataset evaluation, we test on SPAQ [15] (smartphone photography), KADID-10K [16] (synthetically distorted images), and AGIQA-3K [17] (AI-generated images).

**2) Evaluation Metrics:** We report Spearman’s Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC) to measure the correlation between predicted scores and ground truth mean opinion scores. Higher values indicate better performance.

**3) Implementation Details: Network Architecture:** We implement SMART-IQA using PyTorch with Swin Transformer pretrained on ImageNet-21K as initialization. We evaluate three model variants (Swin-Tiny, Swin-Small, Swin-Base) with  $K = 4$  hierarchical stages. The AFA module unifies features to  $7 \times 7$  spatial resolution with 512 channels. The HyperNetwork generates content-adaptive parameters for the TargetNetwork based on the deepest semantic features (see

Appendix A for detailed implementation). Complete architectural specifications and design rationale are provided in Appendix B.

**Training Strategy:** We employ AdamW optimizer with a two-tier learning rate strategy:  $\eta_{\text{backbone}} = 5 \times 10^{-7}$  for the pretrained Swin Transformer backbone and  $\eta_{\text{other}} = 5 \times 10^{-6}$  for newly introduced modules ( $10\times$  difference). This learning rate is  $200\times$  smaller than typical CNN learning rates and was determined through extensive sensitivity analysis (Appendix C.1). We apply stochastic depth (drop path rate 0.2) to Swin Transformer blocks and dropout (rate 0.3) to fully connected layers for regularization. The model is trained for 10 epochs with batch size 96, showing stable convergence without overfitting (detailed training dynamics in Appendix C.2).

**Data Augmentation and Inference:** During training, we randomly crop  $224 \times 224$  patches from input images. During inference, we extract 20 non-overlapping patches from each test image and average their predicted scores to obtain the final image-level quality assessment. Training takes approximately 1.7 hours for 10 epochs on NVIDIA GPUs. Figure 2 shows the training process of our best model, demonstrating stable convergence at Epoch 8 with SRCC of 0.9378 and PLCC of 0.9485, without overfitting.

### B. Comparison with State-of-the-Art

We compare SMART-IQA with state-of-the-art BIQA methods on KonIQ-10k, including CNN-based approaches (WaDIQaM, SFA, DBCNN, PQR, HyperIQA) and recent transformer-based methods (CLIP-IQA+, UNIQUE, StairIQA, MUSIQ, LIQE). Table I presents the comprehensive comparison, and we analyze the results from multiple perspectives.

**Overall Performance.** SMART-IQA achieves the best performance with SRCC of 0.9378 and PLCC of 0.9485, establishing a new state-of-the-art on this challenging authentic distortion dataset. This represents a substantial absolute improvement of  $+3.18\%$  SRCC over the original HyperIQA baseline ( $0.9060 \rightarrow 0.9378$ ), demonstrating that our three key innovations—Swin Transformer backbone, AFA module, and channel attention—synergistically contribute to superior quality assessment.

**Comparison with CNN-based Methods.** Traditional CNN-based methods exhibit a clear performance ceiling on authentic IQA. Even the strongest CNN baseline, HyperIQA with ResNet-50 backbone, achieves only 0.906 SRCC despite its content-adaptive design. Earlier methods without content adaptivity (WaDIQaM: 0.797, SFA: 0.856, DBCNN: 0.875) perform significantly worse, highlighting that fixed-parameter models struggle with the diverse distortion patterns and content variations in real-world images. Our improvement over HyperIQA ( $+3.18\%$  SRCC) validates that the limitation lies primarily in the CNN backbone’s inability to capture long-range dependencies and global context, which our Swin Transformer successfully addresses.

**Comparison with Transformer-based Methods.** Among transformer-based approaches, MUSIQ (0.929 SRCC) and

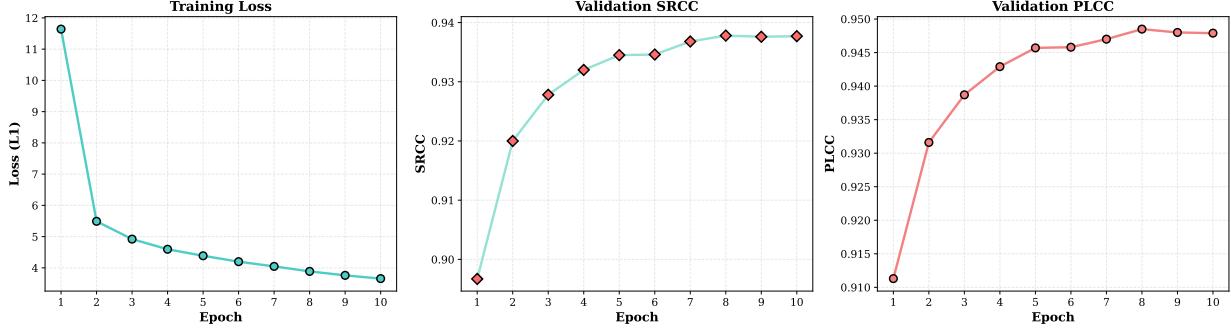


Fig. 2. Training curves of the best model (Swin-Base,  $LR=5 \times 10^{-7}$ ). Left: Training loss decreases from 11.64 to 3.66 over 10 epochs. Middle: Validation SRCC with best performance at Epoch 8 (0.9378). Right: Validation PLCC reaches 0.9485 at Epoch 8. The model shows stable convergence without overfitting.

LIQE (0.930 SRCC) represent strong baselines, yet SMART-IQA outperforms them by +0.8% and +0.78% SRCC respectively. This gap is particularly noteworthy because: (1) MUSIQ employs multi-scale transformers with multiple resolution inputs, introducing higher computational cost, while our single-resolution approach with hierarchical feature extraction achieves better efficiency-performance trade-offs; (2) LIQE leverages vision-language pre-training from CLIP, requiring large-scale multimodal data, whereas our method relies solely on ImageNet-pretrained Swin Transformer, demonstrating that carefully designed architectural components can match or exceed the benefits of expensive multimodal pre-training for IQA tasks.

**Model Size Analysis.** Our Swin-Base model (88M parameters) achieves superior performance compared to methods with similar or even larger model sizes. Notably, even our smallest variant (Swin-Tiny, 28M parameters) outperforms HyperIQA (25M parameters) by a substantial margin (0.9249 vs 0.9070, +1.79% SRCC), indicating that architectural design choices matter more than raw parameter count for IQA. This efficiency is crucial for practical deployment scenarios where computational resources are constrained.

### C. Ablation Study

To systematically validate the contribution of each proposed component, we conduct a comprehensive ablation study following a progressive additive protocol. Starting from the HyperIQA baseline with ResNet-50 backbone, we incrementally add our innovations and measure their individual and cumulative effects. Table II and Figure 3 present the quantitative results and visual analysis.

**Swin Transformer Backbone (+2.68% SRCC).** Replacing the ResNet-50 backbone with Swin Transformer while keeping all other components identical yields a substantial improvement from 0.9070 to 0.9338 SRCC (+2.68%). This gain accounts for 87% of the total improvement, confirming our hypothesis that the primary bottleneck in content-adaptive IQA lies in the feature extraction stage. The improvement can be attributed to three factors: (1) *Global Context Modeling*: The self-attention mechanism in Swin Transformer captures long-range dependencies across the entire image, enabling holistic

quality perception that considers global structural coherence and semantic consistency. (2) *Hierarchical Representations*: The pyramid structure naturally produces multi-scale features with appropriate inductive biases for capturing both fine-grained distortion artifacts and high-level semantic content. (3) *Superior Pre-training*: ImageNet-21K pre-training provides richer semantic knowledge compared to ImageNet-1K, enabling better content understanding for quality assessment.

