No-reference Image Quality Assessment via Non-local Dependency Modeling

Shuyue Jia
Dept. of Computer Science
City University of Hong Kong
Hong Kong, China
shuyuejia3-c@my.cityu.edu.hk

Baoliang Chen
Dept. of Computer Science
City University of Hong Kong
Hong Kong, China
blchen6-c@my.cityu.edu.hk

Dingquan Li
Peng Cheng Laboratory
Shen Zhen, China
lidq01@pcl.ac.cn

Shiqi Wang
Dept. of Computer Science
City University of Hong Kong
Hong Kong, China
shiqwang@cityu.edu.hk

***Abstract*—In this paper, we propose a no-reference image quality assessment method based on non-local features learned by a graph neural network (GNN). The proposed quality assessment framework is rooted in the view that the human visual system perceives image quality with long-dependency constructed among different regions, inspiring us to explore the non-local interactions in quality prediction. Instead of relying on convolutional neural network (CNN) based quality assessment methods that primarily focus on local field features, the GNN aiming for non-local quality perception facilitates modeling such long-dependency. In particular, we first adopt superpixel segmentation for the graph nodes construction. Subsequently, a spatial attention module is proposed to integrate the long- and short-range dependencies among the nodes of the whole image. The learned non-local features are finally combined with the local features extracted by the pre-trained CNN, achieving superior performance to the features utilized individually. Experimental results on intra-dataset and cross-dataset settings verify our proposed method's effectiveness and advanced generalization capability. Source codes are publicly accessible at https://github.com/SuperBruceJia/NLNet-IQA for scientific reproducible research.**

***Index Terms*—No-reference image quality assessment, human visual system, non-local modeling, superpixel, graph neural network.**

I. Introduction

THE image quality assessment (IQA) model objectively measures the input image quality, playing an essential role in various computer vision tasks, *e.g.*, image compression, enhancement, and editing [1], [2]. Compared with the full-reference (FR) IQA where the reference image should be available, the no-reference (NR) IQA is much more practical. However, the absence of reference information brings great challenges to the NR-IQA. Recently, different NR-IQA models have been proposed in the literature [3]–[19].

Early NR-IQA methods mainly explored the shared statistical behaviors of natural images, where the quality of the distorted image can be estimated by evaluating the destruction of natural scene statistics (NSS). The NSS can be constructed in different domains, including the spatial domain [3], discrete cosine transform (DCT) domain [4], and wavelet domain [5]. In [6], [7], the codebook was learned for the quality-aware features extraction, getting rid of the handcrafted feature design. The free-energy principle proposed in brain theory and neuroscience [20] reveals that the image distortion can be measured by the discrepancy between the image and its brain-predicted version. Inspired by such a theory, Zhai *et al*. built an internal description of images via a generative model [8]. Chen *et al*. restored the distorted image by the generative adversarial network (GAN), and an attention-driven approach was introduced to estimate the visual quality [9].

In the deep-learning era, the convolutional neural network (CNN) has demonstrated superior prediction performance for IQA [10]–[15]. To begin with, deep features from CNN were derived for visual quality estimation. Kang *et al*. [10] proposed a shallow CNN to extract the learned quality-aware features. In particular, distortion type identification is also involved, casting the method into a multi-task learning paradigm. Zhang *et al*. [11] introduced a bilinear pooling layer to fuse the distortion type and object semantic information. Su *et al*. employed a self-adaptive hyper network to learn the perceptual rules and content [12]. Wu *et al*. proposed a cascaded CNN model motivated by the hierarchical degradation process of the human visual system (HVS) [13]. Besides, to train a more robust CNN by increasing the amount of data, the rank-based methods [16]–[19] attract much attention as the training samples can be significantly enriched in a paired manner. In [16], Ma *et al*. adopted the discriminative image pairs for quality ranking learning. The Siamese network was utilized in [17] to process a ranked image pair. The learned network was further finetuned for quality regression. Built upon CNN, recently, the graph learning and non-local feature extraction methods were also presented for NR-IQA [14], [15]. Sun *et al*. [14] introduced graph representation learning to model the relationship of different distortions. However, such a model was not specifically designed to extract the non-local features. Golestaneh *et al*. proposed to construct the non-local representation via a Transformer architecture [15], revealing that the non-local information plays an essential role for IQA. Nevertheless, there are still improving spaces for their method since the non-local information was only extracted from the high-level semantic feature maps.

