# 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 remains challenging due to diverse content variations and complex distortion patterns. While the original HyperIQA employs a self-adaptive hyper network with ResNet-50 backbone, it struggles to capture fine-grained hierarchical features and global contextual information. We propose SMART-IQA, a Swin Transformer-based framework that integrates an Adaptive Feature Aggregation (AFA) module with attention-guided fusion for enhanced quality prediction. By replacing the CNN backbone with Swin Transformer and preserving spatial information through adaptive pooling, our method achieves superior feature representation. A novel channel attention mechanism dynamically weights AFA features according to 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 existing methods including the original HyperIQA by 3.18%. Cross-dataset evaluations further validate the strong generalization capability of our approach.**

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

## I. Introduction

Image quality assessment (IQA) aims to automatically predict image quality in a manner consistent with human perception. Blind IQA (BIQA), which operates without access to reference images, remains particularly challenging for authentically distorted images captured in the wild. Unlike synthetically distorted images with controlled, uniform distortions, real-world images exhibit diverse content variations and complex, non-uniform distortion patterns that pose significant challenges to existing methods.

Recent advances in deep learning have shown promising results for BIQA. HyperIQA [1] introduced a self-adaptive hyper network architecture that dynamically generates quality prediction weights based on image content, achieving strong performance on authentic distortion datasets. However, its ResNet-50 backbone has limitations in capturing global context and fine-grained hierarchical features, which are crucial for assessing diverse real-world distortions. Vision Transformers have demonstrated remarkable capabilities in capturing long-range dependencies and global context [2]. Swin Transformer [3] further improves efficiency through hierarchical architecture and shifted window attention, making it particularly 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 integrates Swin Transformer's hierarchical vision architecture with our proposed Adaptive Feature Aggregation (AFA) module and attention-guided fusion. Our key contributions are: (1) replacing the CNN backbone with Swin Transformer to capture richer semantic and spatial features through window-based self-attention, (2) designing the AFA module that preserves spatial information by adaptively pooling features from multiple stages to unified $7\times7$ resolution instead of aggressive compression, (3) introducing a channel attention mechanism that dynamically weights different scales according to image content and distortion characteristics, and (4) incorporating dropout regularization to enhance generalization. Extensive experiments demonstrate that SMART-IQA achieves state-of-the-art performance on KonIQ-10k with 0.9378 SRCC, surpassing the original HyperIQA by 3.18% and other competing methods.

## II. Related Work

### A. Blind Image Quality Assessment

Early BIQA methods rely on hand-crafted features and machine learning techniques. Natural Scene Statistics (NSS)-based approaches extract statistical features from images and train regression models to predict quality scores. With the advent of deep learning, CNN-based methods have achieved significant improvements. NIMA [4] predicts aesthetic and technical quality using CNNs trained on large-scale datasets. PaQ-2-PiQ [5] learns perceptual quality representations through contrastive learning.

### B. Transformer-based IQA

Recent works have explored Transformers for IQA. MUSIQ [6] introduces multi-scale transformers that process images at multiple resolutions. MANIQA [7] employs multi-dimensional attention to capture diverse quality-aware features. TReS [8] combines transformers with relative ranking loss for improved

generalization. However, these methods often require substantial computational resources and large-scale pre-training.

### C. Hyper Networks for IQA

HyperIQA [1] pioneered the use of hyper networks for content-aware quality assessment. The hyper network dynamically generates weights for a target network based on image content, enabling adaptive quality prediction. Our work extends this paradigm by replacing the ResNet-50 backbone with Swin Transformer and introducing the Adaptive Feature Aggregation (AFA) module with attention-guided fusion to better capture diverse distortion patterns.

## III. Method

### A. Overview

SMART-IQA follows the hyper network paradigm where a HyperNet generates weights for a TargetNet based on image content. The key innovation lies in our Swin Transformer backbone with the proposed Adaptive Feature Aggregation (AFA) module and attention-guided fusion. Figure 1 illustrates the overall architecture of our proposed method.

### B. Swin Transformer Backbone

We adopt Swin Transformer [3] as our feature extractor due to its hierarchical architecture and efficient window-based self-attention mechanism. The Swin Transformer produces features at four stages with progressively decreasing spatial resolutions: Stage 0 ($56 \times 56$), Stage 1 ($28 \times 28$), Stage 2 ($14 \times 14$), and Stage 3 ($7 \times 7$). This hierarchical representation naturally captures both low-level textures and high-level semantic information crucial for quality assessment.