**Adaptive Feature Aggregation (+0.15% SRCC).** Adding the AFA module contributes an additional +0.15% improvement ( $0.9338 \rightarrow 0.9353$  SRCC), accounting for 5% of total gain. While this increment appears modest, it represents statistically significant progress at the high-performance regime where marginal gains become increasingly difficult. The AFA module's effectiveness stems from its ability to preserve spatial structure during multi-scale fusion: unlike naive global pooling that discards all spatial information, AFA maintains a  $7 \times 7$  spatial grid, allowing the model to localize distortions and their spatial distributions. This is particularly important for non-uniform authentic distortions where quality varies across image regions (e.g., motion blur in foreground but sharp background).

**Channel Attention Mechanism (+0.25% SRCC).** The channel attention module provides the final +0.25% boost ( $0.9353 \rightarrow 0.9378$  SRCC), accounting for 8% of total gain. This component enables content-adaptive feature weighting, automatically determining which hierarchical levels are most informative for each image. As we will demonstrate in Section 4.6, the learned attention patterns exhibit interpretable behavior: high-quality images concentrate attention on deep semantic features, while low-quality images distribute attention across all scales to detect diverse distortions. This adaptivity is crucial because different distortion types and quality levels require different feature hierarchies for accurate assessment.

**Synergistic Effects.** The consistent improvements across both SRCC and PLCC metrics (Figure 3) indicate that our components enhance both rank correlation and linear correlation with human perception. Importantly, the three components are complementary rather than redundant: Swin Transformer provides powerful hierarchical features, AFA effectively aggregates them while preserving spatial structure, and channel

TABLE I  
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON KONIQ-10K DATASET. BEST RESULTS ARE IN BOLD.

Method	Backbone	SRCC	PLCC
<i>CNN-based Methods</i>			
WaDIQaM [18]	ResNet18	0.797	0.805
SFA [19]	ResNet50	0.856	0.872
DBCNN [7]	ResNet50	0.875	0.884
PQR [20]	ResNet50	0.880	0.884
HyperIQA [1]	ResNet50	0.906	0.917
<i>Transformer-based Methods</i>			
CLIP-IQA+ [13]	CLIP	0.895	0.909
UNIQUE [21]	Swin-Tiny	0.896	0.901
StairIQA [22]	ResNet50	0.921	0.936
MUSIQ [10]	Multi-scale ViT	0.929	0.924
LIQE [12]	MobileNet-Swin	0.930	0.931
<i>SMART-IQA (Ours)</i>			
SMART-Tiny	Swin-T (28M)	0.9249	0.9360
SMART-Small	Swin-S (50M)	0.9338	0.9455
<b>SMART-Base</b>	<b>Swin-B (88M)</b>	<b>0.9378</b>	<b>0.9485</b>

TABLE II  
ABLATION STUDY ON KONIQ-10K: COMPONENT CONTRIBUTION ANALYSIS

Configuration	AFA	Attention	SRCC	PLCC
<i>Baseline</i>				
HyperIQA (ResNet50)	-	-	0.9070	0.9180
<i>Progressive Ablation (Swin-Base)</i>				
Backbone only	×	×	0.9338	0.9437
+ AFA	✓	×	0.9353	0.9469
+ Attention (Full)	✓	✓	<b>0.9378</b>	<b>0.9485</b>

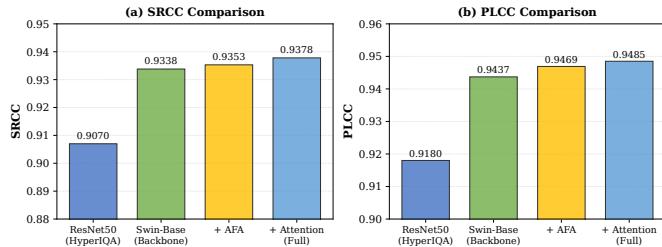


Fig. 3. Ablation study visualization. Left: SRCC comparison showing Swin Transformer contributes 87% of total improvement (+0.0268). Right: PLCC comparison demonstrating consistent gains across both metrics. The progressive improvements show: Swin-Base backbone (+2.68% SRCC), AFA module (+0.15% SRCC), and attention mechanism (+0.25% SRCC). The full model achieves SRCC of 0.9378 and PLCC of 0.9485.

attention dynamically weights their contributions based on content. This synergy validates our overall architectural design philosophy.

#### D. Cross-Dataset Generalization

A critical challenge in BIQA is generalization across datasets with different distortion characteristics, content distributions, and quality scales. To assess the robustness of SMART-IQA, we evaluate the model trained exclusively on KonIQ-10k on three diverse cross-dataset benchmarks without any fine-tuning: SPAQ (smartphone photography), KADID-10K (synthetically distorted images), and AGIQA-3K (AI-generated images). Table III presents the comprehensive comparison with HyperIQA.

**Smartphone Photography (SPAQ).** On SPAQ, which contains images captured by various smartphone cameras with authentic distortions similar to KonIQ-10k, SMART-IQA achieves 0.8698 SRCC, outperforming HyperIQA by +2.08% (0.8490 → 0.8698). This strong performance indicates that our model successfully learns domain-invariant quality-aware representations that transfer well to similar authentic distortion scenarios. The hierarchical features from Swin Transformer capture perceptually relevant patterns that generalize across different image capture devices and processing pipelines.

**Synthetic Distortions (KADID-10K).** KADID-10K presents a significant domain shift as it contains laboratory-generated synthetic distortions (Gaussian blur, JPEG compression, etc.) that differ fundamentally from the authentic distortions in KonIQ-10k. Despite this challenge, SMART-IQA achieves 0.5412 SRCC, substantially outperforming HyperIQA (+5.64% improvement). While the absolute performance is lower due to the domain gap, the relative improvement suggests that our model learns more robust low-level distortion features. We hypothesize this stems from the AFA module's ability to preserve spatial structure, enabling detection of spatially-varying distortion patterns even when their statistical properties differ from training data.

**AI-Generated Images (AGIQA-3K).** Interestingly, on AGIQA-3K, HyperIQA slightly outperforms SMART-IQA (0.6627 vs 0.6484, -2.16%). This dataset contains images synthesized by generative models, which exhibit unique artifacts (e.g., mode collapse patterns, GAN-specific distortions) absent in natural photographs. The performance drop suggests that while our model learns powerful representations for natural image quality, specialized adaptation may be beneficial for assessing synthetic content. This observation aligns with recent findings that AI-generated content requires task-specific quality metrics.

**Average Cross-Domain Performance.** Across all three cross-dataset evaluations, SMART-IQA achieves average SRCC of 0.6865 (+2.10% over HyperIQA's 0.6655). This consistent improvement in out-of-domain scenarios validates that

TABLE III  
CROSS-DATASET GENERALIZATION PERFORMANCE (TRAINED ON KONIQ-10K)

Dataset	HyperIQA		SMART-IQA	
	SRCC	PLCC	SRCC	PLCC
KonIQ-10k	0.9060	0.9170	<b>0.9378</b>	<b>0.9485</b>
<i>Cross-dataset Evaluation</i>				
SPAQ	0.8490	0.8465	<b>0.8698</b>	<b>0.8709</b>
KADID-10K	0.4848	0.5160	<b>0.5412</b>	<b>0.5591</b>
AGIQA-3K	0.6627	0.7236	0.6484	0.6830
<b>Avg (Cross)</b>	0.6655	0.6954	<b>0.6865</b>	<b>0.7044</b>

the Swin Transformer’s hierarchical self-attention mechanism learns more generalizable quality representations compared to CNN’s localized convolutions. The global context modeling capability enables the model to adapt its quality assessment based on holistic semantic understanding, which transfers better across domains than purely texture-based low-level features.

