Triplet-Based Feature Refinement and Quality Difference Perception Network for No-Reference Image Quality Assessment

Chunyu Wu, Yihua Chen, Kejing Wu, Zhenjun Tang*
Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, Guilin 541004, China
Guangxi Key Lab of Multi-Source Information Mining & Security, Guangxi Normal University, Guilin 541004, China

***Abstract*—No-Reference Image Quality Assessment (NR-IQA) is an important technology in the field of computer vision. Most existing NR-IQA methods are constrained by the absence of reference images. These issues lead to the limited IQA performance. To address this, we propose a novel method called Triplet-Based Feature Refinement and Quality Difference Perception Network (TRQDNet) for No-Reference Image Quality Assessment. It introduces the restored image and the degraded image as dual reference images to form a triplet. The method consists of a multi-scale spatial-channel refinement module, a difference-guided distortion perception module and a quality prediction module. Firstly, the multi-scale spatial-channel refinement module refines the features extracted by the pre-trained Swin Transformer from the triplet. Secondly, the difference-guided distortion perception module constructs restoration and degradation difference features and performs cross-attention across the triplet features. It captures quality differences and learns difference-aware representations effectively. Finally, the quality prediction module predicts the score through weighted feature aggregation. Experimental results demonstrate that the TRQDNet outperforms some excellent NR-IQA methods in both prediction accuracy and generalization ability.**

***Index Terms*—No-Reference Image Quality Assessment, Dual Reference Images, Cross-Attention, Quality Difference Perception**

I. Introduction

With the rapid advancement of digital technologies and the explosive growth of online media, images have become a dominant form of information carrier for communication and content sharing [1], [2], [3], [4]. However, visual quality is often unavoidably degraded during acquisition, compression, transmission, resulting in unsatisfactory viewing experiences. Accordingly, researchers have been exploring how to effectively capture quality change for further optimization. Image quality assessment (IQA) is an important technique which aims to model the perceptual characteristics of the human visual system (HVS) for automatic quality evaluation [5]. According to the availability of reference images, IQA methods can be generally categorized into full-reference (FR-IQA) [6], reduced-reference (RR-IQA) [7], and no-reference (NR-IQA) methods [8], [9]. Since the reference image can act as the reference standard, FR-IQA typically achieves better performance. However, the reference images are often difficult to obtain in real-world scenarios. Therefore, NR-IQA has attracted increasing attention in both research and practical applications.

In recent years, many NR-IQA methods based on deep neural networks have been proposed. In [10], Kang et al. introduce CNN networks into the field of NR-IQA for the first time. Inspired by this method, many NR-IQA architectures based on CNN networks have emerged [11], [12]. For example, in [13], Zhang et al. extract both synthetic and authentic distortion features using the CNN network, and then fuse these features to predict the quality score. In [14], Tang et al. adopt the CNN network to extract multi-scale features, then refine and fuse them through a pyramid architecture to predict the quality score. Furthermore, some researchers have incorporated the transformer architecture into NR-IQA to improve the ability to learn high-level information. In [15], Ke et al. introduce the MUSIQ architecture, showcasing the transformer architecture for the first time. In [16], Golestaneh et al. also use the transformer architecture and introduce a relative ranking strategy and a self-consistency strategy to improve the model performance. In [17], Qin et al. also utilize the transformer-based encoder–decoder architecture and employ a quality-aware attention-panel decoder to learn effective quality representations from the limited data. In [18], Xu et al. concatenate multi-scale local features extracted from pre-trained CNN, and gradually inject these features into the pre-trained vision transformer for quality score prediction.

However, these NR-IQA methods have achieved limited performance improvement due to the absence of reference standards resulting from the lack of reference images. This violates the human contrastive perception mechanism [19]. It means that humans tend to assess image quality by comparing different images rather than making absolute judgments on a single one. To address this, we propose a method called Triplet-Based Feature Refinement and Quality Difference Perception Network (TRQDNet) for NR-IQA. Specifically, we introduce the restored image and the degraded image as dual

*Corresponding author: Zhenjun Tang (Email: tangzj230@163.com).
This work was partially supported by the National Natural Science Foundation of China (6227211) and Guangxi "Bagui Scholar" Project.