The HVS perceives the image quality by capturing local distortions and aggregating non-local dependencies [15], [21]. In

This work is supported by the Shenzhen Virtural University Park, The Science Technology and Innovation Committee of Shenzhen Municipality (Project No: 2021Szvup128). Corresponding Author: Dr. Shiqi Wang (email: shiqwang@cityu.edu.hk).

978-1-6654-7189-3/22/$31.00 © 2022 IEEE

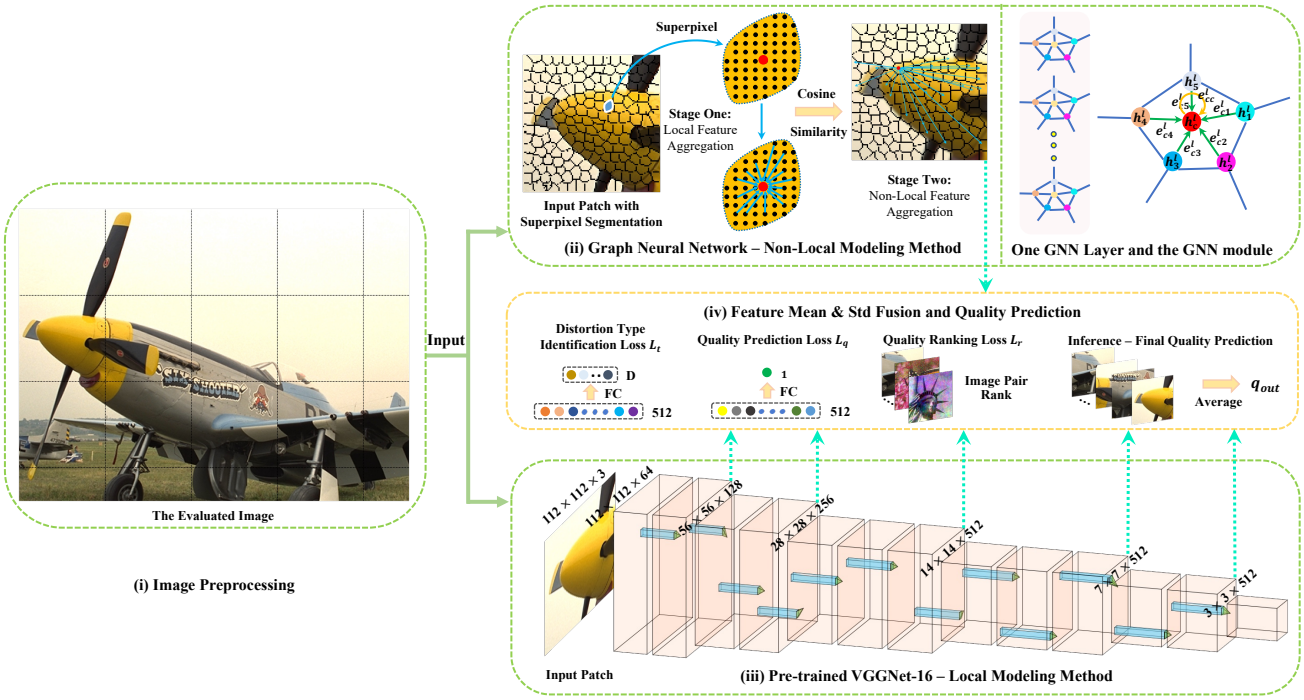

Fig. 1. The overview of the proposed NLNet. (i) The input image is pre-processed. (ii) A two-stage GNN approach is presented for the non-local feature extraction and long-range dependency construction among different regions. The first stage aggregates local features inside superpixels. The following stage learns the non-local features and long-range dependencies among the graph nodes. It then integrates short- and long-range information based on an attention mechanism. The means and standard deviations of the non-local features are obtained from the graph feature signals. (iii) Local feature means and standard deviations are derived from the pre-trained VGGNet-16 considering the hierarchical degradation process of the HVS. (iv) The means and standard deviations of the local and non-local features are fused to deliver a robust and comprehensive representation for quality assessment. Besides, the distortion type identification loss L_t, quality prediction loss L_q, and quality ranking loss L_r are utilized for training the NLNet. During inference, the final quality of the image is the averaged quality of all the non-overlapping patches.