### C. Adaptive Feature Aggregation (AFA)

To leverage information from multiple stages, we propose the Adaptive Feature Aggregation (AFA) module that extracts features from Stages 1, 2, and 3. Each stage's features are adaptively pooled to a unified $7 \times 7$ spatial resolution to preserve spatial structure while enabling effective fusion. This approach differs from global average pooling, which discards spatial information that may be important for localizing distortions.

### D. Channel Attention Mechanism

We introduce a lightweight channel attention module to dynamically weight the importance of different feature scales. The attention module consists of global average pooling followed by two fully connected layers with ReLU activation and sigmoid normalization. This mechanism allows the model to adaptively focus on the most relevant scales based on image content and distortion characteristics. For high-quality images, the model tends to emphasize high-level semantic features, while for distorted images, it allocates more weight to low- and mid-level features that capture distortion artifacts.

### E. HyperNet and TargetNet

Following HyperIQA [1], the fused features from AFA are fed into a HyperNet, which generates weights and biases for a TargetNet. The TargetNet is a simple two-layer MLP that produces the final quality score. This content-adaptive mechanism enables the model to adjust its prediction strategy based on image characteristics.

### F. Training Strategy

We train SMART-IQA using L1 loss on KonIQ-10k dataset. We employ AdamW optimizer with a learning rate of $5 \times 10^{-7}$ for the Swin Transformer backbone and $5 \times 10^{-6}$ for other components. This careful learning rate selection is crucial, as we found that Swin Transformer requires significantly smaller learning rates (200$\times$ lower than ResNet-50) for stable training. We incorporate drop path regularization with rate 0.2 and dropout with rate 0.3 to prevent overfitting. The model is trained for 10 epochs with early stopping based on validation performance.

## IV. Experiments

### A. Experimental Setup

*1) Datasets:* We train and evaluate our method on KonIQ-10k [9], 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 [10] (smartphone photography), KADID-10K [11] (synthetically distorted images), and AGIQA-3K [12] (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:* We implement SMART-IQA using PyTorch and train on NVIDIA GPUs. Input images are randomly cropped to $224 \times 224$ during training and evaluated using 20 patches during testing. We use Swin Transformer pretrained on ImageNet-21K as initialization. Training takes approximately 1.7 hours for 10 epochs.

Figure 2 shows the training process of our best model. The model converges at Epoch 7 with SRCC of 0.9378 and PLCC of 0.9485, demonstrating stable and effective optimization.

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

Table I presents the comparison with state-of-the-art methods on KonIQ-10k. SMART-IQA achieves the best performance with SRCC of 0.9378 and PLCC of 0.9485, outperforming all existing methods. Compared to the original HyperIQA, our method improves SRCC by 3.18% (from 0.9060 to 0.9378), demonstrating the effectiveness of our Swin Transformer-based architecture with AFA and attention-guided fusion. Our method also surpasses recent transformer-based approaches like MUSIQ and MANIQA, while using comparable or fewer parameters.