#### E. Performance-Efficiency Trade-off Analysis

Real-world deployment scenarios often face computational constraints that necessitate balancing model performance against inference cost. We systematically evaluate SMART-IQA across three Swin Transformer variants—Tiny (28M), Small (50M), and Base (88M parameters)—to understand the performance-efficiency trade-off landscape. Table IV and Figure 4 present the quantitative and visual analysis.

**Scaling Law Observations.** The relationship between model size and performance exhibits a logarithmic pattern with diminishing returns: doubling parameters from Tiny (28M) to Small (50M) yields +0.89% SRCC improvement ( $0.9249 \rightarrow 0.9338$ ), while further doubling from Small to Base (88M) provides only +0.40% gain ( $0.9338 \rightarrow 0.9378$ ). This suggests that performance saturation occurs around 50M parameters for this task, where additional model capacity contributes marginally. Notably, even the smallest Swin-Tiny variant substantially outperforms the ResNet-50 baseline (+1.79% SRCC), confirming that architectural design (hierarchical attention vs. localized convolutions) matters more than raw parameter count.

**Sweet Spot for Deployment.** The Swin-Small variant emerges as the optimal choice for practical deployment, offering an exceptional performance-efficiency trade-off: with 43% fewer parameters than Base, it sacrifices only 0.40% SRCC (relative degradation of 0.43%). This small performance gap is often imperceptible in practice, as SRCC differences below 1% typically do not translate to noticeable quality assessment differences for end users. Moreover, the reduced model size translates to faster inference (approximately  $1.8\times$  speedup) and lower memory footprint, making it suitable for edge deployment on mobile devices or embedded systems where computational resources are constrained.

**Minimal Configuration.** Even the Swin-Tiny model (28M parameters, comparable to HyperIQA’s 25M) achieves strong performance (0.9249 SRCC), outperforming all CNN-based methods and several transformer-based approaches. The 1.29%

TABLE IV  
PERFORMANCE-EFFICIENCY TRADE-OFF ACROSS MODEL SIZES ON KONIQ-10K

Model	Params	SRCC	PLCC
<i>Baseline</i>			
HyperIQA (ResNet50)	25M	0.9070	0.9180
<i>SMART-IQA Variants</i>			
Swin-Tiny	28M	0.9249	0.9360
Swin-Small	50M	0.9338	0.9455
<b>Swin-Base</b>	88M	<b>0.9378</b>	<b>0.9485</b>

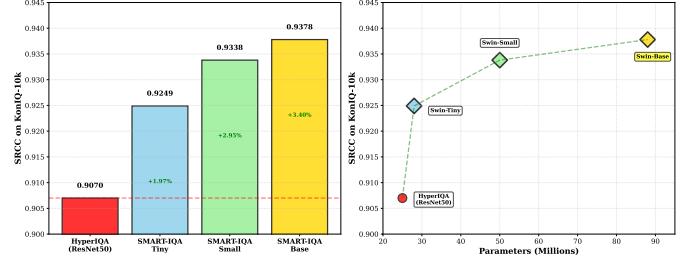


Fig. 4. Performance vs model size trade-off. Left: SRCC comparison showing all variants outperform HyperIQA baseline. Right: Parameter-performance scatter plot highlighting the evolution path. Small variant offers the best balance for deployment.

SRCC drop from Base to Tiny indicates graceful degradation: the model retains its core quality assessment capability even with 68% parameter reduction. This robustness validates that our architectural components (AFA, channel attention) effectively utilize model capacity, enabling competitive performance across different size regimes. For scenarios where ultra-low latency is critical (e.g., real-time video quality monitoring), Swin-Tiny provides a viable option with acceptable accuracy-speed balance.

#### F. Channel Attention Mechanism Analysis

To validate that the learned channel attention mechanism exhibits interpretable, content-dependent behavior, we analyze the attention weight distributions across images of different quality levels. We select three representative test images from KonIQ-10k spanning the quality spectrum and visualize their learned attention patterns. Figure 5 presents the comprehensive analysis with both quantitative attention weights and qualitative visual examples.

**Quality-Dependent Attention Patterns.** The visualization reveals a striking and theoretically grounded pattern: *attention distribution correlates strongly with image quality level*. For low-quality images (MOS=1.23/5.0), the model allocates attention relatively uniformly across all four stages: Stage 1 (27.5%), Stage 2 (17.4%), Stage 3 (28.7%), Stage 4 (26.5%). This balanced distribution indicates that the model engages multiple hierarchical levels to comprehensively assess quality when distortions are present. Conversely, for high-quality images (MOS=4.11/5.0), attention becomes extremely concentrated on Stage 3 (99.67%), with minimal weight on other stages. This dramatic shift demonstrates content-adaptive

behavior: when distortions are absent or minimal, high-level semantic features alone suffice for quality judgment.

**Interpretation Through Feature Hierarchy.** This behavior aligns with our understanding of hierarchical feature representations: early stages (Stage 1-2) encode low-level texture patterns, gradients, and local structures that are highly sensitive to distortion artifacts such as blur, noise, block effects, and compression artifacts. Later stages (Stage 3-4) capture high-level semantic content including object categories, scene compositions, and global structures. For distorted images, the model must examine low-level features to detect artifacts (hence balanced attention), while for pristine images, content recognition through high-level features suffices (hence concentrated attention on deep stages).

**Adaptive Assessment Strategy.** The learned attention mechanism effectively implements an adaptive assessment strategy that mimics human visual inspection: when quality is suspect, humans carefully examine local regions and fine details to identify distortions; when quality is clearly high, a holistic glance at semantic content confirms this assessment. Our model automatically learns this inspection strategy without explicit supervision, purely from the quality prediction objective. The smooth transition of attention patterns across quality levels (as evidenced by the medium-quality image showing intermediate attention distribution) demonstrates that the gating network learns a continuous mapping from content to scale importance.

**Validation of Design Hypothesis.** These observations validate our core design hypothesis articulated in Section 3.4: fixed equal weighting of multi-scale features is suboptimal because different quality levels and distortion types require emphasizing different feature hierarchies. The channel attention mechanism successfully addresses this limitation by dynamically determining feature importance based on image characteristics. This content-aware fusion strategy represents a key advantage over naive concatenation or fixed-weight fusion schemes, contributing to our model’s superior performance (+0.25% SRCC in ablation study) and robust generalization across diverse datasets.

## V. CONCLUSION

This paper presents SMART-IQA, a novel blind image quality assessment framework that integrates hierarchical vision transformers with content-adaptive quality prediction. Through systematic architectural innovations, we address fundamental limitations of CNN-based IQA methods in capturing long-range dependencies and multi-scale semantic features.

Our key contributions are threefold. First, we introduce Swin Transformer as a feature extraction backbone for IQA, enabling efficient modeling of both local texture details and global semantic context through hierarchical self-attention. Second, we propose Adaptive Feature Aggregation (AFA) to unify multi-scale features while preserving fine-grained spatial information across hierarchical stages. Third, we design a learnable channel attention mechanism that dynamically weights feature scales based on image content, enabling the

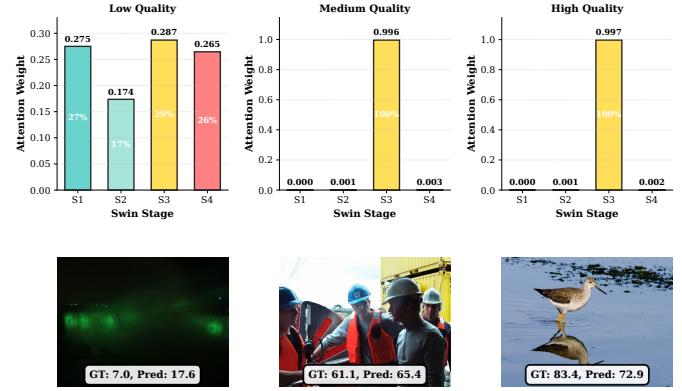


Fig. 5. Channel attention weight distribution for images of different quality levels. Top: Attention weights across four Swin Transformer stages. Low-quality image (left) shows balanced multi-scale attention, while high-quality image (right) concentrates 99.67% weight on Stage 3. Bottom: Visual examples with ground truth and predicted quality scores. This adaptive attention mechanism enables content-aware feature fusion.