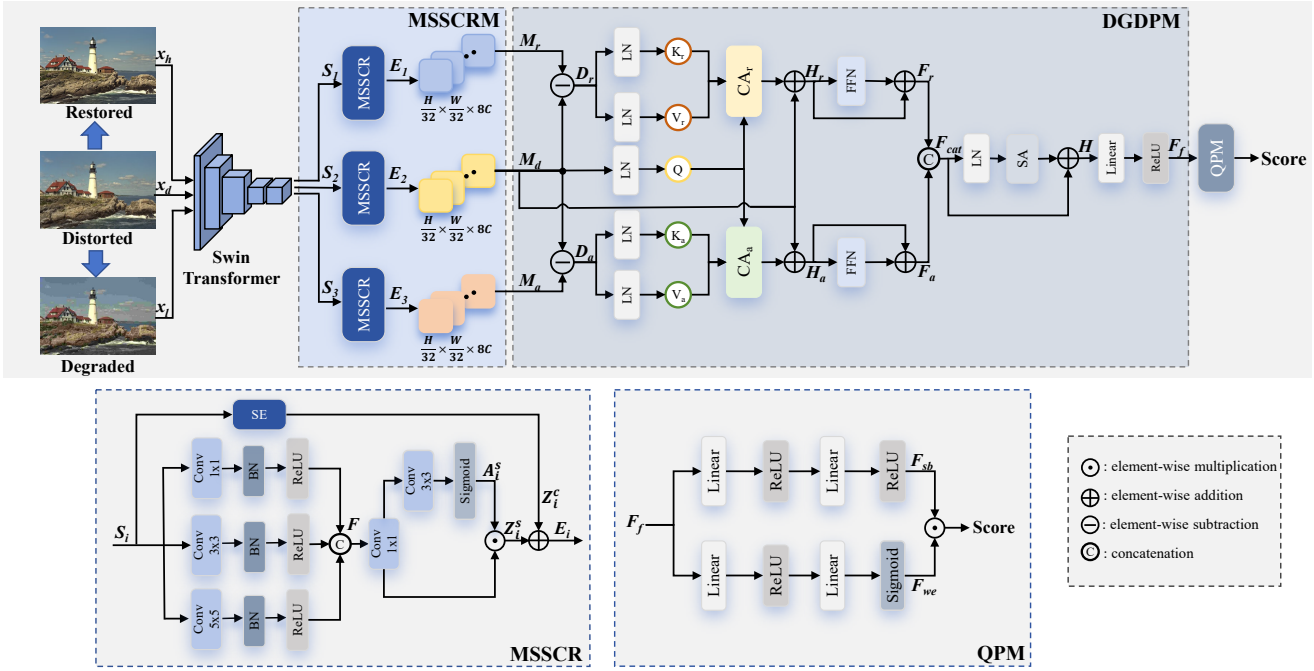

Fig. 1. Overall framework of the TRQDNet.

reference images to form a triplet. Then, we explore how to extract distortion-related features from the triplet. Finally, we perform cross-attention across the triplet features to obtain fused features and regress the fused features to the quality score. The main contributions of this work are as follows:

- We propose a Multi-Scale Spatial-Channel Refinement Module. It enhances features in both spatial and channel dimensions. In this way, distortion-related features are refined and content-irrelevant information is filtered for the triplet.
- We propose a Difference-Guided Distortion Perception Module. It constructs restoration and degradation difference features and performs cross-attention across the distortion-related features. In this way, the module captures quality differences and learns difference-aware features for quality prediction.
- We propose a novel triplet-based NR-IQA method, which introduces the restored image and the degraded image as dual reference image. We conduct extensive experiments on six IQA datasets. The results indicate that the TRQDNet exhibits strong capability in modeling perceptual quality and achieves competitive performance across diverse scenarios.

II. Proposed Method

As shown in Fig. 1, the input of the TRQDNet is a triplet which includes a distored image, a high-quality restored image and a low-quality degraded image. The output of the TRQDNet is the quality score. Specifically, our method consists of three modules: Multi-Scale Spatial-Channel Refinement Module (MSSCRM), Difference-Guided Distortion Perception Module (DGDPM), and Quality Prediction Module (QPM). Firstly, we feed the triplet into the pre-trained Swin Transformer (SwinT) and extract features from the final stage. And we input the features into the MSSCRM. Then the MSSCRM exploits spatial and channel attentions to refine these features, highlighting distortion-related features and reducing quality-irrelevant information. After that, the DGDPM employs cross-attention to model quality differences and distills quality-sensitive discrepancy features into a unified representation. Finally, the QPM aggregates the fused features and outputs the final quality score.