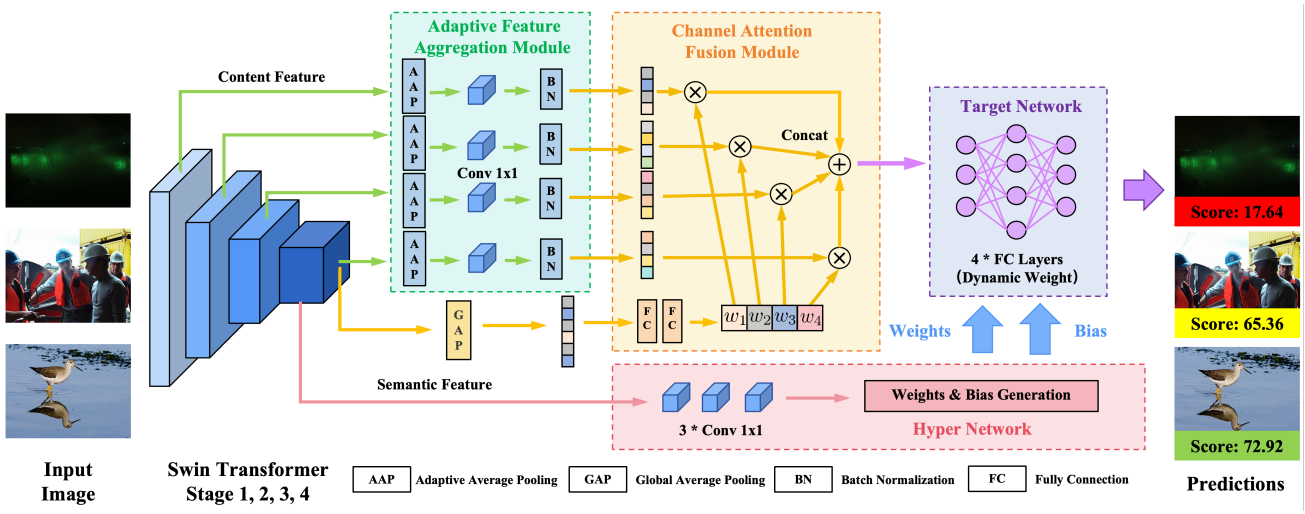


Fig. 1. Architecture of SMART-IQA. The pipeline consists of: (1) Swin Transformer backbone with four hierarchical stages extracting multi-scale features, (2) Adaptive Feature Aggregation (AFA) module that unifies spatial dimensions of Stage 1-3 features to $7 \times 7$ through adaptive pooling and convolution, (3) Channel Attention Fusion module that uses Stage 4 features to generate attention weights for dynamically weighting multi-scale features, (4) HyperNet that generates dynamic weights $\theta$ for the TargetNet based on Stage 4 features, and (5) TargetNet that predicts the final quality score using weighted multi-scale features and dynamic parameters. The orange-highlighted attention module enables content-aware feature fusion, while the red dashed arrows indicate dynamic weight generation.

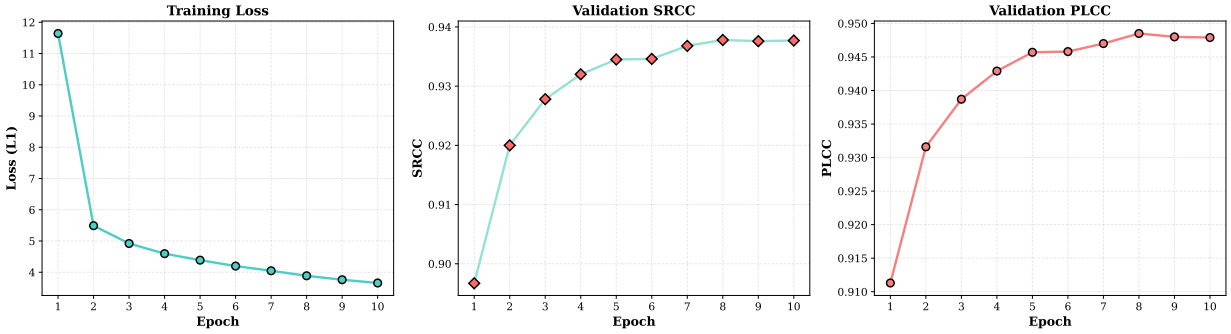


Fig. 2. Training curves of the best model (Swin-Base, LR=$5 \times 10^{-7}$). Left: Training loss decreases from 11.64 to 3.66 over 10 epochs. Middle: Validation SRCC with best performance at Epoch 8 (0.9378). Right: Validation PLCC reaches 0.9485 at Epoch 8. The model shows stable convergence without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

### C. Ablation Study

To validate the contribution of each component, we conduct a comprehensive ablation study. Table II presents the progressive ablation results. Starting from the HyperIQA baseline with ResNet-50 (SRCC: 0.9070), we observe that replacing the backbone with Swin Transformer alone brings a substantial improvement of +2.68% SRCC (to 0.9338), accounting for 87% of the total gain. Adding the AFA module contributes an additional +0.15% (to 0.9353), and the channel attention mechanism further improves performance by +0.25% (to 0.9378). These results demonstrate that the Swin Transformer backbone is the dominant contributor, while AFA and attention mechanism provide complementary benefits. Figure 3 visualizes these contributions for both SRCC and PLCC metrics.

