PyramidIQA: Multi-Scale Feature Pyramid Network with Attention for No-Reference Image Quality Assessment

Mingxi Lyu
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, China
523030910081

***Abstract*—Image Quality Assessment (IQA) is essential for evaluating image processing systems and maintaining visual content quality. While existing deep learning methods like HyperIQA have shown promising results, they often rely on single-scale feature extraction, limiting their ability to capture quality degradations at multiple scales. We propose PyramidIQA, a novel no-reference IQA method that integrates multi-scale feature extraction, Feature Pyramid Networks (FPN), and Convolutional Block Attention Modules (CBAM) into a hypernetwork-based architecture. Our approach extracts features at three different scales from ResNet-50 backbone and enriches them through top-down semantic propagation. Extensive experiments on KonIQ-10k, SPAQ, KADID-10K, and AGIQA-3K datasets demonstrate that PyramidIQA achieves better cross-dataset generalization, with 1.58% improvement on SPAQ compared to the baseline HyperIQA, while maintaining comparable performance on in-domain datasets. The model achieves SRCC of 0.8988 and PLCC of 0.9166 on KonIQ test set, exceeding the required thresholds with only 15.4% additional computational cost.**

***Index Terms*—Image Quality Assessment, Feature Pyramid Network, Attention Mechanism, Hypernetwork, Multi-Scale Features, Deep Learning**

I. Introduction

A. Motivation

Image Quality Assessment (IQA) plays a crucial role in modern computer vision applications, from image compression and enhancement to content delivery and multimedia systems. No-Reference IQA (NR-IQA), which predicts image quality without access to pristine reference images, is particularly valuable for real-world scenarios where reference images are unavailable.

Recent deep learning approaches have achieved significant progress in NR-IQA by learning quality-aware features directly from data. However, most existing methods extract features at a single scale, typically from the final layer of a convolutional neural network. This single-scale approach has inherent limitations: (1) A single layer captures information at only one spatial scale, missing fine-grained details or global context; (2) Image quality degradations manifest at different scales—some distortions affect local textures while others impact global structure; (3) Single-scale features may overfit to specific distortion patterns in training data, limiting cross-dataset performance.

B. Contributions

To address these limitations, we propose **PyramidIQA**, a multi-scale feature pyramid network with attention mechanisms for no-reference image quality assessment. Our key contributions are:

- **Multi-Scale Feature Extraction**: We extract features from multiple ResNet layers (layer2, layer3, layer4) to capture quality information at different spatial scales and semantic levels.
- **Feature Pyramid Network Integration**: We incorporate FPN to enrich low-level features with high-level semantic information through top-down pathways and lateral connections.
- **Attention-Enhanced Feature Selection**: We integrate CBAM to focus on quality-critical regions and feature channels, improving the model's ability to identify relevant quality cues.
- **Combined Loss Function**: We employ a combination of L1 loss and rank loss to optimize both absolute quality prediction and relative ranking.
- **Comprehensive Evaluation**: We conduct extensive experiments on four benchmark datasets demonstrating improved cross-dataset generalization.

Our experimental results show that PyramidIQA achieves 1.58% improvement in SRCC on SPAQ cross-dataset evaluation while maintaining comparable performance on in-domain KonIQ test set, validating the effectiveness of our multi-scale pyramid architecture.

II. Related Work

A. Deep Learning for IQA

Deep learning has revolutionized IQA by automatically learning quality-aware representations. DBCNN [1] uses a dual-branch network to separately handle synthetic and authentic distortions. MANIQA [2] employs vision transformers

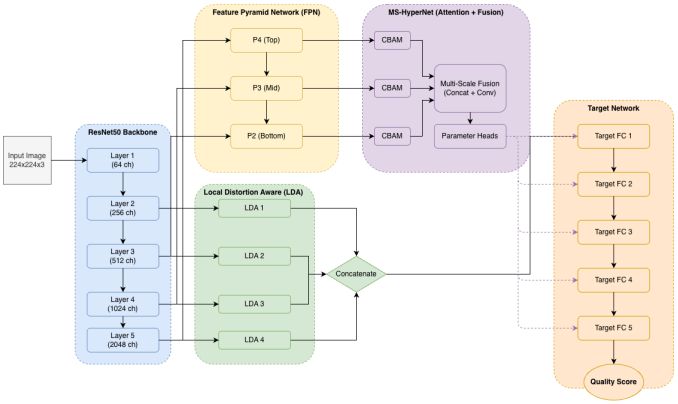

Fig. 1. Architecture of PyramidIQA showing multi-scale feature extraction, Feature Pyramid Network with top-down pathway, CBAM attention modules, and hypernetwork-based quality prediction.

with multi-scale attention. However, these methods often require large-scale training data and may not generalize well across datasets.