A. Dual Reference Image

1) High-Quality Reference Image: Recent blind image restoration methods [20], [21] based on diffusion priors have shown strong ability to recover high-quality images from severely degraded inputs. Inspired by this, we adopt the pre-trained model DiffBIR [22] to construct the high-quality reference image in our method. Given a distorted image $x_d \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the height and width of the input image. Then we feed it into the model to obtain a restored image $x_h \in \mathbb{R}^{H\times W\times 3}$. This model removes unknown degradations (e.g., blur, noise, compression artifacts) and restores realistic details with a diffusion-based generative prior. The restored image x_h is perceptually closer to the clean scene, as shown in Fig. 2. It serves as the high-quality image in our triplet.

$$x_h = \mathrm{DiffBIR}(x_d) \tag{1}$$

2) Low-Quality Reference Image: Min et al. [23] have already demonstrated that the effectiveness of degraded reference images for quality prediction. Inspired by this, we adopt

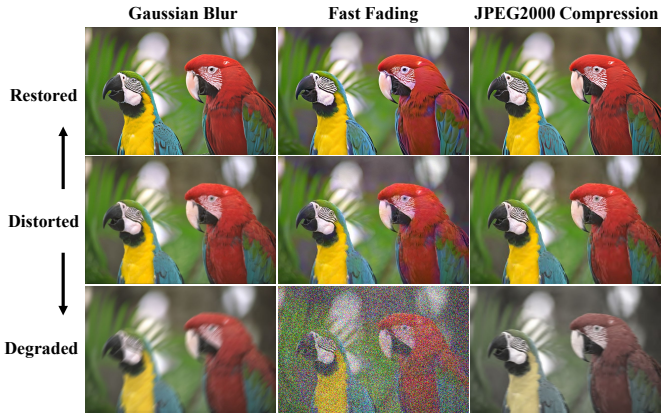

Fig. 2. Examples of triplet under different distortions.

a low-quality reference image generation strategy by further degrading the input image x_d, as shown in Fig. 2. Specifically, any distortion can be represented as:

$$x_\ell = D_i(x_d; w) \tag{2}$$

where D_i denotes the distortion function, w indicates the distortion parameters, and $x_\ell \in \mathbb{R}^{H\times W\times 3}$ represents the low-quality reference image. A composite distortion strategy can be described as:

$$x_\ell = D_0\big(D_1(\cdots D_M(x_d; w_M)\cdots; w_1); w_0\big) \tag{3}$$

where D represents the overall degradation process and M denotes the number of distortion types.

During implementation, the distortion strategy includes JPEG2000 Compression, Gaussian Blur, Gaussian Noise, etc. For each selected distortion, its severity level is randomly sampled from a given range.

B. Multi-Scale Spatial-Channel Refinement Module

The features from the pre-trained SwinT contain rich semantic information, but also include redundant content and are not sufficiently sensitive to distortions. Therefore, we design the MSSCRM to refine these features from the triplet, as shown in Fig. 1.

Specifically, the MSSCRM consists of three components named MSSCR. Suppose that the input of the i-th MSSCR is $S_i \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 8C}$ $(i = 1, 2, 3)$, C is the base channel dimension. In each component, we apply three parallel convolutions with different kernel sizes to the input feature S_i. This allows us to capture spatial information at different receptive-field scales. Then, the three outputs are concatenated along the channel dimension and fused by a 1×1 convolution to obtain a feature map $F \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 8C}$. Subsequently, we feed F into a convolution layer and a Sigmoid function to generate a spatial attention map $A_i^s \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 8C}(i = 1, 2, 3)$, which is then applied to F via element-wise multiplication to obtain the spatially refined feature $Z_i^s \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 8C}(i = 1, 2, 3)$. Meanwhile, we apply the channel attention block to the original input features S_i. This block reweights the channels and produces channel-weighted features $Z_i^c \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 8C}(i = 1, 2, 3)$. Finally, the spatially refined feature Z_i^s and the channel-weighted feature Z_i^c are fused through a residual connection to produce the output feature $E_i \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 8C}$ $(i = 1, 2, 3)$. The process of the i-th component is as follows:

$$F_i^1 = \mathrm{ReLU}\big(\mathrm{BN}(\mathrm{Conv1x1}(S_i))\big) \tag{4}$$

$$F_i^2 = \mathrm{ReLU}\big(\mathrm{BN}(\mathrm{Conv3x3}(S_i))\big) \tag{5}$$

$$F_i^3 = \mathrm{ReLU}\big(\mathrm{BN}(\mathrm{Conv5x5}(S_i))\big) \tag{6}$$

$$F = \mathrm{Conv1x1}\big(F_i^1 \otimes F_i^2 \otimes F_i^3\big) \tag{7}$$

$$A_i^s = \mathrm{Sigmoid}\big(\mathrm{Conv3x3}(F)\big) \tag{8}$$

$$Z_i^s = A_i^s \odot F \tag{9}$$

$$Z_i^c = \mathrm{SE}(S_i) \tag{10}$$

$$E_i = Z_i^s \oplus Z_i^c \tag{11}$$

where $\mathrm{BN}(\cdot)$ denotes the Batch Normalization (BN), $\mathrm{Conv1x1}(\cdot)$ indicates 1×1 convolution with padding = 0, $\mathrm{Conv3x3}(\cdot)$ represents 3×3 convolution with padding = 1, $\mathrm{Conv5x5}(\cdot)$ denotes 5×5 convolution with padding = 2, $\otimes$ represents the concatenation operation, $\odot$ denotes element-wise multiplication, $\oplus$ indicates residual addition, $\mathrm{ReLU}(\cdot)$ denotes ReLU function; $\mathrm{Sigmoid}(\cdot)$ represents the Sigmoid function, $\mathrm{SE}(\cdot)$ denotes the Squeeze-and-Excitation (SE) [24] block.

C. Difference-Guided Distortion Perception Module

To exploit the quality differences within the triplet, we design the DGDPM to learn difference-aware features for quality prediction. The module consists of two parallel difference-guided cross-attention components and a fusion layer.

Suppose that the input of the DGDPM are distortion-related features E_i. They are reshaped along the spatial dimensions into the token sequence $M_i \in \mathbb{R}^{N\times C}$ $(i = r, d, a)$, where $N = \frac{H}{32}\times\frac{W}{32}$. Based on these sequences, we first compute two difference features: the restoration difference $D_r \in \mathbb{R}^{N\times C}$ and the degradation difference $D_a \in \mathbb{R}^{N\times C}$.

$$D_r = M_r - M_d \tag{12}$$

$$D_a = M_a - M_d \tag{13}$$

Next, in order to fuse the restoration and degradation differences efficiently, we perform cross-attention fusion in two directions, namely the restoration direction and the degradation direction.

For the restoration direction, the distorted feature M_d and the restoration difference D_r are taken as inputs. We first apply LayerNorm (LN) to M_d and D_r to obtain normalized features. Next, we use linear projections on the normalized M_d to obtain the query Q, and use linear projections on the normalized D_r to obtain the key and value (K_r, V_r). Cross-attention (CA_r) is computed between Q and (K_r, V_r), and the attention output

is further processed by a feed-forward network (FFN) with a residual connection, yielding the restoration-direction feature $F_r \in \mathbb{R}^{N\times C}$. The specific equation is as follows:

$$Q = \text{Linear}(\text{LN}(M_d)) \tag{14}$$

$$(K_r, V_r) = \text{Linear}(\text{LN}(D_r)) \tag{15}$$

$$\text{CA}_\text{r}(\text{Q}, \text{K}_\text{r}, \text{V}_\text{r}) = \text{softmax}\left(\frac{QK_r^\text{T}}{\sqrt{d}}\right) V_r \tag{16}$$

$$H_r = \text{CA}_\text{r}(Q, K_r, V_r) \oplus M_d \tag{17}$$

$$F_r = \text{FFN}(H_r) \oplus H_r \tag{18}$$

where d is the channel number of K_r.