particular, the strong dependencies among neighborhood pixels carry essential information of the structure of objects, which is sensitively perceptive by HVS for quality evaluation [22]. On the other hand, the non-local features have also been revealed to play complementary roles in human quality rating [23]. Following this vein, for the exploration of local quality clues, a pre-trained VGGNet [24] is adopted, whose effectiveness has been validated in acquiring the highly quality-aware features and is widely used for both FR-IQA tasks [1] and NR-IQA tasks [17]. Nevertheless, there is a strong inductive bias in CNN, *i.e.*, locality. First, the CNN filters extract features mainly from the local neighborhoods. Thus, it is hard to catch the non-local features and long-range dependencies among pixels and regions from the image [25]. Moreover, since the content in an image is multi-scale and space-variant [22], the CNN filters equally process it which should be treated distinguishingly. Furthermore, CNN fails to model the geometric and relational causal dependencies [26]. As such, suffered by the local priors to CNN, the non-local information is usually absent. To account for this, we further design a superpixel-based graph neural network (GNN) approach to capture the non-local features. Finally, the local and non-local features are fused for image quality regression. Experimental results have demonstrated the complementary role of the local and non-local features, and superior performances can be achieved by combining the two types of features. The main contributions of this paper are summarized as follows,

- We propose a novel NR-IQA framework based on the GNN. The non-local behavior of natural images is emphasized and learned in our proposed **N**on-**L**ocal dependency **Net**work (termed as NLNet).
- The spatial attention module is introduced to integrate the information of long- and short-range communications among graph nodes of the entire image.
- Extensive experimental results reveal that the proposed NLNet manages to extract the non-local information for quality prediction, and the superior performance in cross-dataset settings verifies the high generalization capability of our proposed method.

II. Proposed NR-IQA Method

This work presents a hybrid of local and non-local feature extractions to assess the image quality. We propose a superpixel-based GNN to learn the non-local features and capture long-range dependencies. Meanwhile, a pre-trained VGGNet-16 is employed to extract local features from spatially-proximate neighborhoods. Finally, the non-local features from GNN and local features from CNN are fused to predict the quality. In the following paragraphs, we first introduce the superpixel segmentation. Then, the construction

of GNN is provided, and finally, the local and non-local feature fusion is elaborated.

A. Superpixel Segmentation

A superpixel is a group of pixels with similar visual characteristics, such as color, intensity, and spatial adjacency. Superpixels can be represented as graphs and processed via graph learning methods, *e.g.*, GNN [27]. Compared with features derived from the standard pixel grids, superpixels can be adaptive to regional content and generate more meaningful representations. Thus the NSS can be well-exploited, which has been validated for FR-IQA [28], [29]. Our method explores the effectiveness of feature extraction from superpixels for NR-IQA. In this work, we adopt the Simple Linear Iterative Clustering (SLIC) algorithm to generate superpixels on account of its efficient computation and exceptional adherence to the boundaries of objects [28], [29].