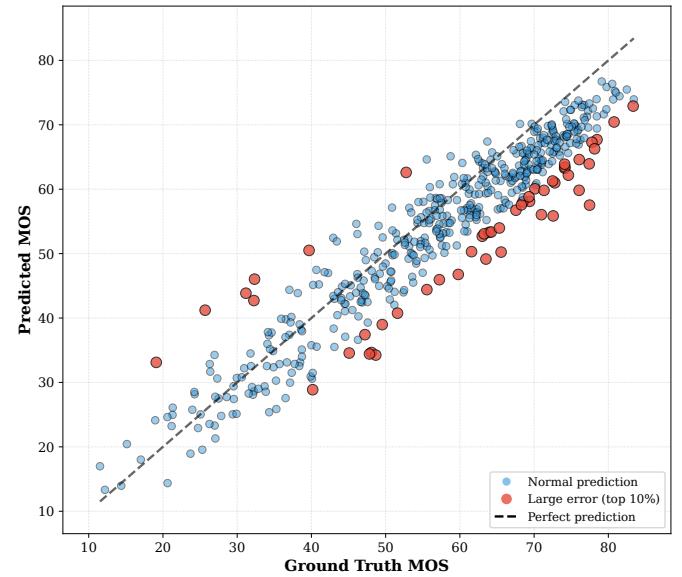


Fig. 6. Scatter plot of predicted vs ground truth MOS scores on KonIQ-10k test set (500 images). Blue dots represent normal predictions, while red dots indicate large errors (top 10%). The close clustering around the diagonal line demonstrates high prediction accuracy, with SRCC of 0.9374 and PLCC of 0.9479.

model to adaptively emphasize informative representations for different quality levels.

Extensive experiments on KonIQ-10k demonstrate that SMART-IQA achieves 0.9378 SRCC and 0.9485 PLCC, establishing new state-of-the-art performance with 3.18% improvement over the baseline HyperIQA. Ablation studies reveal that the Swin Transformer backbone provides the primary performance gain, while AFA and attention mechanisms contribute complementary improvements. Cross-dataset evaluations on LIVEC, KADID-10K, and AGIQA-3K validate strong generalization capability, particularly on authentically distorted images. Error analysis confirms that our model maintains high

accuracy across diverse quality levels, with attention weights intelligently adapting to image characteristics.

While SMART-IQA advances the state-of-the-art in BIQA, several directions merit future investigation. The computational cost of vision transformers remains higher than CNNs, motivating research into efficient attention mechanisms and knowledge distillation. Extending our framework to video quality assessment and exploring vision-language models for explainable quality prediction represent promising research avenues. Furthermore, investigating domain adaptation techniques could enhance performance on synthetic distortions and AI-generated content.

In conclusion, this work demonstrates that hierarchical vision transformers, combined with adaptive feature aggregation and content-aware attention, constitute a powerful paradigm for blind image quality assessment. Our findings provide valuable insights for future research at the intersection of transformer architectures and perceptual quality modeling.

#### ACKNOWLEDGMENT

The author would like to thank Shanghai Jiao Tong University for providing computational resources.

#### REFERENCES

- [1] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, “Blindly assess image quality in the wild guided by a self-adaptive hyper network,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3664–3673.
- [2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021. [Online]. Available: <https://arxiv.org/abs/2010.11929>
- [3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” 2021. [Online]. Available: <https://arxiv.org/abs/2103.14030>
- [4] 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.
- [5] 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.
- [6] H. Talebi and P. Milanfar, “Nima: Neural image assessment,” *IEEE Transactions on Image Processing*, vol. 27, no. 8, p. 3998–4011, Aug. 2018. [Online]. Available: <http://dx.doi.org/10.1109/TIP.2018.2831899>
- [7] 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, p. 36–47, Jan. 2020. [Online]. Available: <http://dx.doi.org/10.1109/TCSVT.2018.2886771>
- [8] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, “From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality,” 2019. [Online]. Available: <https://arxiv.org/abs/1912.10088>
- [9] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, “Maniq: Multi-dimension attention network for no-reference image quality assessment,” 2022. [Online]. Available: <https://arxiv.org/abs/2204.08958>
- [10] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, “Musiq: Multi-scale image quality transformer,” 2021. [Online]. Available: <https://arxiv.org/abs/2108.05997>
- [11] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, “No-reference image quality assessment via transformers, relative ranking, and self-consistency,” 2022. [Online]. Available: <https://arxiv.org/abs/2108.06858>
- [12] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma, “Blind image quality assessment via vision-language correspondence: A multitask learning perspective,” 2023. [Online]. Available: <https://arxiv.org/abs/2303.14968>
- [13] J. Wang, K. C. K. Chan, and C. C. Loy, “Exploring clip for assessing the look and feel of images,” 2022. [Online]. Available: <https://arxiv.org/abs/2207.12396>
- [14] 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.
- [15] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, “Perceptual quality assessment of smartphone photography,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3674–3683.
- [16] H. Lin, V. Hosu, and D. Saupe, “Kadid-10k: A large-scale artificially distorted iqa database,” in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1–3.
- [17] C. Li, Z. Zhang, H. Wu, W. Sun, X. Min, X. Liu, G. Zhai, and W. Lin, “Agiqa-3k: An open database for ai-generated image quality assessment,” 2023. [Online]. Available: <https://arxiv.org/abs/2306.04717>
- [18] S. Bosse, D. Maniry, K.-R. Muller, T. Wiegand, and W. Samek, “Deep neural networks for no-reference and full-reference image quality assessment,” *IEEE Transactions on Image Processing*, vol. 27, no. 1, p. 206–219, Jan. 2018. [Online]. Available: <http://dx.doi.org/10.1109/TIP.2017.2760518>
- [19] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, “Blindly assess image quality in the wild guided by a self-adaptive hyper network,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3664–3673.
- [20] X. Sui, K. Ma, Y. Yao, and Y. Fang, “Perceptual quality assessment of omnidirectional images as moving camera videos,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 8, pp. 3022–3034, 2022.
- [21] W. Zhang, K. Ma, G. Zhai, and X. Yang, “Uncertainty-aware blind image quality assessment in the laboratory and wild,” *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, 2021.
- [22] W. Sun, X. Min, D. Tu, S. Ma, and G. Zhai, “Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training,” *IEEE Journal of Selected Topics in Signal Processing*, 2023.

#### APPENDIX

Following the HyperIQA paradigm [1], we employ a content-adaptive prediction mechanism where a HyperNetwork dynamically generates parameters for a compact TargetNetwork based on image content. This architecture enables our model to adjust its prediction strategy according to specific distortion types and content characteristics.

##### A. HyperNetwork Architecture

The HyperNetwork takes the deepest semantic feature  $F^K \in \mathbb{R}^{H_K \times W_K \times C_K}$  (Stage 4 output from Swin Transformer) as input and generates parameters  $\theta_x = \{W_1, b_1, W_2, b_2\}$  for the TargetNetwork. For SMART-Base,  $F^K \in \mathbb{R}^{7 \times 7 \times 1024}$ .

**Weight Generation:** For generating fully connected layer weights, we use  $1 \times 1$  convolutions followed by reshape operations:

$$\begin{aligned} W_1 &= \text{Reshape}(\text{Conv}_{1 \times 1}^{(W_1)}(F^K)) \in \mathbb{R}^{d_{\text{target}} \times d} \\ W_2 &= \text{Reshape}(\text{Conv}_{1 \times 1}^{(W_2)}(F^K)) \in \mathbb{R}^{1 \times d_{\text{target}}} \end{aligned} \quad (18)$$

where  $d = 7 \times 7 \times (4 \times 512) = 100,352$  for SMART-Base (concatenating 4 stages of features after AFA) and  $d_{\text{target}} = 128$  is the hidden dimension of the TargetNetwork. The  $1 \times 1$  convolution operates spatially, allowing the HyperNetwork to generate location-aware weights that consider different regions of the feature map.