For the degradation direction, the distorted feature M_d and the degradation difference $D_a \in \mathbb{R}^{N\times C}$ are used in the same manner. We also apply LN to M_d and D_a. Then, linear projections are applied to the normalized M_d and D_a to generate the query Q, the key and value (K_a, V_a). Cross-attention (CA_a) is computed between Q and (K_a, V_a), and the result is passed through a FFN with a residual connection to obtain the degradation-direction feature $F_a \in \mathbb{R}^{N\times C}$. The specific equation is shown below:

$$(K_a, V_a) = \text{Linear}(\text{LN}(D_a)) \tag{19}$$

$$\text{CA}_\text{a}(\text{Q}, \text{K}_\text{a}, \text{V}_\text{a}) = \text{softmax}\left(\frac{QK_a^\text{T}}{\sqrt{d}}\right) V_a \tag{20}$$

$$H_a = \text{CA}_\text{a}(Q, K_a, V_a) \oplus M_d \tag{21}$$

$$F_a = \text{FFN}(H_a) \oplus H_a \tag{22}$$

Then, we concatenate F_r and F_a along the channel dimension to form $F_\text{cat} \in \mathbb{R}^{N\times 2C}$, and apply LN and a Multi-Head Self-Attention (MHSA) [25] to F_cat for further interaction between the two directions. Finally, a linear layer with ReLU activation projects the fused feature to the target dimension, yielding the final difference-aware representation $F_f \in \mathbb{R}^{N\times C}$. The specific equation is as follows:

$$F_\text{cat} = F_r \otimes F_a \tag{23}$$

$$H = \text{MHSA}(\text{LN}(F_\text{cat})) \oplus F_\text{cat} \tag{24}$$

$$F_f = \text{ReLU}(\text{Linear}(H)) \tag{25}$$

D. Quality Prediction Module

Different locations of the feature map correspond to different weights and have different effects on the quality score prediction. We utilize a two-branch structure for quality score prediction, as shown in Fig. 1. The input feature $F_f \in \mathbb{R}^{N\times C}$ is fed into the two-branch structure to produce the feature $F_\text{sb} \in \mathbb{R}^{N\times C}$ and the weight $F_\text{wb} \in \mathbb{R}^{N\times C}$. By multiplying the feature and weight of each patch, the score $Score \in \mathbb{R}$ of the corresponding part of the distorted image is computed. The specific process is as follows:

$$F_\text{sb} = \text{ReLU}(\text{Linear}(\text{ReLU}(\text{Linear}(F_f)))) \tag{26}$$

$$F_\text{we} = \text{Sigmoid}(\text{Linear}(\text{ReLU}(\text{Linear}(F_f)))) \tag{27}$$

$$Score = F_\text{sb} \odot F_\text{we} \tag{28}$$

III. Experiments

A. Experimental Setups

1) Datasets and Evaluation Criteria: We conduct extensive experiments on six public datasets including four synthetic datasets and two authentic datasets, such as LIVE [26], CSIQ [27], TID2013 [28], KADID10K [29], LIVEC [30] and KONIQ10K [31]. We test the effectiveness of the method using Spearman's Rank Correlation Coefficient (SRCC) and Pearson's Linear Correlation Coefficient (PLCC). Their range is between -1 and 1, and the higher the absolute value, the more effective the model.

2) Implementation Details: Our experiment is implemented on NVIDIA TESLA V100S with PyTorch 1.8.0 and CUDA 13.0 for training and testing. Since the input size of the SwinT does not match the used dataset, we randomly crop the input image with a size of 224×224 from the dataset, and then randomly turn the cropped image horizontally with a probability of 0.5 to augment the image. Each dataset is randomly divided into an 8:2 ratio, 80% of the images are used for training and the 20% for testing, and select 5 different seeds. During training, the batch size is fixed at 8, and the learning rate is set to 1×10^{-5}. We use the Adam optimizer with a weight decay of 1×10^{-5}. The base channel dimension C is 128. The mean squared error (MSE) loss is utilized for training, and the method is trained for 150 epochs. The number of distortion types M is set to 6 in the distortion strategy. In terms of model complexity, our method costs 54.60 GFLOPs and contains 190.2 M learnable parameters.