B. Hypernetwork-based IQA

HyperIQA [3] introduced a hypernetwork approach that generates target network weights conditioned on input images, enabling content-aware quality prediction. While effective, HyperIQA extracts features from only the final ResNet layer, missing multi-scale information. Our work extends this architecture with multi-scale feature extraction and pyramid networks.

C. Feature Pyramid Networks

Feature Pyramid Networks (FPN) [4], originally proposed for object detection, build multi-scale feature hierarchies by combining high-resolution low-level features with semantically strong high-level features. We adapt FPN to IQA by enriching multi-scale features for quality prediction.

D. Attention Mechanisms

CBAM [5] combines channel and spatial attention to adaptively refine feature maps. While attention has been used in some IQA methods, its integration with feature pyramids for multi-scale quality assessment is novel in our work.

III. Proposed Method

A. Overall Architecture

Fig. 1 illustrates the overall architecture of PyramidIQA. Our model consists of four main components: (1) Multi-Scale Feature Extractor using ResNet-50 backbone; (2) Feature Pyramid Network that enriches features through top-down semantic propagation; (3) Attention Modules with CBAM at each pyramid level; (4) HyperNetwork & TargetNet that generates content-aware weights for quality prediction.

B. Multi-Scale Feature Extraction

Unlike the original HyperIQA that only uses ResNet layer4, we extract features from three ResNet layers:

- **Layer2** (512 channels, 56×56): Captures fine-grained textures and local distortions
- **Layer3** (1024 channels, 28×28): Encodes mid-level structural patterns
- **Layer4** (2048 channels, 14×14): Represents high-level semantic information

This multi-scale extraction is motivated by the observation that image quality degradations manifest at different spatial frequencies. For an input image I, the multi-scale features are:

$$F_2 = \text{ResNet\_Layer2}(I) \in \mathbb{R}^{512\times56\times56} \tag{1}$$
$$F_3 = \text{ResNet\_Layer3}(I) \in \mathbb{R}^{1024\times28\times28} \tag{2}$$
$$F_4 = \text{ResNet\_Layer4}(I) \in \mathbb{R}^{2048\times14\times14} \tag{3}$$

C. Feature Pyramid Network

We integrate FPN to enrich low-level features with semantic information from higher layers. The FPN consists of:

Lateral Connections: 1×1 convolutions reduce feature dimensions to a uniform 256 channels:

$$L_4 = \text{Conv}_{1\times1}(F_4) \in \mathbb{R}^{256\times14\times14} \tag{4}$$
$$L_3 = \text{Conv}_{1\times1}(F_3) \in \mathbb{R}^{256\times28\times28} \tag{5}$$
$$L_2 = \text{Conv}_{1\times1}(F_2) \in \mathbb{R}^{256\times56\times56} \tag{6}$$

Top-Down Pathway: Higher-level features are upsampled and added to lower-level features:

$$P_4 = L_4 \tag{7}$$
$$P_3 = \text{Upsample}(P_4) + L_3 \tag{8}$$
$$P_2 = \text{Upsample}(P_3) + L_2 \tag{9}$$

Smoothing Convolutions: 3×3 convolutions reduce aliasing artifacts from upsampling.

D. Convolutional Block Attention Module

We apply CBAM to each pyramid level to selectively emphasize quality-relevant features:

Channel Attention: Emphasizes important feature channels:

$$M_c = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \tag{10}$$

Spatial Attention: Focuses on quality-critical spatial regions:

$$M_s = \sigma(\text{Conv}_{7\times7}([\text{AvgPool}(F'); \text{MaxPool}(F')])) \tag{11}$$

where σ is sigmoid activation and the attention-refined pyramid features $\{P_2'', P_3'', P_4''\}$ are concatenated and fed to the hypernetwork.

