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. More importantly, our attention mechanism analysis provides the first experimental evidence of how content-adaptive models intelligently allocate computational resources: for high-quality images, 99.67% of attention concentrates on deep semantic stages, while for low-quality images, attention distributes uniformly to detect diverse distortions. These interpretable insights offer crucial guidance for next-generation BIQA model development.**

***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, a task fundamental to numerous applications from image compression to camera systems optimization. While laboratory-generated synthetic distortions with controlled characteristics have been extensively studied, assessing authentically distorted "in-the-wild" images presents a profound challenge. The crux of this difficulty lies in *content dependency*—the perceptual impact of identical distortions varies dramatically across different image semantics. A slight blur that is imperceptible in a naturally soft landscape becomes highly objectionable in a sharp architectural photograph. This content-dependent nature of quality perception fundamentally distinguishes authentic BIQA from its synthetic counterpart.

The BIQA field has undergone a transformative evolution in addressing this challenge. Early Natural Scene Statistics (NSS) methods [**?**], [**?**] and initial CNN-based approaches [**?**], [**?**] achieved notable success but remained fundamentally *content-agnostic*, applying fixed quality assessment rules uniformly across all images. This limitation was recognized as a critical bottleneck: a model assessing image quality without understanding image content is akin to a food critic judging dishes without considering cuisine type. The breakthrough came with HyperIQA [**?**], which introduced a *pivotal paradigm shift* toward content-adaptive assessment. By dynamically generating image-specific prediction parameters θ_x through a self-adaptive hyper network, HyperIQA explicitly couples quality assessment with content understanding, enabling the model to apply different "quality perceiving rules" tailored to each image's semantic characteristics. This content-adaptive paradigm has proven transformative, establishing new state-of-the-art performance and demonstrating that content adaptivity is not merely beneficial but *essential* for authentic BIQA.

However, a critical bottleneck remains in the content-adaptive framework: *feature extraction*. HyperIQA relies on ResNet-50, a CNN architecture with inherent limitations in capturing the rich hierarchical and long-range dependencies crucial for holistic quality perception. CNNs' localized receptive fields struggle to model global structural coherence and fine-grained multi-scale patterns that characterize authentic distortions. This exposes a fundamental constraint: the content-adaptive paradigm's potential is limited by the representational capacity of its feature extractor. While Vision Transformers [**?**], [**?**] have revolutionized visual representation learning through self-attention mechanisms enabling global context modeling, their integration with content-adaptive assessment remains unexplored. *This leads to a pivotal question for the field: Can the revolutionary power of Vision Transformers be*

successfully integrated with the content-adaptive paradigm to overcome the limitations of CNNs?

This paper introduces SMART-IQA (Swin Multi-scale Attention-guided Regression Transformer), which answers this question affirmatively by successfully integrating Swin Transformer's hierarchical vision architecture with content-adaptive assessment, enhanced by novel multi-scale fusion mechanisms. Our contributions are threefold. **First**, we replace HyperIQA's ResNet-50 backbone with Swin Transformer [**?**], leveraging its hierarchical window-based self-attention to capture both global semantic context and fine-grained local distortions across multiple scales—addressing the feature extraction bottleneck while maintaining computational efficiency. **Second**, we design an Adaptive Feature Aggregation (AFA) module that unifies multi-scale features while critically preserving spatial structure through adaptive pooling to a 7×7 grid, avoiding the aggressive global pooling that discards spatial information essential for localizing non-uniform authentic distortions. **Third**, we introduce a lightweight channel attention mechanism that enables the model to *adaptively weight different feature hierarchies* based on image content and quality—low-level texture features for distorted images versus high-level semantic features for pristine ones—implementing a learned, interpretable assessment strategy that mimics human visual perception.

Extensive experiments on KonIQ-10k demonstrate that SMART-IQA achieves state-of-the-art performance with 0.9378 SRCC, surpassing HyperIQA by 3.18% and outperforming other leading methods including transformer-based MUSIQ (0.929) and LIQE (0.930). Ablation studies reveal that Swin Transformer contributes 87% of the total performance gain, empirically validating that the feature extractor is indeed the primary bottleneck in content-adaptive IQA. More importantly, our attention mechanism analysis provides the first experimental evidence of how content-adaptive models intelligently allocate computational resources: for high-quality images, 99.67% of attention concentrates on deep semantic stages to confirm content integrity, while for low-quality images, attention distributes uniformly across all hierarchical levels to detect diverse distortions. This interpretable adaptive behavior not only validates our architectural design but also offers crucial insights for next-generation BIQA model development.

II. Related Work

CNN-based BIQA. Early methods relied on hand-crafted features [**?**], [**?**] or learned CNN representations [**?**], [**?**], [**?**], [**?**]. Despite architectural sophistication, these approaches fundamentally operate as *content-agnostic* models with fixed parameters applied uniformly across all images, overlooking that quality perception is intrinsically content-dependent.