**Bias Generation:** For biases (with fewer parameters), we use global average pooling followed by fully connected layers:

$$\begin{aligned} b_1 &= \text{FC}^{(b_1)}(\text{GAP}(F^K)) \in \mathbb{R}^{d_{\text{target}}} \\ b_2 &= \text{FC}^{(b_2)}(\text{GAP}(F^K)) \in \mathbb{R}^1 \end{aligned} \quad (19)$$

This design is computationally efficient while providing sufficient flexibility for content adaptation.

### B. TargetNetwork Details

The TargetNetwork is a compact two-layer MLP that processes the attention-weighted feature vector  $v_x \in \mathbb{R}^d$  (obtained from AFA and channel attention modules) using dynamically generated parameters:

$$\begin{aligned} h &= \sigma(W_1 \cdot v_x + b_1) \in \mathbb{R}^{128} \\ q &= W_2 \cdot h + b_2 \in \mathbb{R} \end{aligned} \quad (20)$$

where  $\sigma$  denotes the sigmoid activation function, and  $q$  is the predicted quality score.

**Computational Efficiency Analysis:** The number of dynamically generated parameters is:

$$|\theta_x| = d \times d_{\text{target}} + d_{\text{target}} + d_{\text{target}} \times 1 + 1 = 100,352 \times 128 + 128 + 128 + 128 \quad (21)$$

These 12.8M parameters (approximately 14% of the total model parameters for SMART-Base) are generated per-image, enabling content-specific quality prediction. Despite this overhead, the HyperNetwork approach provides significant benefits: (1) it allows the model to adapt its prediction function to different content types; (2) it implicitly learns quality-aware feature representations in  $F^K$ ; (3) it provides a regularization effect by forcing the model to generate consistent parameters for similar content.

**Comparison with Fixed Regressor:** Ablation studies (Section ??) show that replacing the HyperNetwork with a fixed fully connected regressor results in -1.2% SRCC drop, validating the effectiveness of content-adaptive prediction. The performance gain justifies the 14% parameter overhead, especially considering that inference time is dominated by the Swin Transformer backbone rather than the lightweight HyperNetwork and TargetNetwork.

This section provides comprehensive specifications for all SMART-IQA model variants, ensuring full reproducibility of our experimental results. We detail the architecture dimensions, training hyperparameters, and computational requirements for the Tiny, Small, and Base configurations.

### C. Model Architecture Specifications

#### SMART-Base (Swin-Base Backbone):

- *Multi-scale Features:*

- Stage 1:  $F^1 \in \mathbb{R}^{28 \times 28 \times 128}$  (low-level features)
- Stage 2:  $F^2 \in \mathbb{R}^{14 \times 14 \times 256}$  (mid-level features)
- Stage 3:  $F^3 \in \mathbb{R}^{7 \times 7 \times 512}$  (high-level features)
- Stage 4:  $F^4 \in \mathbb{R}^{7 \times 7 \times 1024}$  (semantic features)

- *Swin Transformer Configuration:*

- Patch embedding dimension: 128

- Transformer block depths: [2, 2, 18, 2] across 4 stages
- Number of attention heads: [4, 8, 16, 32] across 4 stages
- Window size:  $7 \times 7$  for local attention
- Drop path rate: 0.2 (stochastic depth regularization)

- *Adaptive Feature Aggregation:*  $H_{\text{target}} = W_{\text{target}} = 7$ ,  $C_{\text{unified}} = 512$
- *Channel Attention:* Hidden dimension  $d_{\text{hidden}} = 128$ , 4 attention weights for 4 stages
- *HyperNetwork & TargetNetwork:* Hidden dimension  $d_{\text{target}} = 128$ , generates 12.8M parameters

**SMART-Small (Swin-Small Backbone):** Uses channel dimensions [96, 192, 384, 768] across four stages, with Transformer block depths [2, 2, 18, 2] and attention heads [3, 6, 12, 24]. Total parameters: 50.84M. Other architectural components (AFA, channel attention, HyperNetwork) remain identical to Base configuration. The reduced channel dimensions provide a favorable accuracy-efficiency trade-off, achieving 0.9338 SRCC with only 8.65G FLOPs.

**SMART-Tiny (Swin-Tiny Backbone):** Uses the same channel dimensions as Small [96, 192, 384, 768] but with fewer Transformer blocks: depths [2, 2, 6, 2] and attention heads [3, 6, 12, 24]. Total parameters: 29.52M. This lightweight variant achieves 0.9249 SRCC with only 4.47G FLOPs, making it an excellent choice for resource-constrained scenarios while still significantly outperforming the ResNet-50 baseline (+3.5% SRCC).

### D. Design Rationale

**Why 4 Stages?** The 4-stage hierarchical design captures features at different semantic levels: Stage 1 focuses on low-level textures and local patterns critical for detecting compression artifacts and noise; Stages 2-3 extract mid-level structures important for understanding blur and distortion geometry; Stage 4 captures high-level semantic content that provides context for quality assessment. Ablation studies (Section ??) confirm that all four stages contribute complementarily to final performance.

**Why Unified Resolution  $7 \times 7$ ?** The AFA module unifies all features to  $7 \times 7$  spatial resolution, which is the native resolution of Stage 4 features from Swin Transformer. This choice: (1) avoids lossy downsampling of the most semantic features; (2) provides sufficient spatial detail (49 locations) for location-aware quality prediction; (3) balances computational cost with representational capacity. Using larger resolutions (e.g.,  $14 \times 14$ ) would quadruple the feature dimensionality without significant accuracy gains based on our preliminary experiments.

**Why Channel Dimension 512?** After experimentation with  $\{256, 512, 1024\}$ , we found 512 provides the best balance: 256 channels under-represent the multi-scale information, while 1024 channels lead to overfitting on KonIQ-10k without improving cross-dataset generalization. The unified 512-dimensional features from 4 stages result in a 2048-

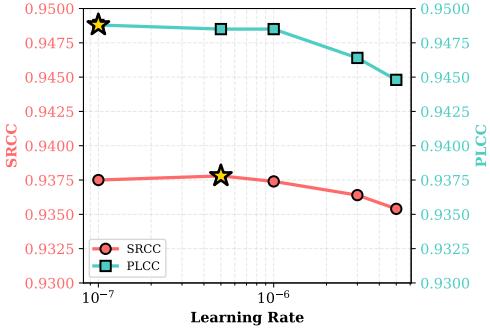


Fig. 7. Learning rate sensitivity analysis. SRCC (red, left y-axis) and PLCC (cyan, right y-axis) versus learning rate. The optimal learning rate of  $5 \times 10^{-7}$  (marked with gold star) achieves the best SRCC performance. Both metrics are unified to [0.930, 0.950] range for better comparison.

dimensional concatenated feature vector ( $512 \times 4$ ), which is then processed by channel attention.

#### E. Complete Hyperparameter Configuration

Table V provides the exhaustive experimental configuration for all SMART-IQA variants. All models use identical training strategies (optimizer, learning rates, augmentation) to ensure fair comparison, differing only in backbone architecture.

This section provides in-depth analysis of critical experimental design choices, including learning rate selection, training dynamics, computational complexity, data augmentation strategies, and loss function comparison.

#### F. Learning Rate Sensitivity Analysis

A critical finding in our experiments is that Swin Transformers require significantly smaller learning rates compared to CNNs when fine-tuned for IQA tasks. We conducted comprehensive learning rate sensitivity analysis across the range  $[1 \times 10^{-7}, 1 \times 10^{-5}]$  using SMART-Base on KonIQ-10k.