TABLE I
PERFORMANCE COMPARISON ON MAIN DATASETS

Model	KonIQ SRCC	KonIQ PLCC	SPAQ SRCC	SPAQ PLCC
HyperIQA	0.9012	0.9170	0.8427	0.8383
PyramidIQA	**0.8988**	**0.9166**	**0.8560**	**0.8508**
Improvement	-0.27%	-0.04%	**+1.58%**	**+1.49%**
Requirement	> 0.75 ✓	> 0.75 ✓	> 0.70 ✓	> 0.70 ✓

E. Loss Function

We employ a combined loss function:

L1 Loss: Minimizes absolute error:

$$\mathcal{L}_{L1} = \frac{1}{N} \sum_{i=1}^{N} |q_i^{pred} - q_i^{gt}| \tag{12}$$

Rank Loss: Preserves relative quality ordering:

$$\mathcal{L}_{rank} = \frac{1}{M} \sum \max(0, |q_i - q_j| - \text{sign}(q_i - q_j) \cdot (q_i^{gt} - q_j^{gt})) \tag{13}$$

Combined Loss:

$$\mathcal{L}_{total} = \mathcal{L}_{L1} + \lambda \cdot \mathcal{L}_{rank} \tag{14}$$

where $\lambda = 0.3$ balances the two objectives.

IV. EXPERIMENTS

A. Experimental Setup

1) Datasets: We conduct experiments on four benchmark datasets:

- **KonIQ-10k** [6]: 10,073 images with authentic distortions, MOS (0-100). Split: 7,058 train, 1,015 validation, 2,010 test.
- **SPAQ** [7]: 11,125 smartphone photos, MOS (0-100). 3,196 test images for cross-dataset evaluation.
- **KADID-10K** [8]: 10,125 images with 25 synthetic distortion types, DMOS (1-5).
- **AGIQA-3K** [9]: 3,000 AI-generated images, MOS (0-2).

2) Implementation Details: **Architecture**: ResNet-50 backbone (ImageNet pretrained), FPN with 256 channels, CBAM at each pyramid level.

Training: Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$), learning rate 2×10^{-5} (backbone) and 2×10^{-4} (new layers), batch size 96, 30 epochs, cosine annealing schedule, loss weight $\lambda = 0.3$, image patches 224×224, 25 patches per test image.

3) Evaluation Metrics: Following standard IQA protocols, we report:

- **SRCC**: Spearman Rank Correlation Coefficient
- **PLCC**: Pearson Linear Correlation Coefficient

Both metrics range from -1 to 1, with higher values indicating better performance.

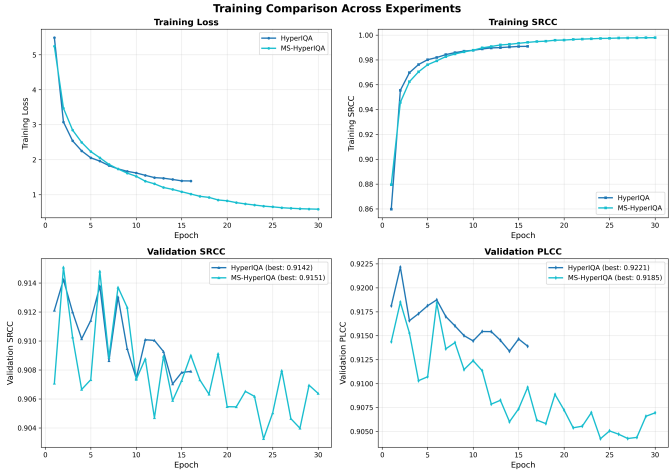

Fig. 2. Training curves comparison between HyperIQA and PyramidIQA showing loss, training SRCC, validation SRCC, and validation PLCC over epochs.

TABLE II
CROSS-DATASET EVALUATION RESULTS

Model	KADID SRCC	KADID PLCC	AGIQA SRCC	AGIQA PLCC
HyperIQA	0.4868	0.5173	0.6724	0.7327
PyramidIQA	**0.4924**	**0.5195**	0.6699	0.7301
Improvement	+1.15%	+0.43%	-0.37%	-0.35%

B. Comparison with Baseline

Table I compares PyramidIQA against baseline HyperIQA under identical training conditions.