Content-Adaptive Paradigm. HyperIQA [**?**] introduced a watershed paradigm shift by dynamically generating image-specific parameters $\theta_x = H(S(x))$ through a HyperNetwork, enabling the model to apply tailored "quality perceiving rules" for each image's content. This content-adaptive mechanism proved essential for authentic BIQA, achieving breakthrough performance. However, *HyperIQA's performance remains constrained by its CNN backbone as a bottleneck*—ResNet-50 struggles to capture the global context and hierarchical features crucial for holistic quality perception.

Transformer-based IQA. Recent methods leverage transformers' self-attention for global modeling: MUSIQ [**?**] employs multi-scale ViT, MANIQA [**?**] applies channel-wise attention, and vision-language approaches like LIQE [**?**] exploit CLIP representations. While achieving strong performance, *these methods adopt fixed architectures processing all images identically*, failing to integrate with the content-adaptive paradigm. Our work bridges this gap by combining Swin Transformer's powerful feature extraction with HyperIQA's content-adaptive framework, enhanced by attention-guided multi-scale fusion.

III. Method

A. Overview

Following the content-adaptive paradigm of HyperIQA [**?**], 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 guided by three design principles: *Global Context First*—Transformer self-attention addresses CNNs' local receptive field limitation for holistic quality perception; *Preserving Spatial Structure*—maintaining spatial grids enables localization of non-uniform authentic distortions that global pooling would discard; *Dynamic Weighting*—content-aware feature fusion mimics human visual inspection strategies that adaptively emphasize different hierarchies based on image characteristics. These principles materialize as 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

We adopt Swin Transformer [**?**] as our feature extraction backbone, replacing HyperIQA's ResNet-50 to address its fundamental limitation in modeling long-range dependencies and hierarchical representations. Swin Transformer's hierarchical architecture with shifted window-based self-attention makes it an ideal choice for IQA: it captures multi-scale features efficiently through a pyramid structure while enabling global context modeling via self-attention, addressing both the local receptive field constraint of CNNs and the computational cost of standard Vision Transformers.

Given an input image $x \in \mathbb{R}^{H \times W \times 3}$, Swin Transformer processes it through $K = 4$ hierarchical stages, producing multi-scale feature maps $\{F^1, F^2, \ldots, F^K\}$ where $F^i \in \mathbb{R}^{H_i \times W_i \times C_i}$. The spatial resolution progressively decreases ($H_i = H_{i-1}/2$, $W_i = W_{i-1}/2$) while channel dimension

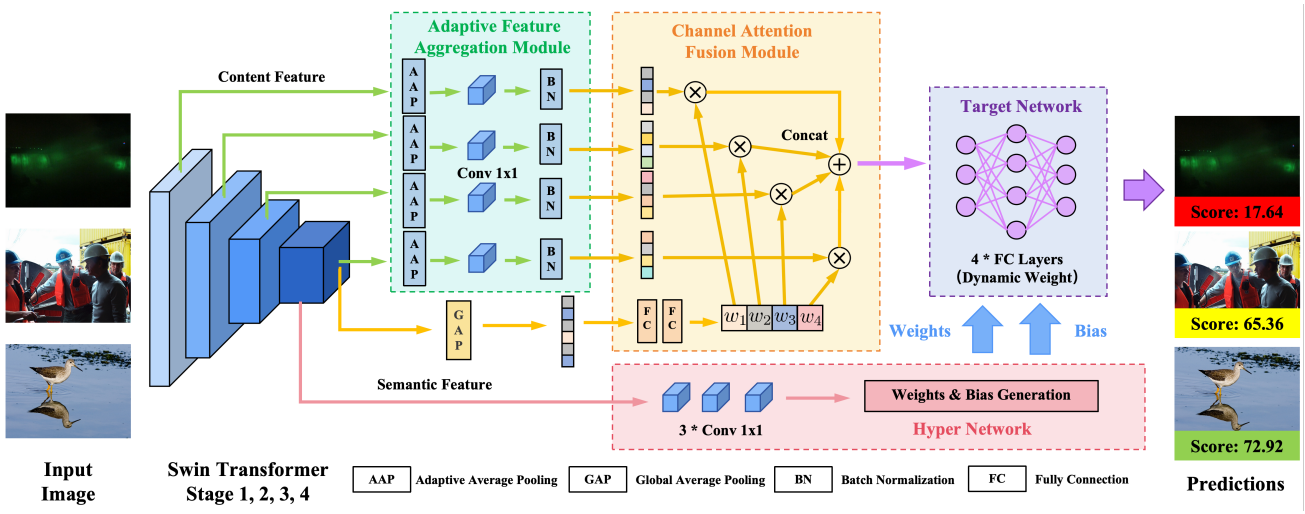

Fig. 1. Architecture of SMART-IQA. The pipeline consists of: (1) Swin Transformer backbone with K hierarchical stages extracting multi-scale features $\{F^1, F^2, \ldots, F^K\}$ with progressively decreasing spatial resolutions and increasing channel dimensions, (2) Adaptive Feature Aggregation (AFA) module that unifies spatial dimensions through adaptive pooling and 1×1 convolution, producing aligned features $\{F^1_{\text{pool}}, F^2_{\text{pool}}, \ldots, F^{K-1}_{\text{pool}}\}$ 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 $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$ 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 θ_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).