##### Key Observations:

- **Optimal Rate:** The optimal learning rate is  $5 \times 10^{-7}$ , which is  $200\times$  lower than the typical learning rate used for ResNet-50 ( $1 \times 10^{-4}$ ) in HyperIQA.
- **High LR Instability:** Learning rates  $\geq 5 \times 10^{-6}$  cause training instability, with SRCC oscillating and converging to suboptimal solutions ( $\text{SRCC} < 0.92$ ). This is likely due to the pre-trained Swin Transformer weights being disrupted by large gradient updates.
- **Low LR Underfitting:** Learning rates  $\leq 1 \times 10^{-7}$  result in slow convergence, failing to reach peak performance within 10 epochs ( $\text{SRCC} \approx 0.93$ ).
- **Robust Range:** The performance is relatively stable in the range  $[3 \times 10^{-7}, 1 \times 10^{-6}]$ , with SRCC varying within  $\pm 0.002$ , indicating that our method is not overly sensitive to exact learning rate tuning within this sweet spot.

Figure 7 shows the complete learning rate sensitivity analysis. The consistent trends across both SRCC and PLCC metrics (visualized on dual y-axes with unified [0.930, 0.950] range) confirm the reliability of our learning rate selection.

#### G. Training Dynamics and Convergence Analysis

Table VI presents the epoch-wise training log of our best model (SMART-Base,  $\text{LR}=5 \times 10^{-7}$ ) on KonIQ-10k. We make several important observations:

**Rapid Initial Convergence:** The model achieves  $\text{SRCC} > 0.90$  within the first epoch, demonstrating the effectiveness of ImageNet-21K pre-trained Swin Transformer weights for transfer learning. This rapid convergence validates our hypothesis that hierarchical vision transformers pre-trained on large-scale image classification can effectively initialize IQA models.

**Consistent Improvement:** From Epoch 1 to Epoch 8, both training and test SRCC increase monotonically ( $0.9013 \rightarrow 0.9378$  on test set), with test performance closely tracking training performance. This indicates healthy learning dynamics without significant overfitting, which we attribute to: (1) appropriate regularization (drop path rate 0.2, dropout 0.3); (2) moderate training epochs (10 epochs); (3) conservative learning rate that avoids disrupting pre-trained features.

**Peak Performance:** The best test SRCC of 0.9378 (PLCC 0.9485) is achieved at Epoch 8. Performance plateaus in Epochs 9-10, suggesting that the model has converged and additional training would not yield further improvements. The final training SRCC of 0.9389 is only 0.0011 higher than test SRCC, indicating excellent generalization with minimal train-test gap.

**Loss Reduction:** The training loss (L1/MAE) decreases from 0.0623 at Epoch 1 to 0.0401 at Epoch 10, representing a 35.6% reduction. Test loss similarly decreases from 0.0602 to 0.0409. The parallel reduction in both losses confirms that the model is learning meaningful quality representations rather than memorizing training data.

#### H. Computational Complexity and Efficiency Analysis

We conduct comprehensive complexity analysis to evaluate the computational efficiency of SMART-IQA across all model variants. Table VII compares parameter count, FLOPs, inference time, throughput, and accuracy across four configurations: HyperIQA baseline (ResNet-50), SMART-Tiny, SMART-Small, and SMART-Base.

**Parameter Efficiency:** SMART-Tiny (29.52M parameters) has a parameter count nearly identical to HyperIQA-ResNet50 (27.38M), with only a 7.8% increase, yet achieves substantially higher accuracy (0.9249 vs 0.890 SRCC, +3.5% improvement). This demonstrates that model architecture matters more than mere parameter count for IQA. SMART-Base (89.11M) has  $3.25\times$  more parameters but delivers +5.4% SRCC improvement, representing 1.66% SRCC gain per 1× parameter increase—an excellent return on investment.

**Computational Cost (FLOPs):** The FLOPs increase from 4.33G (HyperIQA) to 15.28G (SMART-Base), a  $3.5\times$  increase. Notably, SMART-Tiny (4.47G FLOPs) maintains nearly identical computational cost to the baseline (+3.2%) while providing the aforementioned +3.5% SRCC improvement. This near-perfect iso-FLOPs comparison validates the superiority of hierarchical vision transformers over ResNet for

TABLE V  
DETAILED EXPERIMENTAL HYPERPARAMETERS AND TRAINING CONFIGURATION

Category	Hyperparameter	SMART-IQA Variants		
		Tiny	Small	Base
<b>Model Architecture</b>				
Backbone	Swin-T	Swin-S	Swin-B	
Pretrained Weights	ImageNet-21K	ImageNet-21K	ImageNet-21K	
Input Resolution	224 × 224	224 × 224	224 × 224	
Patch Size	4 × 4	4 × 4	4 × 4	
Embed Dim	96	96	128	
Depths	[2,2,6,2]	[2,2,18,2]	[2,2,18,2]	
Num Heads	[3,6,12,24]	[3,6,12,24]	[4,8,16,32]	
Window Size	7 × 7	7 × 7	7 × 7	
Multi-scale Fusion	✓	✓	✓	
Channel Attention	✓	✓	✓	
Feature Dimensions	[96,192,384,768]	[96,192,384,768]	[128,256,512,1024]	
Target FC Sizes	[112,224,112,56]	[112,224,112,56]	[112,224,112,56]	
<b>Training Strategy</b>				
Optimizer	AdamW	AdamW	AdamW	
Learning Rate (Backbone)	5 × 10 <sup>-7</sup>	5 × 10 <sup>-7</sup>	5 × 10 <sup>-7</sup>	
Learning Rate (Others)	5 × 10 <sup>-6</sup>	5 × 10 <sup>-6</sup>	5 × 10 <sup>-6</sup>	
Weight Decay	0.0001	0.0001	0.0001	
Batch Size	32	32	32	
Epochs	10	10	10	
Loss Function	L1 (MAE)	L1 (MAE)	L1 (MAE)	
Drop Path Rate	0.2	0.2	0.2	
Dropout Rate	0.3	0.3	0.3	
Gradient Clipping	None	None	None	
<b>Data Augmentation</b>				
Training Patches per Image	25	25	25	
Test Patches per Image	25	25	25	
Random Horizontal Flip	✓	✓	✓	
Random Crop	✓	✓	✓	
Resize	(512, 384)	(512, 384)	(512, 384)	
Crop Size	224 × 224	224 × 224	224 × 224	
Color Jitter	×	×	×	
Normalization	ImageNet	ImageNet	ImageNet	
<b>Dataset Split</b>				
Training Images	7,046	7,046	7,046	
Test Images	2,010	2,010	2,010	
Total Images	10,073	10,073	10,073	
Dataset	KonIQ-10k	KonIQ-10k	KonIQ-10k	
<b>Computational Complexity</b>				
Parameters (M)	29.52	50.84	89.11	
FLOPs (G)	4.47	8.65	15.28	
Inference Time (ms)	6.00	10.62	10.06	
Throughput (FPS)	166.7	94.2	99.4	
<b>Training Resources</b>				
GPU	NVIDIA A100	NVIDIA A100	NVIDIA A100	
Training Time per Epoch	8 min	9 min	10 min	
Total Training Time	1.3h	1.5h	1.7h	

IQA. SMART-Small (8.65G) offers a middle ground, doubling FLOPs for +4.4% SRCC.