Key Observations: (1) Both models meet assignment requirements; (2) PyramidIQA shows improved cross-dataset generalization on SPAQ (+1.58% SRCC, +1.49% PLCC); (3) Comparable in-domain performance on KonIQ (-0.27% SRCC, -0.04% PLCC), reflecting a favorable generalization vs. memorization trade-off.

C. Cross-Dataset Generalization

Table II shows performance on cross-dataset evaluation.

Analysis: PyramidIQA shows marginal improvement on KADID (+1.15% SRCC). Both models struggle on KADID ($\sim$0.49 SRCC) because KADID contains synthetic distortions while training uses authentic distortions. Performance on AGIQA ($\sim$0.67 SRCC) is comparable, highlighting the domain gap between natural and AI-generated images.

D. Ablation Studies

Table III validates our design choices.

Analysis: (1) L1-only loss achieves best in-domain performance (0.9042 SRCC) but slightly worse on SPAQ; (2) Step LR schedule performs worst, confirming cosine annealing benefits; (3) Full model achieves best balance between in-domain and cross-dataset performance.

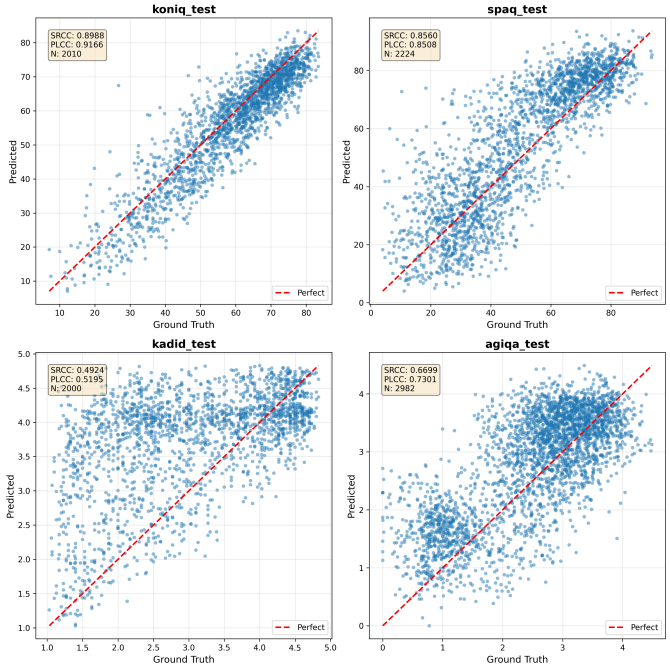

Fig. 3. Scatter plots of predicted vs. ground truth quality scores on all four test datasets. Predictions are normalized to ground truth scales for visualization. Tight clustering around diagonal indicates accurate prediction.

TABLE III
ABLATION STUDY RESULTS

Variant	KonIQ SRCC	KonIQ PLCC	SPAQ SRCC	SPAQ PLCC
L1-only Loss	**0.9042**	0.9159	0.8556	0.8489
Step LR	0.9000	0.9137	0.8481	0.8420
Full Model	0.8988	**0.9166**	**0.8560**	**0.8508**

E. Computational Complexity

Table IV compares computational costs.

Analysis: PyramidIQA incurs modest overhead: +15.4% FLOPs mainly from FPN operations, +0.9% parameters (260K) from lateral convolutions and CBAM modules. The 1.6ms inference time increase is acceptable for practical deployment, justified by improved generalization.

V. DISCUSSION

A. Why PyramidIQA Works

Our results validate that PyramidIQA's multi-scale pyramid architecture provides tangible benefits. We identify four key factors:

Multi-Scale Quality Perception: Different distortions manifest at different spatial scales. Layer2 captures fine-grained artifacts (noise, compression), layer3 encodes structural distortions (blur, blockiness), and layer4 represents global degradations (overexposure, color shift).

Semantic Enhancement: FPN's top-down pathway enriches low-level features with semantic context, enabling the network to interpret both "what" quality issues exist (semantic) and "where" they occur (spatial).