B. Non-local Modeling via GNN

We propose to extract the non-local features by GNN in a two-stage manner. In the first stage, a GNN layer is constructed to aggregate features within superpixels. In the second stage, the learned spatial features are integrated with a multi-head self-attention. We elaborate on the two stages as follows.

a) Graph Construction: Supposing N superpixels are constructed by the segmentation, we construct an undirected and weighted graph denoted as $\mathbf{G}$, and the superpixels are treated as nodes. Thus, $\mathbf{G} = \{\mathbf{V}, \mathbf{E}, \mathbf{A}\}$. In particular, $\mathbf{V}$ denotes nodes (vertices), and $|\mathbf{V}| = N$. $\mathbf{E}$ denotes edges (links) that connect nodes $\mathbf{V}$. The adjacency matrix $\mathbf{A}$ contains weights (correlations) between nodes $\mathbf{V}$.

b) Self-Attention for Nodes Integration: To capture the long and short-range dependencies between different nodes, the self-attention mechanism is introduced in GNN, which is defined as follows,

$$\mathbf{h}_i^{l+1} = \mathrm{ELU}\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^l \mathbf{W}^l \mathbf{h}_j^l\right). \tag{1}$$

α_{ij}^l denotes the normalized attentional weight between node j and node i in the l^{th} layer. Herein, $\mathcal{N}(i)$ means the neighborhood of the i^{th} node, and $\mathbf{h}_j^l$ are the features of the j^{th} node in the l^{th} layer. $\mathbf{W}^l$ is a trainable matrix, and we consider the Exponential Linear Unit (ELU) [30] as the activation function. In Eqn. (1), α_{ij}^l can be computed as follows,

$$\alpha_{ij}^l = \frac{\exp\left(a_{ij}^l\right)}{\sum\limits_{k\in\mathcal{N}(i)} \exp\left(a_{ik}^l\right)}, \tag{2}$$

$$a_{ij}^l = \mathrm{LeakyReLU}\left(\mathrm{FC}\left(\left[\mathbf{W}^l\mathbf{h}_i^l \| \mathbf{W}^l\mathbf{h}_j^l\right]\right)\right). \tag{3}$$

a_{ij}^l is an attentional coefficient representing the importance of node j to the center node i. a_{ij}^l is derived by the concatenation of the mapped features $\mathbf{W}^l\mathbf{h}_i^l$ and $\mathbf{W}^l\mathbf{h}_j^l$ with a fully-connected neural network $\mathrm{FC}(\cdot)$. The Leaky Rectified Linear Unit (Leaky ReLU) [31] is considered here as the activation function. We further adopt the multi-head self-attention mechanism for stable training [32]:

$$\mathbf{h}_i^{l+1} = \Big\|_{m=1}^{M} \mathrm{ELU}\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij}^{l,m} \mathbf{W}^{l,m} \mathbf{h}_j^l\right), \tag{4}$$

where M is the number of heads. It should be noted that although the learnable weight $\mathbf{W}^{l,m}$ is shared among nodes, the graph attention layer aggregates features of each node with distinctions via different attentional coefficients $\alpha_{ij}^{l,m}$ [32], highly improving the learning capacity of the GNN. After the aggregation, we normalize the acquired features of each node to alleviate the variance inflammation and prevent gradient vanishing [33] as follows,

$$\mathrm{NodeNorm}\left(\mathbf{x}_i\right) = \frac{x_{i,j}}{\left\{\left(\frac{1}{F-1}\sum_{j=1}^{F}(x_{i,j}-\bar{x}_i)^2\right)^{1/2}\right\}^{1/p} + C'}. \tag{5}$$

$\mathbf{x}_i \in \mathbb{R}^{1\times F}$ represents graph signals of the i^{th} node. $x_{i,j}$ denotes the j^{th} feature of the i^{th} node, and $\bar{x}_i = \sum_{j=1}^{F} x_{i,j}/F$ is the standard deviation of features in the i^{th} node. p is a constant, and C' is a small positive constant to stabilize training. Finally, the means and standard deviations of the deep visual features are obtained from the graph attention layers,

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{i,j}, \tag{6}$$

$$\sigma_j = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_{i,j}-\mu_j)^2\right)^{\frac{1}{2}}, \tag{7}$$

in which $\boldsymbol{\mu}_{\mathbf{x}} = [\mu_1, \cdots, \mu_F] \in \mathbb{R}^{1\times F}$ denote the mean intensities of each graph feature layer, and $\boldsymbol{\sigma}_{\mathbf{x}} = [\sigma_1, \cdots, \sigma_F] \in \mathbb{R}^{1\times F}$ represent the standard deviations of features. By combining the feature means $\boldsymbol{\mu}_{\mathbf{x}}$ and standard deviations $\boldsymbol{\sigma}_{\mathbf{x}}$ along the feature dimension, the non-local features $\mathbf{f_n}$ are extracted for the overall image quality estimation.