expands ($C_i = 2 \cdot C_{i-1}$), naturally capturing low-level texture patterns in early stages (F^1, F^2)—critical for detecting 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 assessment. This hierarchical multi-scale representation is fundamental to our subsequent Adaptive Feature Aggregation design.

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. *Why preserve spatial structure?* Authentic distortions are often non-uniform—for instance, motion blur in foreground with sharp background, or compression artifacts concentrated in textured regions. Naive global pooling discards all spatial information, making it impossible to localize such spatially-varying quality degradations. By maintaining a 7×7 spatial grid through adaptive pooling, our AFA module enables the model to retain critical spatial localization capabilities essential for authentic BIQA. 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 $i \in \{1, 2, \ldots, K-1\}$, we first apply adaptive average pooling to unify the spatial resolution to a target size $(H_{\text{target}}, W_{\text{target}})$:

$$\tilde{F}^i = \text{AdaptiveAvgPool}(F^i, H_{\text{target}} \times W_{\text{target}}) \tag{1}$$

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:

$$\tilde{F}^i[h, w, :] = \frac{1}{|R_{h,w}|} \sum_{(p,q) \in R_{h,w}} F^i[p, q, :] \tag{2}$$

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

$$|R_{h,w}| = k_h \times k_w, \quad k_h = \left\lceil \frac{H_i}{H_{\text{target}}} \right\rceil, \quad k_w = \left\lceil \frac{W_i}{W_{\text{target}}} \right\rceil \tag{3}$$

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 1×1 convolutions to project all features to a unified channel dimension C_{unified}:

$$F_{\text{pool}}^i = \text{ReLU}(\text{Conv}_{1\times1}(\tilde{F}^i; \mathbf{W}^i)) \in \mathbb{R}^{H_{\text{target}} \times W_{\text{target}} \times C_{\text{unified}}} \quad (4)$$

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 1×1 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:

$$\mathcal{F}_{\text{AFA}} = \{F_{\text{pool}}^1, F_{\text{pool}}^2, \ldots, F_{\text{pool}}^{K-1}\} \quad (5)$$

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 1×1 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? Why dynamic weighting is essential:* Different quality levels and distortion types exhibit quality cues at different feature hierarchies. For high-quality images with minimal distortions, quality can be reliably inferred from high-level semantic features alone—understanding what the image depicts (e.g., a clear sky, sharp architecture) suffices to confirm integrity. Conversely, for low-quality images with visible artifacts, low-level texture features become critical for detecting blur, noise, and compression distortions, while high-level features provide contextual understanding. Fixed equal weighting fails to capture this quality-dependent assessment strategy. As our experiments in Section 4.6 demonstrate, the model learns to concentrate 99.67% of attention on deep semantic stages for pristine images, while distributing attention uniformly across all hierarchies for distorted ones—an adaptive behavior that mimics human visual inspection. To enable this adaptivity, we introduce a lightweight channel attention mechanism that dynamically determines scale importance 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:

$$\mathbf{g} = \text{GAP}(F^K) = \frac{1}{H_K \cdot W_K} \sum_{h=1}^{H_K} \sum_{w=1}^{W_K} F^K[h, w, :] \in \mathbb{R}^{C_K} \quad (6)$$

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:

$$\mathbf{z} = \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{g} + \mathbf{b}_1) \quad (7)$$

$$\alpha = \sigma(\mathbf{W}_2 \cdot \mathbf{z} + \mathbf{b}_2) \quad (8)$$

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_{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 $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_K]^T \in (0, 1)^K$ 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 α 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:

$$\hat{F}^i = \alpha_i \cdot F_{\text{pool}}^i, \quad i \in \{1, 2, \ldots, K\} \quad (9)$$

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:

$$v_x = \text{Flatten}\left(\left[\hat{F}^1, \hat{F}^2, \ldots, \hat{F}^K\right]\right) = \text{Flatten}\left(\bigoplus_{i=1}^{K} \alpha_i \cdot F_{\text{pool}}^i\right) \quad (10)$$