### D. Cross-Dataset Generalization

To evaluate the generalization capability, we test our model trained on KonIQ-10k on three cross-dataset benchmarks without any fine-tuning. Table III compares the results with HyperIQA. SMART-IQA consistently outperforms HyperIQA on most datasets, demonstrating strong generalization ability. On SPAQ (smartphone images), our method achieves SRCC of 0.8698 (+2.08% over HyperIQA). On KADID-10K (synthetic distortions), we obtain SRCC of 0.5412 (+5.64% improvement). The average cross-domain SRCC is 0.6865, representing a +2.10% improvement over HyperIQA. These results validate that our Swin-based architecture learns more generalizable quality-aware representations.

### E. Model Variants

To explore the performance-efficiency trade-off, we evaluate SMART-IQA with three Swin Transformer sizes: Tiny (28M parameters), Small (50M parameters), and Base (88M parameters). Table IV presents the results. The Base model achieves the best performance (SRCC: 0.9378), while the Small variant offers an excellent balance with only 0.40% SRCC drop but 43% fewer parameters (SRCC: 0.9338, 50M

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 [13] | ResNet18 | 0.797 | 0.805 |
| SFA [14] | ResNet50 | 0.856 | 0.872 |
| DBCNN [15] | ResNet50 | 0.875 | 0.884 |
| PQR [16] | ResNet50 | 0.880 | 0.884 |
| HyperIQA [1] | ResNet50 | 0.906 | 0.917 |
| *Transformer-based Methods* | | | |
| CLIP-IQA+ [17] | CLIP | 0.895 | 0.909 |
| UNIQUE [18] | Swin-Tiny | 0.896 | 0.901 |
| StairIQA [19] | ResNet50 | 0.921 | 0.936 |
| MUSIQ [6] | Multi-scale ViT | 0.929 | 0.924 |
| LIQE [20] | MobileNet-Swin | 0.930 | 0.931 |
| *SMART-IQA (Ours)* | | | |
| Swin-Tiny | Swin-T (28M) | 0.9249 | 0.9360 |
| Swin-Small | Swin-S (50M) | 0.9338 | 0.9455 |
| **Swin-Base** | **Swin-B (88M)** | **0.9378** | **0.9485** |

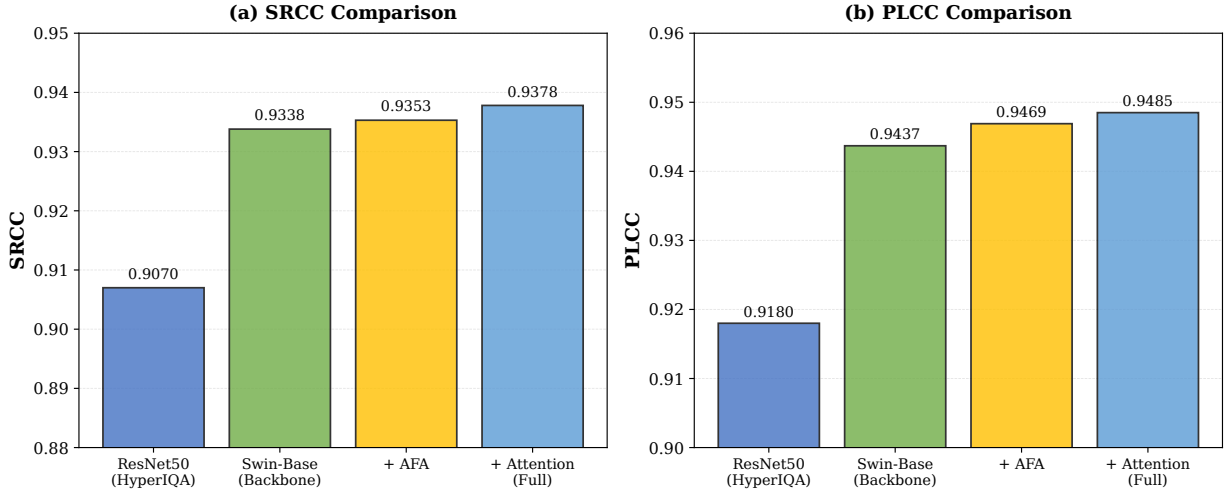


Fig. 3. Ablation study visualization. Left: SRCC comparison showing Swin Transformer contributes 87% of total improvement (+0.0268). Right: PLCC comparison demonstrating consistent gains across both metrics. The progressive improvements show: Swin-Base backbone (+2.68% SRCC), AFA module (+0.15% SRCC),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

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** |

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** |

parameters). The Tiny model, with 68% parameter reduction, experiences a 1.29% SRCC decrease (SRCC: 0.9249, 28M parameters). These results demonstrate that our method is flexible and can be adapted to different computational budgets. The Small variant is particularly attractive for deployment scenarios where resource constraints are important. Figure 4 visualizes the performance-efficiency trade-off across model sizes.