C. Local Modeling via CNN

In addition to the non-local features $\mathbf{f_n}$ learned by the GNN, we further adopt the VGGNet-16 which is pre-trained on the ImageNet [24] as the backbone for local feature extraction. In particular, we discard the fully-connected layers, and the multi-scale features at five stages are utilized. We denote the extracted local features as $\mathbf{f_l}$.

D. Feature Fusion and Objective Functions

Finally, we concatenate the non-local features $\mathbf{f_n}$ and local features $\mathbf{f_l}$ along the channel dimension to form a robust and comprehensive representation of the overall image quality. The combined features are denoted as $\mathbf{f_g} = [\mathbf{f_n} \,\|\, \mathbf{f_l}]$, and further processed by a fully-connected layer for quality prediction.

The local modeling and non-local modeling modules are jointly trained in an end-to-end manner. For the model learning, the quality prediction loss L_q, ranking loss L_r, and distortion type classification loss L_t are conducted. In detail,

we employ the Huber Loss [34] (denoted as $\mathrm{HuberLoss}$) for quality evaluation, which is less sensitive to noise than the L_2 loss [13]. The L_q, L_r, and L_t are defined as follows,

$$L_q = \frac{1}{B}\sum_{k}\mathrm{HuberLoss}(\hat{q}_k - q_k), \tag{8}$$

$$L_r = \frac{1}{B(B-1)/2}\sum_{j<k}\mathrm{HuberLoss}\left((\hat{q}_j - \hat{q}_k) - (q_j - q_k)\right), \tag{9}$$

$$L_t = -\frac{1}{B}\sum_{i=1}^{B}\sum_{d=1}^{D} p_{id}\ln\hat{p}_{id}, \tag{10}$$

where B denotes the batch size, $\hat{q}_j$ and $\hat{q}_k$ are the predicted quality scores of two different images, and q_j and q_k are their corresponding MOSs. In Eqn. (10), for the i^{th} image inside the batch, p_{id} is the label probability of the d^{th} distortion type, and $\hat{p}_{id}$ is the predicted probability of the d^{th} type. Herein, we adopt the cross-entropy loss for the distortion type classification. In particular, we map the $\mathbf{f_g}$ to a hidden representation via a fully-connected layer. Then, the hidden features are mapped to D neurons, where D denotes the number of distortion types. In summary, the overall objective function is as follows,

$$L = \theta \times L_q + L_r + L_t + \frac{\rho}{2P}\|\mathbf{W}\|^2, \tag{11}$$

where θ is a hyper-parameter that leverages the importance of quality prediction loss, $\mathbf{W}$ are the network parameters, P denotes the number of network parameters, and ρ is the weight decay rate.

III. Experimental Validations and Analysis

A. Implementation Details

In this work, we perform the superpixel segmentation on the cropped patches (112×112) with the size of superpixels set by 8×8. In the GNN construction phase, we uniformly sample 60 nodes from each superpixel to aggregate local features within superpixels. We apply the cosine similarity among the aggregated features of superpixels to measure the correlations and similarities. To build a graph for non-local feature integration, we empirically set a threshold of 0.70. There is an edge if the similarity between two superpixels is greater than the threshold. Otherwise, there is no connection. We employ the initial connection for each layer to prevent the over-smoothing problem [35]. Besides, half of the total nodes, *i.e.*, 100 nodes, are uniformly sampled to aggregate the non-local features [25].