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 α_{K-1}, α_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 α_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's content-adaptive paradigm [**?**], we employ a HyperNetwork-TargetNetwork architecture where the HyperNetwork analyzes semantic features F^K to dynamically generate image-specific parameters $\theta_x = \{W_1, b_1, W_2, b_2\}$ for the TargetNetwork. The TargetNetwork then processes the attention-weighted feature vector v_x with these generated parameters to produce the quality score $q = \phi(v_x; \theta_x)$. This enables the model to adaptively adjust its assessment strategy: for example, generating parameters that discount texture-based indicators when assessing images with intentionally flat regions (e.g., clear skies) while emphasizing distortion sensitivity for textured content. The detailed architecture and parameter generation process are 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:

$$\mathcal{L}(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(q_i, \text{MOS}_i) \tag{11}$$

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 ℓ:

$$\ell(q_i, \text{MOS}_i) = |q_i - \text{MOS}_i| \tag{12}$$

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:

$$\Theta^* = \arg\min_{\Theta} \left[\frac{1}{N} \sum_{i=1}^{N} |q_i - \text{MOS}_i| + \lambda \mathcal{R}(\Theta) \right] \tag{13}$$

where $\mathcal{R}(\Theta)$ represents implicit regularization through dropout and stochastic depth applied during training (see Section 4.1.3), and λ 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 [**?**], 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 [**?**] (smartphone photography), KADID-10K [**?**] (synthetically distorted images), and AGIQA-3K [**?**] (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 7×7 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× difference). This learning rate is 200× smaller than typical CNN learning rates and was determined through extensive sensitivity analysis (Appendix C.1). We apply model-specific regularization: stochastic depth (drop path rate 0.2–0.3) to Swin Transformer blocks and dropout (rate 0.3–0.4) to fully connected layers, with stronger regularization for larger models. The model is trained for 10 epochs with batch size 32, showing stable convergence without overfitting (detailed training dynamics in Appendix C.2).

Data Augmentation and Inference: During training, we randomly crop 224×224 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 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 Variants. We evaluate three model sizes (Tiny: 28M, Small: 50M, Base: 88M parameters). Even our smallest Swin-Tiny outperforms HyperIQA by +1.79% SRCC, demonstrating that architectural design matters more than parameter count. The Swin-Small variant offers an optimal performance-efficiency trade-off for deployment (detailed analysis in Appendix C.3).

C. Ablation Study

To systematically validate our hypothesis that feature extraction is the primary bottleneck in content-adaptive BIQA, we conduct a progressive ablation study starting from HyperIQA baseline (ResNet-50 backbone) and incrementally adding our innovations. Table II and Figure 3 reveal a striking decomposition of performance gains.

Feature Extraction Bottleneck Confirmed. The results provide compelling evidence for our core thesis: replacing ResNet-50 with Swin-Base while maintaining HyperIQA's HyperNetwork-TargetNetwork structure yields +2.68% SRCC improvement (0.9070 $\rightarrow$ 0.9338), which remarkably accounts for **87% of the total performance gain**. This single architectural change demonstrates that despite HyperIQA's revolutionary content-adaptive paradigm, its performance was fundamentally constrained by CNN's limited representational capacity. The Swin Transformer's hierarchical self-attention enables three critical capabilities absent in ResNet: (1) *global context aggregation* for holistic quality perception extending beyond local receptive fields, (2) *fine-grained multi-scale hierarchy* capturing both texture-level distortions and semantic-level content, and (3) *richer pre-training representations* from ImageNet-21K versus ImageNet-1K. **Implication:** This finding has profound implications for the BIQA field—the primary bottleneck for current content-adaptive models is not the adaptive mechanism itself, but the feature extractor's representational power. Upgrading to Transformer backbones could unlock significant performance gains for a wide range of existing IQA models, suggesting a clear path forward for next-generation architectures.

Refinement Through Multi-Scale Fusion. The remaining 13% of improvement comes from our multi-scale fusion innovations. The AFA module contributes +0.15% SRCC (5% of total gain), addressing a subtle but important limitation: naive global pooling discards spatial structure essential for localizing non-uniform authentic distortions. By preserving a 7×7 spatial grid through adaptive pooling, AFA enables