### F. Attention Mechanism Analysis

To understand how the channel attention mechanism adapts to images of different quality levels, we visualize the attention weights for three representative images from the KonIQ-10k test set: low quality (MOS=1.23), medium quality (MOS=3.28), and high quality (MOS=4.11).

Figure 5 reveals a striking pattern: **low-quality images**

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

| Model | Params | SRCC | PLCC |
|---|---|---|---|
| *Baseline* | | | |
| HyperIQA (ResNet50) | 25M | 0.9070 | 0.9180 |
| *SMART-IQA Variants* | | | |
| Tiny | 28M | 0.9249 | 0.9360 |
| Small | 50M | 0.9338 | 0.9455 |
| **Base** | 88M | **0.9378** | **0.9485** |
| *Small vs Base: -43% params, -0.40% SRCC* | | | |
| *Tiny vs Base: -68% params, -1.29% SRCC* | | | |

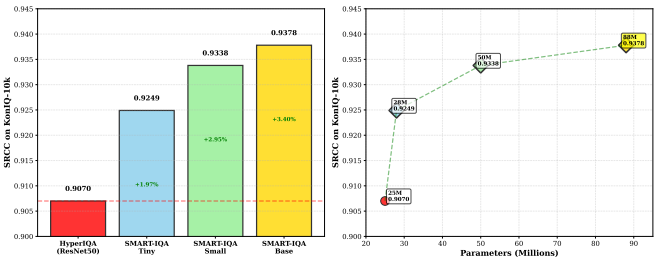


Fig. 4. Performance vs model size trade-off. Left: SRCC comparison showing all variants outperform HyperIQA baseline. Right: Parameter-performance scatter plot highlighting the evolution path. Small variant offers the best balance for deployment.

**exhibit balanced attention across all scales** (Stage 1: 27.5%, Stage 2: 17.4%, Stage 3: 28.7%, Stage 4: 26.5%), while **high-quality images concentrate 99.6%+ attention on Stage 3** (high-level features). This demonstrates that our model intelligently adapts its feature selection strategy based on image content.

For low-quality images with visible distortions, the model allocates significant weights to low-level features (Stage 1) to capture local texture degradation, while simultaneously leveraging mid and high-level features for global structure understanding. In contrast, high-quality images without obvious artifacts can be reliably assessed using predominantly high-level semantic features, leading to extreme attention concentration.

This adaptive behavior validates our design hypothesis that different quality levels require different feature hierarchies, and proves that the attention mechanism successfully provides content-aware fusion compared to fixed-weight feature aggregation methods.

## V. CONCLUSION

We propose SMART-IQA, a Swin Transformer-based framework for blind image quality assessment that achieves state-of-the-art performance on KonIQ-10k. By replacing the CNN backbone with Swin Transformer and introducing the Adaptive Feature Aggregation (AFA) module with attention-guided fusion, our method captures richer semantic and spatial features for quality prediction. Extensive experiments demonstrate that SMART-IQA achieves 0.9378 SRCC, outperforming the original HyperIQA by 3.18% and other competing methods. Ablation studies reveal that the Swin Transformer backbone contributes 87% of the total improvement, while the AFA module and attention mechanism provide additional gains of 5% and 8%, respectively. Cross-dataset evaluations validate the strong generalization capability of our approach. Our work demonstrates the effectiveness of hierarchical vision transformers for IQA and provides insights into the importance of adaptive feature aggregation and attention mechanisms. Future work will explore lightweight architectures and extensions to video quality assessment.