We utilize 32 hidden neurons inside $\mathrm{FC}(\cdot)$ in Eqn. (3) to map the input normalized RGB values to a hidden representation. Besides, the number of heads M is set as 4, and the GNN layer is parallelly implemented for stable and faster training. In Eqn. (5), we set $p = 2$, and C' to 1×10^{-5}. The number of neurons inside the fully-connected layer for quality prediction and distortion type classification is 512 before the final output layer. The batch size B is 4. We train the model using the Adam optimizer [36] for 100 epochs over all the experiments with a learning rate of 1×10^{-4} which is reduced by 5 every 20 epochs. In Eqn. (11), the weight decay rate ρ is 5×10^{-4}, and θ is set as 100.

TABLE I
Brief summary of the LIVE, CSIQ, and TID2013 databases

Database	LIVE	CSIQ	TID2013
Number of Reference Images	29	30	25
Number of Images	779	866	3,000
Number of Distortion Types	5	6	24
Number of Distortion Levels	$5 \sim 8$	$4 \sim 5$	5
Annotation	DMOS	DMOS	MOS
Range	[0, 100]	[0, 1]	[0, 9]

TABLE II
Performance comparisons on the LIVE, CSIQ, and TID2013 databases

Method	LIVE		CSIQ		TID2013	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE (2012) [3]	0.939	0.935	0.746	0.829	0.604	0.694
CORNIA (2012) [6]	0.947	0.950	0.678	0.776	0.678	0.768
M3 (2015) [40]	0.951	0.950	0.795	0.839	0.689	0.771
HOSA (2016) [7]	0.946	0.947	0.741	0.823	0.735	0.815
FRIQUEE (2017) [41]	0.940	0.944	0.835	0.874	0.68	0.753
DIQaM-NR (2018) [42]	0.960	**0.972**	-	-	0.835	0.855
DB-CNN (2020) [11]	0.968	**0.971**	**0.946**	**0.959**	0.816	0.865
HyperIQA (2020) [12]	0.962	0.966	0.923	0.942	0.729	0.775
GraphIQA (2022) [14]	**0.968**	0.970	0.920	0.938	-	-
TReS (2022) [15]	**0.969**	0.968	0.922	0.942	**0.863**	**0.883**
NLNet (Proposed)	0.962	0.963	**0.941**	**0.958**	**0.856**	**0.880**

B. Evaluation Databases and Criteria

1) Evaluation Databases: The proposed NLNet is evaluated on three natural image IQA benchmarks, including the LIVE [37], CSIQ [38], and TID2013 [39] databases. In Table I, a lower Difference Mean Opinion Score (DMOS) indicates a better quality, whereas the Mean Opinion Score (MOS) is the opposite.

2) Experiments Settings: For intra-database experiments, we randomly split the reference images into 60% training, 20% validation, and 20% testing, and 10 random splits of the reference indices are performed to avoid bias. We report the median performances on the testing set. Furthermore, for the cross-database experiments, one database is used as the training set, and the other databases are the testing sets. The performance of the model in the last epoch is reported. The images are cropped to several patches with the size of 112×112 during training. All the cropped patches are utilized in the testing phase, and the final prediction results are their average. The Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank-order Correlation Coefficient (SRCC) are used to evaluate model performance.

C. Performance Evaluations on Each Individual Database

In Table II, we present the performance comparisons with several NR-IQA methods. The reported performances are derived from the corresponding papers. We report the results of the DIQaM-NR from [42] which is the no-reference DIQaM

TABLE III
CROSS-DATABASE PERFORMANCE COMPARISONS

Training	LIVE		CSIQ		TID2013	
Testing	CSIQ	TID2013	LIVE	TID2013	LIVE	CSIQ
BRISQUE (2012) [3]	0.562	0.358	0.847	0.454	0.790	0.590
CORNIA (2012) [6]	0.649	0.360	0.853	0.312	0.846	0.672
M3 (2015) [40]	0.621	0.344	0.797	0.328	0.873	0.605
HOSA (2016) [7]	0.594	0.361	0.773	0.329	0.846	0.612
FRIQUEE (2017) [41]	0.722	0.461	0.879	0.463	0.755	0.635
DIQaM-NR (2018) [42]	0.681	0.392	-	-	-	0.717
DB-CNN (2020) [11]	**0.758**	**0.524**	0.877	**0.540**	**0.891**	**0.807**
HyperIQA (2020) [12]	0.697	**0.538**	**0.905**	**0.554**	0.839	0.543
NLNet (Proposed)	**0.771**	0.497	**0.923**	0.516	**0.895**	**0.730**