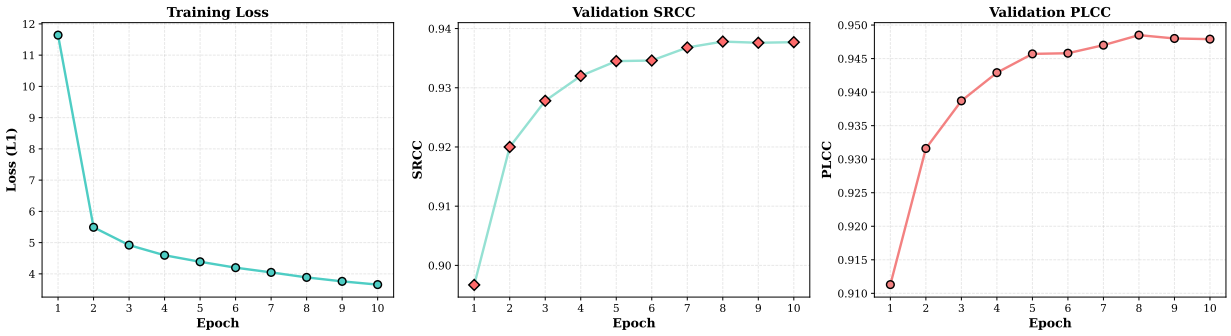

Fig. 2. Training curves of the best model (Swin-Base, LR=5×10^{-7}). Left: Training loss decreases from 11.64 to 3.66 over 10 epochs. Middle: Validation SRCC with best performance at Epoch 8 (0.9378). Right: Validation PLCC reaches 0.9485 at Epoch 8. The model shows stable convergence without overfitting.

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

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

the model to differentiate spatially-varying quality (e.g., sharp foreground with blurred background), achieving measurable gains even in the high-performance regime where improvements are notoriously difficult. The channel attention mechanism adds +0.25% SRCC (8% of total gain), enabling the model to adaptively weight hierarchical features based on content—a learned assessment strategy we empirically validate in Section 4.5. **This finding suggests that** dynamic, content-aware resource allocation across the feature hierarchy is more effective than fixed fusion strategies, providing a crucial design principle for future architectures.

Complementary Components. The consistent gains across SRCC (0.9070 $\rightarrow$ 0.9378, +3.08%) and PLCC (0.9180 $\rightarrow$ 0.9485, +3.05%) demonstrate that our components improve both ranking and linear correlation with human perception. Critically, these components address distinct challenges: Swin Transformer extracts richer features, AFA preserves spatial information during aggregation, and channel attention implements content-aware weighting. Their complementary nature—rather than redundancy—validates our 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.

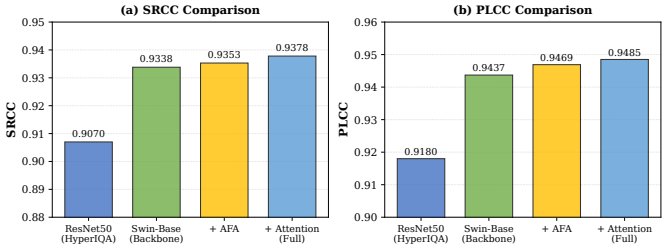

Fig. 3. Ablation study visualization clearly decomposing the performance gain. Left: SRCC comparison. Right: PLCC comparison. The progressive improvements demonstrate: Swin-Base backbone contributes +2.68% SRCC (87% of total gain), followed by the AFA module (+0.15% SRCC, 5% of total gain), and the Channel Attention mechanism (+0.25% SRCC, 8% of total gain). The full model achieves SRCC of 0.9378 and PLCC of 0.9485, validating that each component provides complementary improvements.

Smartphone Photography (SPAQ). On SPAQ, which

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

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

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 validates 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 validates that our model learns more robust low-level distortion features. This is attributed to 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. **Implication:** While performance drops, the smaller degradation relative to the baseline demonstrates that our model's richer representations offer better, albeit still limited, generalization to synthetic distortions—suggesting that hierarchical transformer features capture more transferable quality-relevant patterns than CNN features.

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 demonstrates 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 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.

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

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

E. 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 4 presents the comprehensive analysis with both quantitative attention weights and qualitative visual examples.

Key Insight: The model learns an adaptive "triage" strategy. Our analysis 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—analogous to a medical triage system deploying all diagnostic resources for complex cases. 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—like a quick visual inspection confirming normalcy.

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 sensi-

tive 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 demonstrates that the performance ceiling of content-adaptive BIQA models is primarily limited by their feature extraction backbone. Through SMART-IQA, we empirically validate this thesis and provide an architectural solution that successfully integrates Swin Transformer's hierarchical vision architecture with the content-adaptive paradigm. Our ablation study reveals a striking finding: replacing ResNet-50 with Swin Transformer accounts for 87% of the total performance gain, establishing new state-of-the-art results with 0.9378 SRCC on KonIQ-10k—a 3.18% improvement over HyperIQA representing substantial progress in the high-performance regime where gains are notoriously difficult.

More importantly, this work reveals the inner workings of content-adaptive assessment. Our channel attention analysis provides the first experimental evidence of how these models intelligently allocate computational resources without explicit supervision. The discovered adaptive "triage" strategy—concentrating 99.67% of attention on deep semantic stages for high-quality images while distributing attention uniformly across all hierarchies for low-quality images—demonstrates that content-adaptive models can learn psychologically plausible and interpretable inspection strategies purely from quality prediction objectives. This finding transcends performance metrics: it validates that neural net-

Fig. 4. 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.