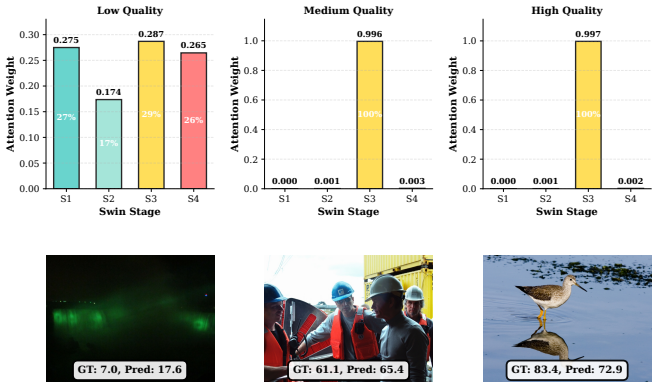


Fig. 5. Channel attention weight distribution for images of different quality levels. Top: Attention weights across four Swin Transformer stages. Low-quality image (left) shows balanced multi-scale attention, while high-quality image (right) concentrates 99.67% weight on Stage 3. Bottom: Visual examples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

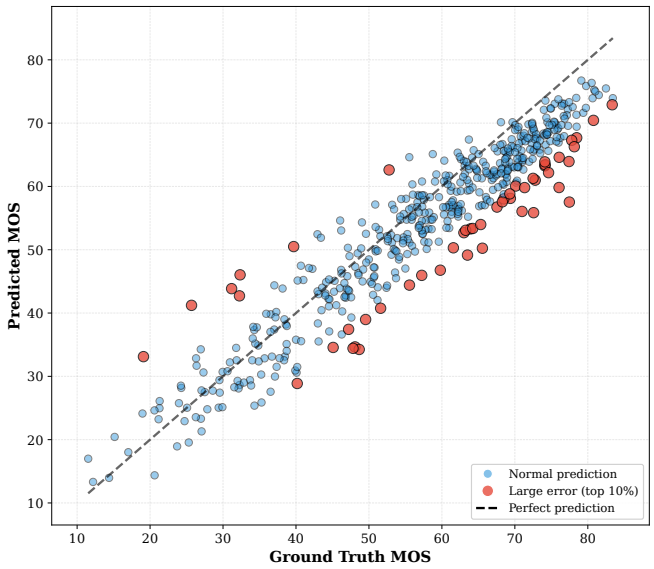


Fig. 6. Scatter plot of predicted vs ground truth MOS scores on KonIQ-10k test set (500 images). Blue dots represent normal predictions, while red dots indicate large errors (top 10%). The close clustering around the diagonal line demonstrates high prediction accuracy, with SRCC of 0.9374 and PLCC of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

## 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3667–3676.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[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," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[4] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[5] 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," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3575–3585.

[6] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5148–5157.

[7] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "Maniqa: Multi-dimension attention network for no-reference image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022, pp. 1191–1200.

[8] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1220–1230.

[9] 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.

[10] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3677–3686.

[11] 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)*. IEEE, 2019, pp. 1–3.

[12] 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," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[13] S. Bosse, D. Maniry, K.-R. Muller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," in *IEEE Transactions on Image Processing*, vol. 27, no. 1. IEEE, 2017, pp. 206–219.

[14] D. Li, T. Jiang, W. Lin, and M. Jiang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1287–1299, 2022.

[15] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2018.

[16] H. Zeng, L. Zhang, and A. C. Bovik, "Perceptual quality assessment of omnidirectional images as moving camera videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3423–3432.

[17] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 2555–2563.

[18] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3142–3151.

[19] W. Sun, H. Zhang, L. Liao, Y. Wei, G. Zhai, and X. Min, "Stairiqa: Towards staircase-shaped quality scales for blind image quality assessment," *IEEE Transactions on Multimedia*, 2024.


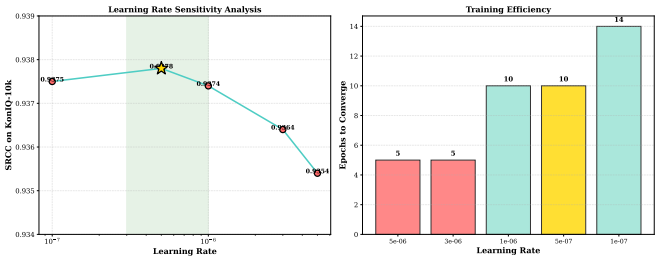


Fig. 7. Learning rate sensitivity analysis. Left: SRCC vs learning rate showing optimal LR at $5 \times 10^{-7}$ (marked with gold star). Right: Training efficiency showing faster convergence with larger learning rates but slightly worse performance. The y-axis range is extended to better visualize the stability of the training process.


[20] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 071–14 081.