B. Evaluation on the Same Dataset

On CSIQ, LIVE, TID2013, KADID10K, LIVEC and KONIQ10K datasets, we perform experiment comparisons with other 16 well-known methods. The comparison results are given in Table I. The results of the compared NR-IQA methods are taken from their original papers. Notably, Our SRCC and PLCC values are greater than those of all compared methods on LIVE, CSIQ, TID2013, LIVEC and KONIQ10K datasets. On the KONIQ10K dataset, the SRCC of LoDa is the biggest value, our SRCC is the second biggest value and their difference is only 0.003. In conclusion, the TRQDNet performs better than the compared methods in terms of prediction accuracy.

C. Evaluation on the Cross-Dataset

To test the generalization ability of our method, we compare it with other NR-IQA methods under the cross-dataset. In this experiment, training is performed on one specific dataset, and testing is performed on another dataset without any finetuning or parameter adjustment. The quantitative results are summarized in Table II. Notably, our method achieves the highest SRCC values. It demonstrates that the TRQDNet is better than some baseline methods in terms of generalization ability.

TABLE I
PERFORMANCE COMPARISON ON THE SAME DATASETS (BEST RESULTS ARE IN BOLD)

Method	Source	LIVE		CSIQ		TID2013		KADID10K		LIVEC		KONIQ10K	
		PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
BRISQUE [32]	TIP 2012	0.944	0.929	0.748	0.812	0.571	0.626	0.567	0.528	0.629	0.629	0.685	0.681
HOSA [33]	TIP 2016	0.949	0.948	0.841	0.781	0.764	0.688	0.693	0.659	0.640	0.678	0.684	0.703
BIECON [34]	JSTSP 2017	0.961	0.958	0.823	0.815	0.762	0.717	0.648	0.623	0.613	0.613	0.537	0.523
BMPRI [23]	TBC 2018	0.9329	0.9310	0.9339	0.9085	0.9466	0.9287	-	-	-	-	-	-
MEON [35]	TIP 2018	0.955	0.951	0.864	0.852	0.824	0.808	0.691	0.604	0.710	0.697	0.628	0.611
DBCNN [13]	TCSVT 2020	0.971	0.968	0.959	0.946	0.865	0.816	0.856	0.851	0.869	0.851	0.884	0.875
HyperIQA [36]	CVPR 2020	0.966	0.962	0.942	0.923	0.858	0.840	0.845	0.852	0.882	0.859	0.917	0.906
MUSIQ [15]	ICCV 2021	0.940	0.911	0.893	0.871	0.815	0.773	0.872	0.875	0.746	0.702	0.928	0.916
TReS [16]	WACV 2022	0.968	0.969	0.942	0.922	0.883	0.863	0.858	0.859	0.877	0.846	0.928	0.915
VCRNet [37]	TIP 2022	0.974	0.973	0.955	0.943	0.875	0.846	0.706	0.697	0.865	0.856	0.909	0.894
Re-IQA [38]	CVPR 2023	0.971	0.970	0.960	0.947	0.861	0.804	0.885	0.872	0.854	0.840	0.923	0.914
DEIQT [39]	AAAI 2023	0.982	0.980	0.963	0.946	0.908	0.892	0.887	0.889	0.894	0.875	0.934	0.921
TempQT [40]	TMM 2024	0.977	0.976	0.960	0.950	0.906	0.883	-	-	0.886	0.870	0.920	0.903
LoDa [18]	CVPR 2024	0.979	0.975	-	-	0.901	0.869	0.936	0.931	0.899	0.876	0.944	**0.932**
FsPN [14]	SPL 2025	0.979	0.979	0.962	0.955	0.897	0.912	0.909	0.911	0.884	0.860	0.932	0.918
AKD-IQA [41]	TBC 2025	0.968	0.968	0.959	0.952	0.944	0.943	-	-	0.881	0.850	0.904	0.896
TRQDNet	Proposed	**0.986**	**0.982**	**0.985**	**0.981**	**0.967**	**0.962**	**0.958**	**0.953**	**0.905**	**0.890**	**0.944**	0.929

TABLE II
PERFORMANCE COMPARISON ON CROSS-DATASETS IN TERMS OF SRCC

Training	KADID10K	TID2013		KONIQ10K	LIVEC
Testing	CSIQ	CSIQ	KADID10K	LIVEC	KONIQ10K
HyperIQA [36]	0.782	0.667	0.516	0.743	0.690
VCRNet [37]	0.530	0.350	0.402	0.660	0.579
DBCNN [13]	0.766	0.689	0.509	0.736	0.687
TReS [16]	0.799	0.703	0.504	0.777	0.724
DEIQT [39]	0.810	0.559	0.369	0.733	0.623
TRQDNet	**0.813**	**0.753**	**0.526**	**0.862**	**0.737**