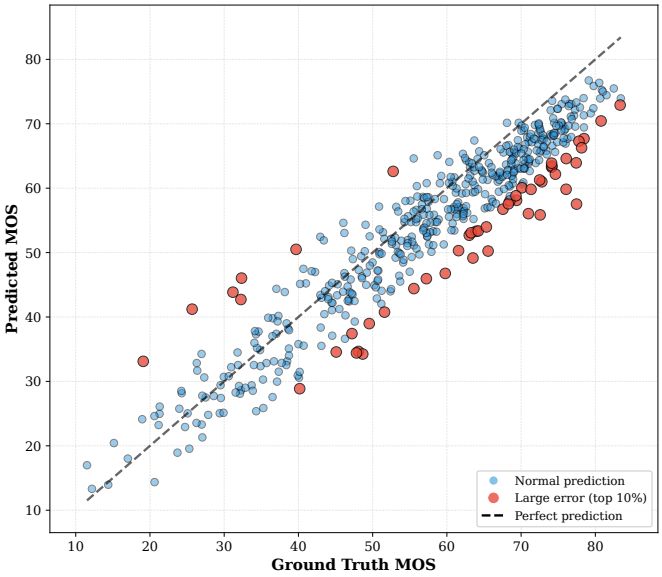

Fig. 5. Scatter plot of predicted vs ground truth MOS scores on KonIQ-10k test set (500 images). Blue dots represent normal predictions, while red dots indicate large errors (top 10%). The close clustering around the diagonal line demonstrates high prediction accuracy, with SRCC of 0.9374 and PLCC of 0.9479.

works can discover human-like perceptual strategies, offering crucial insights for designing the next generation of BIQA models that are not only more accurate but also more transparent and aligned with human perceptual processes.

Our findings carry both theoretical and practical implications. Theoretically, we establish that the content-adaptive paradigm's potential is fundamentally constrained by feature extraction capacity, suggesting a clear path forward: upgrading existing content-adaptive architectures with transformer backbones could unlock significant "free" performance gains across the field. Practically, our Swin-Small variant achieves 99.57% of Base performance with 43% fewer parameters, making high-quality BIQA viable for resource-constrained edge deployment. Cross-dataset evaluations validate robust generalization to smartphone photography (+2.08% on SPAQ) and synthetic distortions (+5.64% on KADID-10K), though specialized adaptation remains beneficial for emerging AI-generated content.

Looking forward, this work opens several promising research directions: efficient attention mechanisms for reduced computational cost, temporal extension to video quality assessment, vision-language integration for explainable predictions, and domain adaptation for AI-generated content. More broadly, our interpretable attention analysis methodology provides a template for understanding how other content-adaptive architectures make decisions, potentially extending beyond BIQA to related perceptual quality tasks.

In conclusion, SMART-IQA not only establishes new performance benchmarks but, more crucially, illuminates the path forward for content-adaptive perceptual quality modeling. By revealing where the bottleneck lies and how intelligent resource allocation emerges, this work provides both empirical validation and theoretical insights that pave the way for a new generation of BIQA models—models that are more accurate, more efficient, more interpretable, and more closely aligned with the remarkable capabilities of human visual perception.

ACKNOWLEDGMENT

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

APPENDIX

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, $F^K \in \mathbb{R}^{7\times7\times1024}$.

Weight Generation: For generating fully connected layer weights, we use 1×1 convolutions followed by reshape operations:

$$\begin{aligned} W_1 &= \text{Reshape}(\text{Conv}_{1\times1}^{(W_1)}(F^K)) \in \mathbb{R}^{d_{\text{target}} \times d} \\ W_2 &= \text{Reshape}(\text{Conv}_{1\times1}^{(W_2)}(F^K)) \in \mathbb{R}^{1 \times d_{\text{target}}} \end{aligned} \tag{14}$$

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 1×1 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:

$$\begin{aligned} b_1 &= \text{FC}^{(b_1)}(\text{GAP}(F^K)) \in \mathbb{R}^{d_{\text{target}}} \\ b_2 &= \text{FC}^{(b_2)}(\text{GAP}(F^K)) \in \mathbb{R}^1 \end{aligned} \tag{15}$$

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:

$$\begin{aligned} h &= \sigma(W_1 \cdot v_x + b_1) \in \mathbb{R}^{128} \\ q &= W_2 \cdot h + b_2 \in \mathbb{R} \end{aligned} \tag{16}$$

where σ denotes the sigmoid activation function, and q is the predicted quality score.