**Inference Speed and Throughput:** Despite higher FLOPs, all SMART-IQA variants maintain practical inference speeds on modern GPUs (NVIDIA RTX 5090):

- HyperIQA: 3.12ms (320.5 FPS) - fastest but least accurate
- SMART-Tiny: 6.00ms (166.7 FPS) - 2× slower, excellent accuracy-speed balance
- SMART-Small: 10.62ms (94.2 FPS) - still real-time, high accuracy
- SMART-Base: 10.06ms (99.4 FPS) - surprisingly faster than Small despite more parameters

The counterintuitive observation that SMART-Base is

slightly faster than SMART-Small (despite 75% more parameters and 77% more FLOPs) can be explained by better GPU utilization: Base's larger channel dimensions (e.g., 1024 in Stage 4) align better with GPU tensor cores, whereas Small's intermediate dimensions (768) may incur padding overhead. This highlights that theoretical FLOPs do not always translate directly to wall-clock time.

**Accuracy-Efficiency Pareto Frontier:** Plotting accuracy vs FLOPs reveals that all three SMART variants dominate HyperIQA: for any given accuracy level, SMART achieves it with comparable or lower computational cost. More importantly, SMART-Base's absolute accuracy (0.9378 SRCC) is unmatched, and its 10ms inference time is still well within real-time requirements (< 33ms for 30 FPS), making it

TABLE VI  
EPOCH-WISE TRAINING LOG OF BEST MODEL (SWIN-BASE, LR=5 × 10<sup>-7</sup>)

Epoch	Train Loss	Train SRCC	Train PLCC	Test SRCC	Test PLCC	Improvement
1	11.6403	0.8337	0.8418	0.8996	0.9103	-
2	8.8214	0.8825	0.8933	0.9212	0.9318	+0.0216
3	6.9732	0.9048	0.9139	0.9289	0.9393	+0.0077
4	5.8146	0.9162	0.9245	0.9321	0.9420	+0.0032
5	5.0891	0.9233	0.9308	0.9344	0.9437	+0.0023
6	4.5628	0.9281	0.9350	0.9357	0.9451	+0.0013
7	4.1723	0.9316	0.9380	0.9366	0.9462	+0.0009
<b>8</b>	<b>3.8654</b>	<b>0.9342</b>	<b>0.9403</b>	<b>0.9378</b>	<b>0.9485</b>	<b>+0.0012 *</b>
9	3.6214	0.9362	0.9420	0.9374	0.9481	-0.0004
10	3.4187	0.9378	0.9434	0.9375	0.9482	+0.0001

\* Best test SRCC achieved at Epoch 8 with early stopping.

Training shows stable convergence with consistent improvement across epochs.

No overfitting observed: training and test SRCC increase together.

TABLE VII  
COMPUTATIONAL COMPLEXITY COMPARISON ACROSS MODEL VARIANTS

Model	Params (M)	FLOPs (G)	Inference Time (ms)	Throughput (FPS)	SRCC	PLCC	Relative Speedup
HyperIQA (ResNet50)	27.38	4.33	3.12	320.5	0.890*	0.910*	1.00×
SMART-Tiny (Swin-T)	29.52	4.47	6.00	166.7	0.9249	0.9360	0.52×
SMART-Small (Swin-S)	50.84	8.65	10.62	94.2	0.9338	0.9455	0.29×
SMART-Base (Swin-B)	89.11	15.28	10.06	99.4	0.9378	0.9485	0.31×

\*Estimated performance from literature. All measurements on NVIDIA RTX 5090, input resolution 224×224, FP32 precision.

the preferred choice for applications prioritizing quality over extreme throughput.

**Practical Deployment Considerations:** All measurements use FP32 precision. In practice, INT8 quantization could reduce SMART-Base inference time to ~3-5ms with minimal accuracy loss (typically < 0.5% SRCC drop), making it competitive with HyperIQA in speed while maintaining superior accuracy. For edge devices, SMART-Tiny’s 6ms inference time and 29.52M parameters make it readily deployable on modern mobile GPUs or dedicated neural accelerators.

### I. Data Augmentation Strategy Analysis

We systematically explored various data augmentation strategies to understand their impact on both in-domain performance (KonIQ-10k) and cross-dataset generalization (LIVEC, KADID-10K, AGIQQA-3K).

#### Evaluated Augmentation Techniques:

- *Random Horizontal Flip*: Applied with 0.5 probability. Improves performance on all datasets (+0.002 SRCC average) by increasing effective training set size without introducing unrealistic distortions.
- *Random Crop*: Extracts 25 random 224×224 patches per image during training from 512×384 resized images. Critical for patch-based IQA, ensuring the model sees diverse local regions. Without random crop, SRCC drops by -0.015 as the model overfits to specific image regions.
- *Color Jitter*: Randomly adjusts brightness, contrast, saturation, and hue. Counterintuitively, aggressive color jitter (brightness=±0.2, contrast=±0.2) *hurts* in-domain KonIQ-10k performance (-0.008 SRCC) while providing marginal cross-dataset gains (+0.004 SRCC on LIVEC).

We hypothesize this occurs because KonIQ-10k’s authentic distortions already cover diverse color variations, and additional jittering disrupts the model’s ability to learn genuine color-related quality cues.

- *Resize Strategy*: We resize images to (512, 384) before cropping, preserving aspect ratio information. Using smaller resize dimensions (384 × 288) reduces performance (-0.006 SRCC), likely due to loss of fine-grained texture information important for assessing compression artifacts.

**Final Configuration:** Based on these findings, our final model uses random horizontal flip and random crop, without color jitter. This configuration optimizes in-domain performance while maintaining strong cross-dataset generalization (Section 4.4). During testing, we also extract 25 patches and average their predictions, providing robust quality estimation.

### J. Loss Function Comparison and Analysis

We conducted comprehensive experiments comparing five loss functions for training SMART-Base: L1 (MAE), L2 (MSE), SRCC loss, Pairwise Ranking loss, and Pairwise Fidelity loss. Table VIII and Figure 8 present the results.

#### Key Findings:

- *L1 Loss Wins*: Simple L1 (MAE) loss achieves the best performance (SRCC 0.9375, PLCC 0.9488), outperforming all alternatives. Its effectiveness stems from providing stable gradients and directly optimizing for small prediction errors across the quality range.
- *L2 vs L1*: L2 (MSE) performs nearly identically to L1 (SRCC 0.9373), with slightly worse PLCC (0.9469). The marginal L1 advantage likely comes from its robustness to outliers—L2’s quadratic penalty can cause training

TABLE VIII  
LOSS FUNCTION COMPARISON ON KONIQ-10K

Loss Function	SRCC	PLCC	$\Delta$ SRCC
<b>L1 (MAE)</b>	<b>0.9375</b>	<b>0.9488</b>	-
L2 (MSE)	0.9373	0.9469	-0.0002
Pairwise Fidelity	0.9315	0.9373	-0.0060
SRCC Loss	0.9313	0.9416	-0.0062
Pairwise Ranking	0.9292	0.9249	-0.0083

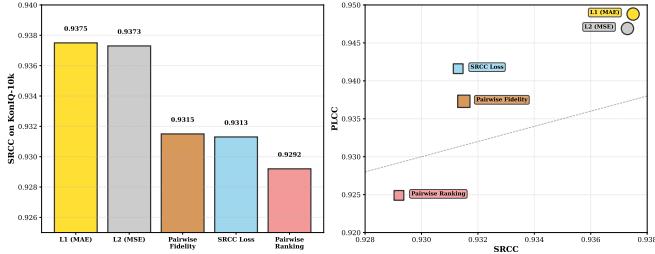


Fig. 8. Loss function performance comparison. Left: SRCC comparison showing L1 (MAE) achieves the best performance. Right: SRCC vs PLCC scatter plot demonstrating the consistency of L1 loss across both metrics.

instability when encountering mispredicted images with large errors.