TABLE IV
THE RESULTS OF ABLATION STUDY

Non-local block	L_r	L_t	PLCC ↑	Δ	SRCC ↑	Δ
✓	✓	✓	0.941	-	0.958	-
	✓	✓	0.936	-0.005	0.951	-0.007
✓			0.916	-0.025	0.938	-0.020
✓	✓		0.929	-0.012	0.945	-0.013
✓		✓	0.934	-0.007	0.947	-0.011

model. From the table, we can observe that the deep-learning based methods, such as the DB-CNN [11], HyperIQA [12], and TReS [15], usually achieve superior performances than the handcrafted feature based methods, as more quality-aware features can be learned from the data. Compared with the GraphIQA [14] where the GNN is also utilized, we achieve a significant SRCC improvement (0.941 *vs.* 0.920) on the CSIQ database. Our method presents a higher subjective opinions consistency than the latest method TReS [15] on both the LIVE and CSIQ databases. The comparable performances on the TID2013 database further verify the effectiveness of our method. The reason may lie in that the non-local features deliver a robust representation of the visual quality.

D. Cross-Database Evaluations

We further analyze the generalization capability of our model in a cross-database manner. As shown in Table III, our method achieves the best performances in terms of both SRCC and PLCC on the LIVE (Train) → CSIQ (Test), CSIQ (Train) → LIVE (Test), and TID2013 (Train) → LIVE (Test) settings. As shown in Table III, we can observe the setting that training on the CSIQ or LIVE database and testing on the TID2013 database is much more challenging, as many distortion types in the TID2013 database are unseen during training. However, our method still achieves a comparable performance. The superior performance in the cross-database settings reveals the high generalization capability of our method, making our method to be more practical in real applications.

E. Ablation Study

To verify the functionalities and effectiveness of different components in our model, we conduct an ablation study on the CSIQ database. The experimental results are shown in Table IV. As shown in the table, when we ablate the non-local block from the NLNet, a performance drop can be observed, revealing that the non-local modeling plays an essential and complementary role to the local modeling. In addition, we explore the importance of the quality ranking loss L_r or the distortion type identification loss L_t. The ablation result shown in Table IV reveals that the rank learning and distortion type identification contribute to the final performance improvement. We believe the reason may lie in that more discriminative features can be obtained when the quality ranking and distortion identification are performed. Overall, we can conclude that each component of our method provides an influential contribution to the final performance achievement.

TABLE V
VALIDATION PERFORMANCES REGARDING DIFFERENT HYPERPARAMETERS IN NLNET

Setting	Num. of heads	Num. of neurons	Num. of layers	SRCC	PLCC
1	**1**	32	3	0.936	0.953
2	**4**	32	3	0.941	0.958
3	**8**	32	3	0.938	0.949
4	4	**16**	3	0.933	0.953
5	4	**32**	3	0.941	0.958
6	4	**64**	3	0.932	0.950
7	4	32	**2**	0.932	0.949
8	4	32	**3**	0.941	0.958
9	4	32	**4**	0.936	0.952

Moreover, the multi-head self-attention module is introduced in our method. To verify its effectiveness, we further compare our method with only single-head attention utilized (Num. of heads is 1) in Table V. Again, the performance drops in terms of both PLCC and SRCC. However, the performance will not be improved when the head number increases which may be led by the over-fitting problem due to more parameters introduced. The best number of heads we finally adopt is 4. Furthermore, as shown in the settings $4 \sim 6$ and settings $7 \sim 9$ of Table V, we also explore the optimal numbers of neurons and GNN layers on the CSIQ database, and the best performance can be achieved when we set the numbers of neurons and GNN layers as 32 and 3, respectively.