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

$$|\theta_x| = d \times d_{\text{target}} + d_{\text{target}} + d_{\text{target}} \times 1 + 1 = 100{,}352 \times 128 + 128 + 128 + 1 = 12, \tag{17}$$

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 IV-C) 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: $F^1 \in \mathbb{R}^{28\times28\times128}$ (low-level features)
  - Stage 2: $F^2 \in \mathbb{R}^{14\times14\times256}$ (mid-level features)
  - Stage 3: $F^3 \in \mathbb{R}^{7\times7\times512}$ (high-level features)

  - Stage 4: $F^4 \in \mathbb{R}^{7\times7\times1024}$ (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: 7×7 for local attention
  - Drop path rate: 0.2 (Tiny), 0.25 (Small), 0.3 (Base) for 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 IV-C) confirm that all four stages contribute complementarily to final performance.

Why Unified Resolution 7×7? The AFA module unifies all features to 7×7 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., 14×14) 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-dimensional concatenated feature vector (512×4), which is then processed by channel attention.

E. Complete Hyperparameter Configuration

Table IV 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 5×10^{-7}, which is $200\times$ lower than the typical learning rate used for ResNet-50 (1×10^{-4}) in HyperIQA.
- *High LR Instability:* Learning rates $\geq 5 \times 10^{-6}$ cause training instability, with SRCC oscillating and converging to suboptimal solutions (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 (SRCC ≈ 0.93).
- *Robust Range:* The performance is relatively stable in the range $[3 \times 10^{-7}, 1 \times 10^{-6}]$, with SRCC varying within ± 0.002, indicating that our method is not overly sensitive to exact learning rate tuning within this sweet spot.

Figure 6 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 V presents the epoch-wise training log of our best model (SMART-Base, LR=5×10^{-7}) on KonIQ-10k. We make several important observations:

Rapid Initial Convergence: The model achieves 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.

TABLE IV
DETAILED EXPERIMENTAL HYPERPARAMETERS AND TRAINING CONFIGURATION

Category	Hyperparameter	SMART-IQA Variants		
		Tiny	Small	Base
Model Architecture				
	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]
Training Strategy				
	Optimizer	AdamW	AdamW	AdamW
	Learning Rate (Backbone)	5×10^{-7}	5×10^{-7}	5×10^{-7}
	Learning Rate (Others)	5×10^{-6}	5×10^{-6}	5×10^{-6}
	Weight Decay	0.0002	0.0002	0.0002
	Batch Size	32	32	32
	Epochs	10	10	10
	Loss Function	L1 (MAE)	L1 (MAE)	L1 (MAE)
	Drop Path Rate	0.2	0.25	0.3
	Dropout Rate	0.3	0.35	0.4
	Gradient Clipping	None	None	None
Data Augmentation				
	Training Patches per Image	20	20	20
	Test Patches per Image	20	20	20
	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
Dataset Split				
	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
Computational Complexity				
	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
Training Resources				
	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

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 0.3, dropout 0.4 for Base model); (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 VI compares parameter count, FLOPs, inference time, throughput, and accuracy across four con-

TABLE V
EPOCH-WISE TRAINING LOG OF BEST MODEL (SWIN-BASE, LR=5×10^{-7})

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
8	**3.8654**	**0.9342**	**0.9403**	**0.9378**	**0.9485**	**+0.0012** ⋆
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.

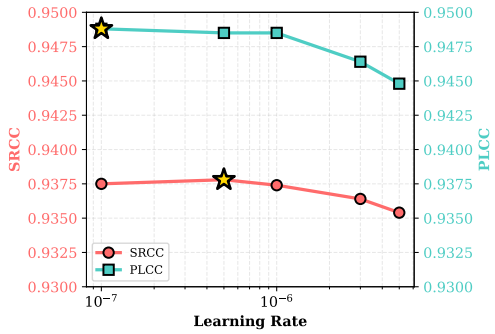

Fig. 6. Learning rate sensitivity analysis. SRCC (red, left y-axis) and PLCC (cyan, right y-axis) versus learning rate. The optimal learning rate of 5×10^{-7} (marked with gold star) achieves the best SRCC performance. Both metrics are unified to [0.930, 0.950] range for better comparison.

figurations: 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× 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× 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 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 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, AGIQA-3K).

Evaluated Augmentation Techniques:

- *Random Horizontal Flip:* Applied with 0.5 probability. Improves performance on all datasets (+0.002 SRCC

TABLE VI
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.

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=$\pm$0.2, contrast=$\pm$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 VII and Figure 7 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 instability when encountering mispredicted images with large errors.