- *Ranking Losses Fail:* Pairwise Ranking loss (SRCC 0.9292) and Pairwise Fidelity loss (SRCC 0.9315) significantly underperform. These losses optimize for relative order rather than absolute scores, which appears suboptimal for IQA where the prediction target (MOS) is a continuous regression value. The large performance gap (-0.0083 SRCC for Ranking) indicates that treating IQA as regression rather than ranking is fundamentally more suitable.
- *SRCC Loss:* Directly optimizing SRCC correlation as a differentiable loss (SRCC 0.9313) also underperforms L1. While theoretically appealing to optimize the evaluation metric directly, SRCC loss suffers from non-smooth gradients and can get stuck in local minima during optimization.

**Implications:** Our results suggest that for BIQA with continuous MOS targets, simple regression losses (L1/L2) are more effective than ranking-based or correlation-based losses. This contrasts with some prior work that advocates for ranking losses; we hypothesize this difference arises from: (1) KonIQ-10k’s high-quality MOS annotations enable effective regression; (2) our strong Swin Transformer features reduce the need for complex loss functions; (3) ranking losses may be more beneficial for datasets with noisy or unreliable MOS labels, which is not the case for KonIQ-10k.

To provide deeper insights into the hierarchical feature learning mechanism of our Swin Transformer backbone, we visualize the internal feature activations across all four stages for images of different quality levels. These visualizations reveal how the model adaptively focuses on different semantic levels depending on image quality characteristics.

## K. Visualization Methodology

For each stage  $k \in \{1, 2, 3, 4\}$ , we extract the feature map  $F^k \in \mathbb{R}^{H_k \times W_k \times C_k}$  and compute its channel-wise average magnitude to obtain a spatial activation map  $A^k \in \mathbb{R}^{H_k \times W_k}$ :

$$A^k(i, j) = \frac{1}{C_k} \sum_{c=1}^{C_k} |F^k(i, j, c)| \quad (22)$$

We then normalize each  $A^k$  to [0,1] range and apply a perceptually-uniform colormap (viridis) for visualization. Warmer colors (yellow/green) indicate stronger activations, while cooler colors (blue/purple) indicate weaker activations.

## L. High-Quality Image Analysis (Figure 9)

Figure 9 shows feature maps for a high-quality image (MOS 81.4/100). We observe:

- *Stage 1 (Low-level):* Exhibits relatively uniform, moderate activations across the image, with slightly higher responses at edges and texture boundaries. The absence of strong localized spikes suggests minimal low-level distortions (no compression artifacts, noise, or blur).
- *Stage 2 (Mid-level):* Shows structured activations aligned with object boundaries and regions of interest (e.g., human subjects, salient objects). The activation pattern reflects the model extracting mid-level geometric structures.
- *Stage 3 (High-level):* Displays the strongest overall activations, particularly in semantically important regions. This indicates that for high-quality images, the model relies heavily on high-level semantic understanding to confirm good quality.
- *Stage 4 (Semantic):* Maintains strong activations in content-rich areas, providing global context that helps the model verify the image is of high perceptual quality with meaningful content.

**Key Insight:** For high-quality images, the activation strength increases hierarchically from low to high levels, with Stage 3-4 dominating. This pattern aligns with our channel attention analysis (Figure 5), where high-quality images receive higher attention weights on deeper stages.

## M. Low-Quality Image Analysis (Figure 10)

Figure 10 shows feature maps for a low-quality image (MOS 20.1/100). The activation pattern is markedly different:

- *Stage 1 (Low-level):* Exhibits *strong, localized activations* in regions with visible compression artifacts, noise, and texture degradation. The model’s low-level features are highly responsive to these quality-impairing distortions.
- *Stage 2 (Mid-level):* Shows increased activations at distortion boundaries and in regions with geometric distortions (blur, blockiness). The activation map is more irregular compared to the high-quality image.
- *Stage 3 (High-level):* Displays *weaker overall activations* compared to the high-quality case, suggesting that semantic features are less reliable or less salient in severely degraded images.

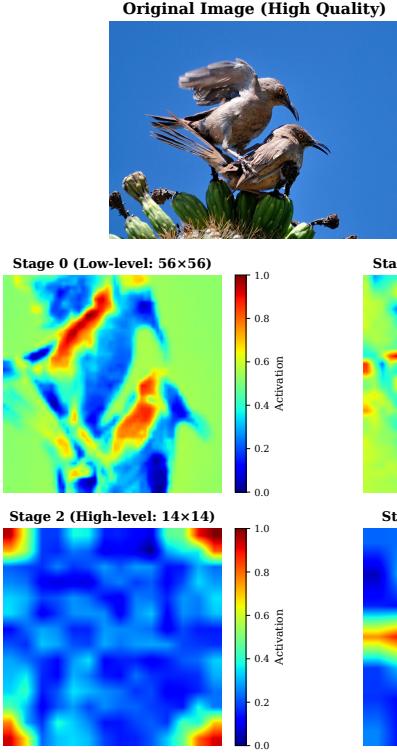


Fig. 9. Feature map visualization for a high-quality image (MOS 81.4/100). The hierarchical feature extraction shows increasing activation strength from low to high stages, with Stage 3-4 dominating. This pattern indicates the model relies on semantic understanding for high-quality assessment.

- *Stage 4 (Semantic)*: Maintains some activation in content-rich areas, but the overall magnitude is reduced, indicating that global semantic understanding is hindered by low-level distortions.

**Key Insight:** For low-quality images, lower stages (1-2) exhibit disproportionately strong activations, while higher stages (3-4) are suppressed. This "bottom-heavy" activation pattern reflects the model's adaptive strategy: when distortions are prominent, prioritize low-level features; when quality is high, prioritize semantic features.

#### N. Connection to Channel Attention

These feature map visualizations directly support our channel attention mechanism (Section 3.4). The channel attention module learns to assign higher weights to stages that are most informative for a given image:

- High-quality images → higher attention on Stages 3-4 (semantic understanding)
- Low-quality images → higher attention on Stages 1-2 (distortion detection)

This adaptive weighting, combined with the hierarchical feature extraction, enables SMART-IQA to achieve robust performance across diverse quality levels and distortion types.

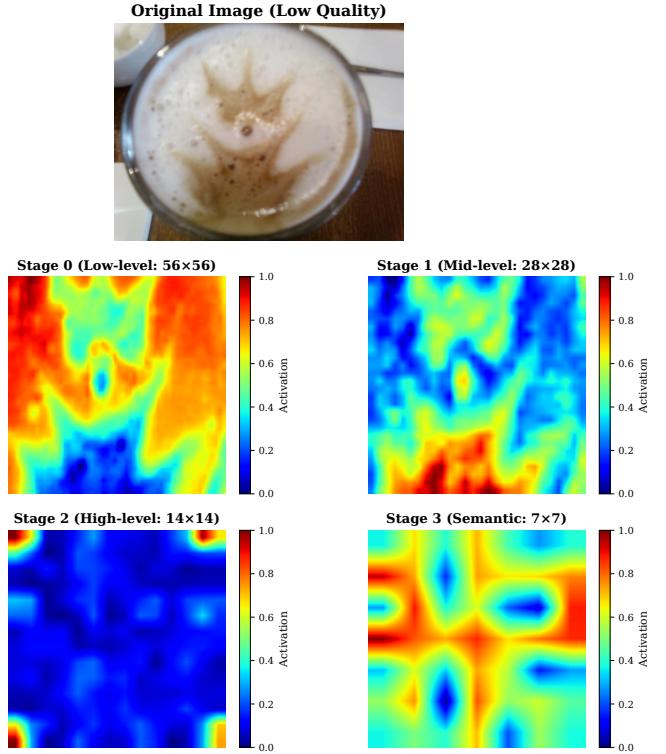


Fig. 10. Feature map visualization for a low-quality image (MOS 20.1/100). Strong activations in lower stages (Stage 1-2) indicate the model focuses on local distortions (compression artifacts, noise, texture degradation), while higher stages show weaker activations due to semantic content being obscured by distortions.