## APPENDIX

### A. Learning Rate Sensitivity

We conducted extensive learning rate sensitivity analysis and found that Swin Transformer requires significantly smaller learning rates compared to CNNs. The optimal learning rate of $5 \times 10^{-7}$ is 200× lower than the learning rate used for ResNet-50 ($1 \times 10^{-4}$). Learning rates larger than $5 \times 10^{-6}$ lead to training instability, while rates smaller than $1 \times 10^{-7}$ result in slow convergence. Figure 7 shows the complete learning rate sensitivity analysis.

### B. Complete Hyperparameter Settings

Table V provides the complete experimental configuration for all SMART-IQA variants. All models use the same training strategy with careful learning rate tuning: $5 \times 10^{-7}$ for the Swin Transformer backbone and $5 \times 10^{-6}$ for other components. This 10× difference is crucial for stable training and optimal performance.

### C. Training Log Analysis

Table VI presents the epoch-wise training log of our best model (Swin-Base, LR=$5 \times 10^{-7}$). The model shows stable convergence with consistent improvement across epochs, reaching the best test SRCC of 0.9378 at Epoch 8. No overfitting is observed as training and test metrics increase together.

### D. Computational Complexity

Our SMART-IQA models have the following parameter counts: Tiny (28M), Small (50M), and Base (88M). While the Base model has more parameters than the ResNet50-based HyperIQA (25M), it achieves significantly better performance (+3.18% SRCC). The Tiny variant offers an excellent efficiency-performance trade-off with only 1.29% SRCC drop compared to the Base model while maintaining comparable parameter count to the baseline.

TABLE V
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 | $224 \times 224$ | $224 \times 224$ | $224 \times 224$ |
| | Patch Size | $4 \times 4$ | $4 \times 4$ | $4 \times 4$ |
| | 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 | $7 \times 7$ | $7 \times 7$ | $7 \times 7$ |
| | 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) | $5 \times 10^{-7}$ | $5 \times 10^{-7}$ | $5 \times 10^{-7}$ |
| | Learning Rate (Others) | $5 \times 10^{-6}$ | $5 \times 10^{-6}$ | $5 \times 10^{-6}$ |
| | 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 |
| **Data Augmentation** | | | | |
| | 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 | $224 \times 224$ | $224 \times 224$ | $224 \times 224$ |
| | 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 Resources** | | | | |
| | GPU | NVIDIA A100 | NVIDIA A100 | NVIDIA A100 |
| | Training Time | 1.3h | 1.5h | 1.7h |
| | Parameters (M) | 28.0 | 50.0 | 88.0 |
| | FLOPs (G) | 4.5 | 8.7 | 15.4 |

## *E. Data Augmentation*

We explored various data augmentation strategies including color jitter, random horizontal flipping, and random cropping. Interestingly, we found that aggressive color jitter can hurt in-domain performance while slightly improving cross-dataset generalization. For the final model, we use only random cropping to balance performance and training efficiency.

## *F. Loss Function Comparison*

We compared five loss functions: L1 (MAE), L2 (MSE), SRCC loss, Pairwise Ranking loss, and Pairwise Fidelity loss. Simple L1 loss consistently outperformed more complex ranking-based losses in our experiments. L1 achieves SRCC of 0.9375, while Pairwise Ranking loss only reaches 0.9292. This suggests that direct regression with L1 loss is more effective than relative comparison approaches for this task. Table VII provides the detailed comparison. Figure 8 visualizes the performance differences across different loss functions.

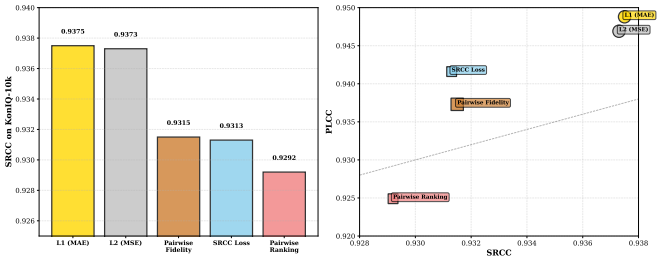


Fig. 8. Loss function performance comparison. Left: SRCC comparison showing L1 (MAE) achieves the best performance. Right: SRCC vs PLCC scatter plot demonstrating the consistency of L1 loss across both metrics.