TABLE III
THE SRCC RESULTS OF ABLATION STUDIES ON THE LIVEC DATASET

Dist.	Restored.	Degraded.	MSSCRM	DGDPM	SRCC
✓					0.833
✓	✓				0.824
✓		✓			0.816
✓	✓	✓			0.833
✓	✓	✓	✓		0.848
✓	✓	✓		✓	0.867
✓	✓	✓	✓	✓	**0.890**

TABLE IV
THE PLCC RESULTS OF ABLATION STUDIES ON THE LIVEC DATASET

Dist.	Restored.	Degraded.	MSSCRM	DGDPM	PLCC
✓					0.859
✓	✓				0.864
✓		✓			0.857
✓	✓	✓			0.872
✓	✓	✓	✓		0.881
✓	✓	✓		✓	0.894
✓	✓	✓	✓	✓	**0.905**

D. Ablation Study

In this section, we conduct ablation studies on the LIVEC dataset to validate the effectiveness of our design. The results in terms of SRCC and PLCC are reported in Table III and Table IV, respectively. 'Dist.' denotes the distorted image in the triplet, 'Restored.' represents the restored image in the triplet, and 'Degraded.' denotes the degraded image in the triplet. Only using the distorted image to regress the quality score provides a strong baseline. However, simply concatenating restored and degraded features with distorted features does not lead to consistent improvements. This is mainly because simple concatenation changes the feature distribution of the distorted image, which can negatively affect performance. We further add two important modules, MSSCRM and DGDPM. Each module individually brings a clear improvement in both SRCC and PLCC. When the two modules are used together, the best results are obtained. These results demonstrate the effectiveness of the TRQDNet.

IV. CONCLUSION

We have proposed a novel NR-IQA method called TRQDNet, which introduces dual reference images to form a triplet. An important contribution is the MSSCRM, which refines the features from the triplet by suppressing content-irrelevant information and enhancing distortion-related representations. Another important contribution is the DGDPM, which constructs restoration and degradation difference features and performs cross-attention across the triplet features, capturing quality differences and learning difference-aware representations for quality prediction. Extensive experimental results on six public IQA datasets have demonstrated that the TRQDNet achieves competitive performance and exhibits strong generalization ability across diverse distortion scenarios.

REFERENCES

[1] X. Liang, Z. Tang, X. Zhang, X. Zhang, and C.-N. Yang, "Robust image hashing with weighted saliency map and laplacian eigenmaps," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 665–676, 2025.

[2] Y. Chen, L. Chen, X. Liang, H. Tang, and Z. Tang, "Unifying statistical and refined semantic features for lightweight no-reference image quality assessment," *IEEE Internet of Things Journal*, vol. 12, no. 15, pp. 31 536–31 547, 2025.

[3] H. Tang, Y. Chen, H. Zhang, L. Chen, R. Sun, and Z. Tang, "Integrating subjective and objective features for image aesthetics assessment," in *Proceedings of the International Joint Conference on Neural Networks*, 2024, pp. 1–8.

[4] H. Li, H. Liu, M. Liu, Y. Xiao, P. Li, and G. Zan, "U-Net-Like Spiking Neural Networks for Single Image Dehazing," in *Proceedings of the International Joint Conference on Neural Networks*, 2025, pp. 1–9.

[5] B. Qu, H. Li, and W. Gao, "Bringing Textual Prompt to AI-Generated Image Quality Assessment," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2024, pp. 1–6.

[6] Y. Cao, X. Min, W. Sun, and G. Zhai, "Attention-guided neural networks for full-reference and no-reference audio-visual quality assessment," *IEEE Transactions on Image Processing*, vol. 32, pp. 1882–1896, 2023.

[7] M. Yu, Z. Tang, X. Zhang, B. Zhong, and X. Zhang, "Perceptual hashing with complementary color wavelet transform and compressed sensing for reduced-reference image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7559–7574, 2022.