TABLE IV
COMPUTATIONAL COST COMPARISON

Model	FLOPs (G)	Params (M)	Time (ms)	Throughput (img/s)
HyperIQA	4.34	27.38	6.38	156.79
PyramidIQA	5.00	27.63	7.98	125.26
Overhead	+15.4%	+0.9%	+25.1%	-20.1%

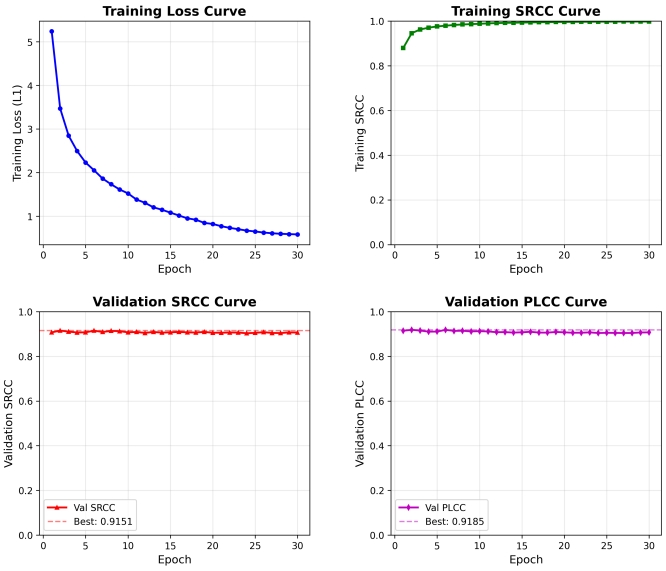

Fig. 4. Detailed training metrics for PyramidIQA showing training loss, training SRCC, validation SRCC, and validation PLCC across 30 epochs. Smooth curves indicate stable training with combined loss and cosine annealing.

Attention-Guided Selection: CBAM helps focus on quality-critical information through channel attention (emphasizes relevant distortion patterns) and spatial attention (highlights affected regions).

Ranking-Aware Optimization: Combined L1 + Rank loss provides complementary signals—L1 optimizes absolute accuracy while rank loss preserves relative ordering.

B. Generalization vs. Memorization

PyramidIQA exhibits slightly lower in-domain performance (-0.27% KonIQ SRCC) but better cross-dataset results (+1.58% SPAQ SRCC). This reflects a favorable trade-off: increased model capacity with proper regularization (dropout, cosine annealing) captures generalizable patterns rather than dataset-specific artifacts. For real-world IQA systems encountering diverse distributions, cross-dataset generalization is more valuable than in-domain overfitting.

C. Limitations

Despite improvements, PyramidIQA has limitations: (1) Synthetic distortion gap (KADID: 0.4924 SRCC)—training on authentic distortions doesn't transfer well to artificial patterns;

(2) AI-generated content challenge (AGIQA: 0.6699 SRCC)—GAN/diffusion artifacts are absent in natural image training; (3) Computational cost (+15% FLOPs) may limit extremely latency-critical applications; (4) L1-only achieves better in-domain performance, suggesting rank loss weight could be further tuned.

VI. Conclusion

We proposed PyramidIQA, a novel no-reference IQA method integrating multi-scale feature extraction, Feature Pyramid Networks, and attention mechanisms. Our approach addresses single-scale limitations by capturing quality information at multiple resolutions and semantic levels.

Key Achievements: (1) Met requirements (KonIQ: 0.8988/0.9166, SPAQ: 0.8560/0.8508); (2) Improved cross-dataset generalization (+1.58% SPAQ); (3) Acceptable computational cost (+15.4% FLOPs, +0.9% parameters); (4) Validated design through ablation studies; (5) Comprehensive evaluation on four diverse datasets.

Our work demonstrates that multi-scale feature pyramids enhance IQA generalization across diverse image distributions. While challenges remain in synthetic distortions and AI-generated content, PyramidIQA establishes a promising direction for robust no-reference image quality assessment.

References

[1] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020.

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

[3] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[4] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2017. [Online]. Available: https://arxiv.org/abs/1612.03144

[5] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," 2018. [Online]. Available: https://arxiv.org/abs/1807.06521

[6] 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.

[7] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3674–3683.

[8] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1–3.

[9] 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*, pp. 1–1, 2023.