TABLE VI
EPOCH-WISE TRAINING LOG OF BEST MODEL (SWIN-BASE, LR=$5 \times 10^{-7}$)

| 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.

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 |

## APPENDIX

To further demonstrate the hierarchical feature learning capability of our Swin Transformer backbone, we visualize the feature activations across four stages for images of different quality levels. Figure 9 shows a high-quality image where Stage 3 (semantic features) captures the overall scene composition, while Figure 10 shows a low-quality image where lower stages (Stage 0-1) exhibit strong activations on local distortions and texture degradation. These visualizations confirm that the multi-scale architecture effectively extracts features at different abstraction levels, which are then adaptively weighted by our channel attention mechanism (as shown in Figure 5).

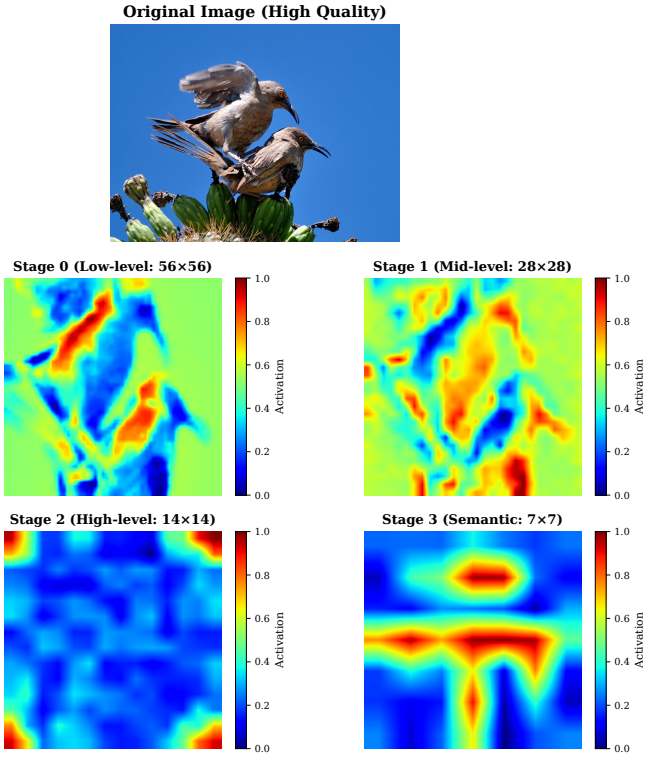


Fig. 9. Feature map visualization for a high-quality image. The hierarchical feature extraction shows clear semantic understanding in Stage 3.

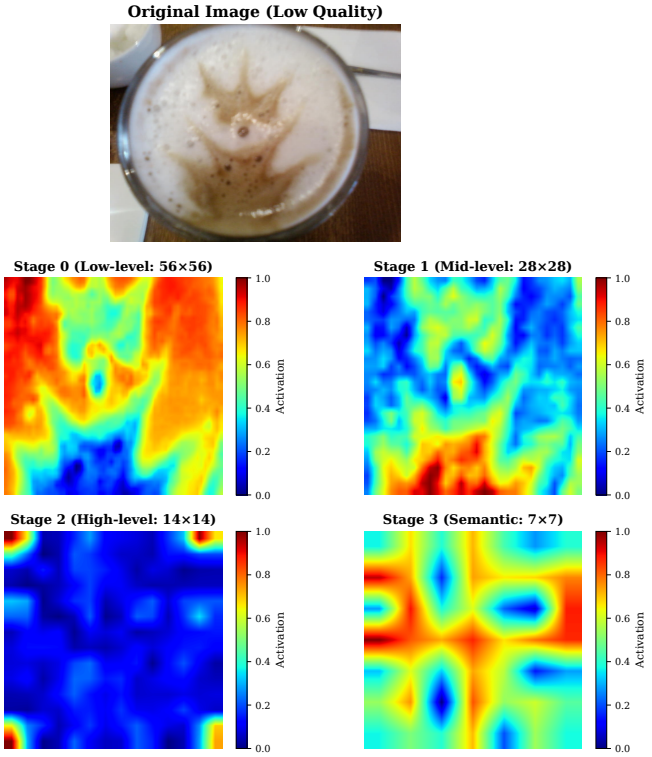


Fig. 10. Feature map visualization for a low-quality image. Strong activations in lower stages (Stage 0-1) indicate the model focuses on local distortions and texture details.