[8] L. Tang, L. Yuan, G. Zheng, Z. Wang, and G. Zhai, "DTSN: No-Reference Image Quality Assessment via Deformable Transformer and Semantic Network," in *Proceedings of the IEEE International Conference on Image Processing*, 2024, pp. 1207–1211.

[9] W. Zhou, J. Xu, Q. Jiang, and Z. Chen, "No-reference quality assessment for 360-degree images by analysis of multifrequency information and local-global naturalness," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1778–1791, 2022.

[10] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional Neural Networks for No-Reference Image Quality Assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[11] L.-M. Po, M. Liu, W. Y. F. Yuen, Y. Li, X. Xu, C. Zhou, P. H. W. Wong, K. W. Lau, and H.-T. Luk, "A novel patch variance biased convolutional neural network for no-reference image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1223–1229, 2019.

[12] Q. Yan, D. Gong, and Y. Zhang, "Two-stream convolutional networks for blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2200–2211, 2019.

[13] 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.

[14] L. Tang, Y. Han, L. Yuan, and G. Zhai, "Fspn: Blind image quality assessment based on feature-selected pyramid network," *IEEE Signal Processing Letters*, vol. 32, pp. 1–5, 2025.

[15] 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. 5128–5137.

[16] 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. 3989–3999.

[17] G. Qin, R. Hu, Y. Liu, X. Zheng, H. Liu, X. Li, and Y. Zhang, "Data-efficient image quality assessment with attention-panel decoder," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 2091–2100.

[18] K. Xu, L. Liao, J. Xiao, C. Chen, H. Wu, Q. Yan, and W. Lin, "Boosting image quality assessment through efficient transformer adaptation with local feature enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2662–2672.

[19] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "Tid2008 – a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, pp. 30–45, 2009.

[20] J. Wei, G. Yang, Z. Wang, S. Tao, A. Liu, and X. Chen, "Degradation-aware prompted transformer for unified medical image restoration," *IEEE Transactions on Image Processing*, vol. 34, pp. 8583–8598, 2025.

[21] K. Zhu, H. Liu, and L. Yue, "Underwater Image Enhancement and Restoration Method Based on Neural Network," in *Proceedings of the International Conference on Image Processing, Computer Vision and Machine Learning*, 2024, pp. 626–629.

[22] X. Lin, J. He, Z. Chen, Z. Lyu, B. Dai, F. Yu, Y. Qiao, W. Ouyang, and C. Dong, "DiffBIR: Toward Blind Image Restoration with Generative Diffusion Prior," in *Proceedings of the European Conference on Computer Vision*, 2024, pp. 430–448.

[23] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 508–517, 2018.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[25] T. Yao, Y. Li, Y. Pan, Y. Wang, X.-P. Zhang, and T. Mei, "Dual vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 870–10 882, 2023.

[26] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[27] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, 2010.

[28] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo, "Image database tid2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

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

[30] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016.

[31] 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, p. 4041–4056, 2020.

[32] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[33] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016.

[34] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 206–220, 2017.

[35] J. Wu, J. Ma, F. Liang, W. Dong, G. Shi, and W. Lin, "End-to-end blind image quality prediction with cascaded deep neural network," *IEEE Transactions on Image Processing*, vol. 29, pp. 7414–7426, 2020.

[36] 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. 3664–3673.

[37] Z. Pan, F. Yuan, J. Lei, Y. Fang, X. Shao, and S. Kwong, "Vcrnet: Visual compensation restoration network for no-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 31, pp. 1613–1627, 2022.

[38] A. Saha, S. Mishra, and A. C. Bovik, "Re-iqa: Unsupervised learning for image quality assessment in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5846–5855.

[39] G. Qin, R. Hu, Y. Liu, X. Zheng, H. Liu, X. Li, and Y. Zhang, "Data-efficient image quality assessment with attention-panel decoder," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 2091–2100.

[40] J. Shi, P. Gao, and A. Smolic, "Blind image quality assessment via transformer predicted error map and perceptual quality token," *IEEE Transactions on Multimedia*, vol. 26, pp. 4641–4651, 2024.

[41] B. Hu, W. Chen, J. Zheng, L. Li, W. Lu, and X. Gao, "No-reference image quality assessment via inter-level adaptive knowledge distillation," *IEEE Transactions on Broadcasting*, vol. 71, no. 2, pp. 581–592, 2025.