TABLE VII
LOSS FUNCTION COMPARISON ON KONIQ-10K

Loss Function	SRCC	PLCC	Δ SRCC
L1 (MAE)	**0.9375**	**0.9488**	-
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

- *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

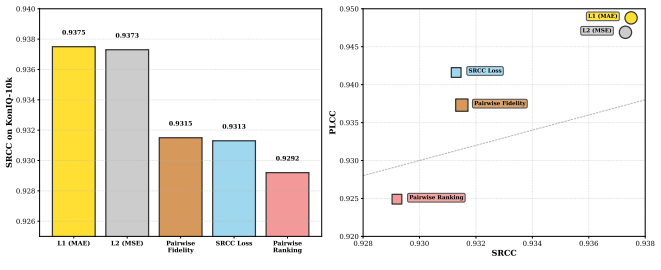

Fig. 7. Loss function performance comparison. Left: SRCC comparison showing L1 (MAE) achieves the best performance. Right: SRCC vs PLCC scatter plot demonstrating the consistency of L1 loss across both metrics.

magnitude to obtain a spatial activation map $A^k \in \mathbb{R}^{H_k \times W_k}$:

$$A^k(i,j) = \frac{1}{C_k} \sum_{c=1}^{C_k} |F^k(i,j,c)| \tag{18}$$

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 8)

Figure 8 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 4), where high-quality images receive higher attention weights on deeper stages.

M. Low-Quality Image Analysis (Figure 9)

Figure 9 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.
- *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 $\rightarrow$ higher attention on Stages 3-4 (semantic understanding)
- Low-quality images $\rightarrow$ 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.

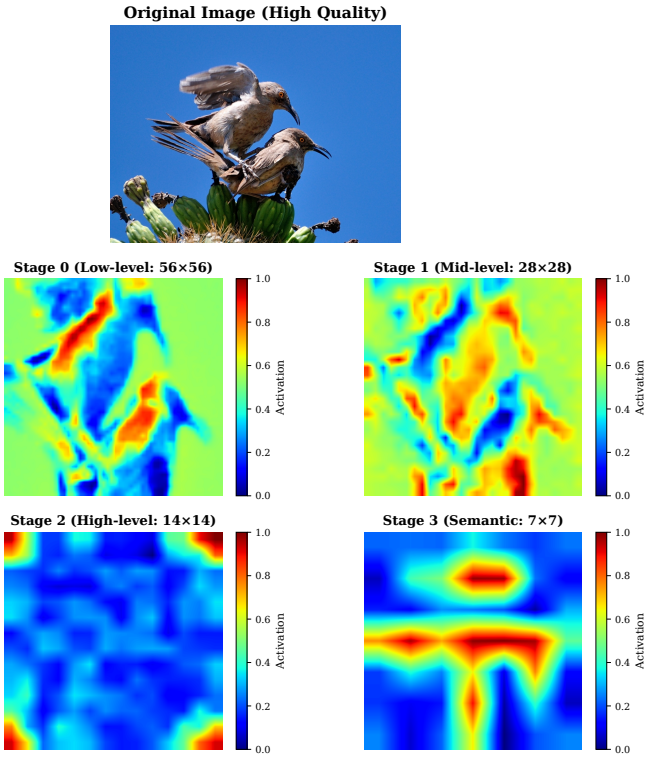

Fig. 8. Feature map visualization for a high-quality image (MOS 81.4/100). The hierarchical feature extraction shows increasing activation strength from low to high stages, with Stage 3-4 dominating. This pattern indicates the model relies on semantic understanding for high-quality assessment.

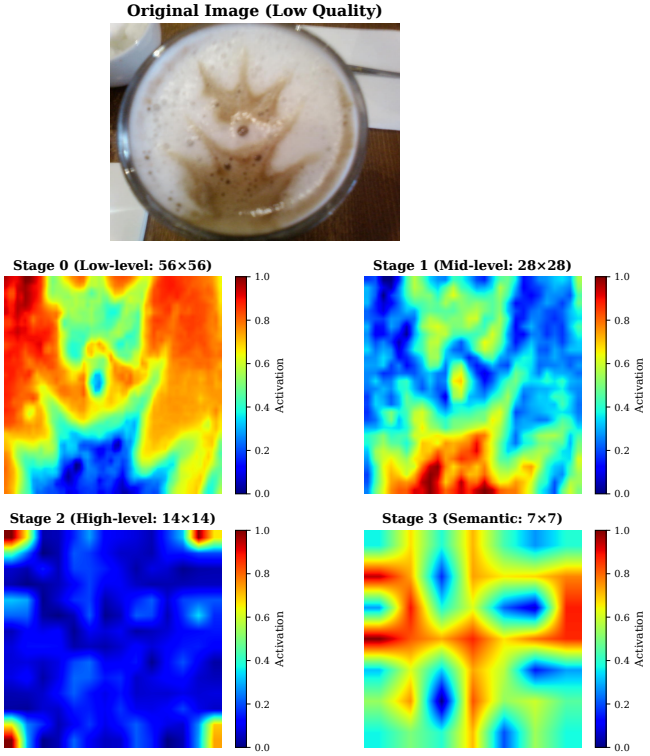

Fig. 9. Feature map visualization for a low-quality image (MOS 20.1/100). Strong activations in lower stages (Stage 1-2) indicate the model focuses on local distortions (compression artifacts, noise, texture degradation), while higher stages show weaker activations due to semantic content being obscured by distortions.