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
nuoyanchen@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 multi-scale features and global contextual information. We propose SMART-IQA, a Swin Transformer-based framework that integrates multi-scale spatial features 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 multi-scale 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, Multi-scale Feature Fusion, 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 multi-scale 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 multi-scale attention-guided feature 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) preserving spatial information by adaptively pooling multi-scale features to 7×7 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 multi-scale 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 multi-scale attention fusion.

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 (56×56), Stage 1 (28×28), Stage 2 (14×14), and Stage 3 (7×7). This multi-scale representation naturally captures both low-level textures and high-level semantic information crucial for quality assessment.

C. Multi-scale Feature Fusion

To leverage information from multiple scales, we extract features from Stages 1, 2, and 3. Each stage's features are adaptively pooled to a unified 7×7 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 multi-scale features 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 5×10^{-7} for the Swin Transformer backbone and 5×10^{-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 224×224 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.

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 multi-scale attention fusion. Our method also surpasses recent transformer-based approaches like MUSIQ and MANIQA, while using comparable or fewer parameters.

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 multi-scale fusion contributes an additional +0.15% (to 0.9353), and the channel attention mechanism further improves performance by +0.25% (to

TABLE I
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON KONIQ-10K DATASET

Method	Year	Backbone	SRCC	PLCC
NIMA [4]	2018	InceptionNet	0.558	0.590
PaQ-2-PiQ [5]	2020	ResNet18	0.892	0.904
HyperIQA [1]	2020	ResNet50	0.906	0.917
MUSIQ [6]	2021	Multi-scale ViT	0.915	0.930
TReS [8]	2022	Transformer	0.908	0.922
MANIQA [7]	2022	ViT-Small	0.920	0.930
SMART-IQA (Ours)	2024	Swin-Base	**0.9378**	**0.9485**

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

Configuration	Multi-Scale	Attention	SRCC
Baseline			
HyperIQA (ResNet50)	-	-	0.9070
Progressive Ablation (Swin-Base)			
Backbone only	×	×	0.9338
+ Multi-Scale	✓	×	0.9353
+ Attention (Full)	✓	✓	**0.9378**
Component Contributions			
Swin Transformer: +0.0268 SRCC (87% of gain)			
Multi-Scale Fusion: +0.0015 SRCC (5% of gain)			
Attention Mechanism: +0.0025 SRCC (8% of gain)			
Total Improvement	**+0.0308 SRCC (+3.18%)**		

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**

0.9378). These results demonstrate that the Swin Transformer backbone is the dominant contributor, while multi-scale fusion and attention mechanism provide complementary benefits.

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 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.

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			

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 multi-scale 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 multi-scale fusion 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 multi-scale feature fusion and adaptive attention mechanisms. Future work will explore lightweight architectures and extensions to video quality assessment.

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.

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 5×10^{-7} is 200$\times$ lower than the learning rate used for ResNet-50 (1×10^{-4}). Learning rates larger than 5×10^{-6} lead to training instability, while rates smaller than 1×10^{-7} result in slow convergence.

B. 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.

C. 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.