IV. CONCLUSION

In this paper, we propose a non-local modeling method for NR-IQA inspired by HVS usually perceives image quality with long-range dependencies. The non-local features and long-range dependencies are learned via a two-stage superpixel-based graph neural network, which plays a complementary role with the local modeling. Experimental results on different IQA datasets demonstrate the effectiveness of the proposed method. The superior performance in the cross-dataset setting reveals the high generalization capability of our method, shedding light on the exploration of the generalized NR-IQA models.

REFERENCES

[1] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Comparison of full-reference image quality models for optimization of image processing

systems," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1258–1281, Jan. 2021.

[2] H. Li, J. Qin, Z. Yang, P. Wei, J. Pan, L. Lin, and Y. Shi, "Real-world image super-resolution by exclusionary dual-learning," *IEEE Transactions on Multimedia*, pp. 1–13, June 2022.

[3] 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, Dec. 2012.

[4] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[5] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[6] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1098–1105, June 2012.

[7] 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, Sept. 2016.

[8] G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang, "A psychovisual quality metric in free-energy principle," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 41–52, Jan. 2012.

[9] D. Chen, Y. Wang, and W. Gao, "No-reference image quality assessment: An attention driven approach," *IEEE Transactions on Image Processing*, vol. 29, pp. 6496–6506, 2020.

[10] L. Kang, P. Ye, Y. Li, and D. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *IEEE International Conference on Image Processing*, pp. 2791–2795, Sept. 2015.

[11] 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, Jan. 2020.

[12] 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 Conference on Computer Vision and Pattern Recognition*, pp. 3664–3673, June 2020.

[13] 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.

[14] S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen, "GraphIQA: Learning distortion graph representations for blind image quality assessment," *IEEE Transactions on Multimedia*, Feb. 2022.

[15] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *IEEE Winter Conference on Applications of Computer Vision*, pp. 3209–3218, Jan. 2022.

[16] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.

[17] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *IEEE International Conference on Computer Vision*, pp. 1040–1049, Oct. 2017.

[18] L. Ma, L. Xu, Y. Zhang, Y. Yan, and K. N. Ngan, "No-reference retargeted image quality assessment based on pairwise rank learning," *IEEE Transactions on Multimedia*, vol. 18, no. 11, pp. 2228–2237, Nov. 2016.

[19] L. Xu, J. Li, W. Lin, Y. Zhang, L. Ma, Y. Fang, and Y. Yan, "Multi-task rank learning for image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 9, pp. 1833–1843, Sept. 2017.

[20] K. Friston, "The free-energy principle: A unified brain theory?," *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, Feb. 2010.

[21] M. Liu, L.-M. Po, X. Xu, K. W. Cheung, Y. Zhao, K. W. Lau, and C. Zhou, "Long-range dependencies and high-order spatial pooling for deep model-based full-reference image quality assessment," *IEEE Access*, vol. 8, pp. 72007–72020, Apr. 2020.

[22] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[23] M. Zontak and M. Irani, "Internal statistics of a single natural image," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 977–984, June 2011.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[25] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, June 2018.

[26] D. She, Y.-K. Lai, G. Yi, and K. Xu, "Hierarchical layout-aware graph convolutional network for unified aesthetics assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8471–8480, June 2021.

[27] J. H. Giraldo, S. Javed, and T. Bouwmans, "Graph moving object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2485–2503, May 2022.

[28] W. Sun, Q. Liao, J.-H. Xue, and F. Zhou, "SPSIM: A superpixel-based similarity index for full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4232–4244, Sept. 2018.

[29] S. Mahmoudpour and P. Schelkens, "Synthesized view quality assessment using feature matching and superpixel difference," *IEEE Signal Processing Letters*, vol. 27, pp. 1650–1654, Sept. 2020.

[30] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," *arXiv preprint arXiv:1511.07289*, 2015.

[31] A. L. Maas, A. Y. Hannun, A. Y. Ng, *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *International Conference on Machine Learning*, vol. 30, pp. 3–8, 2013.

[32] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[33] K. Zhou, Y. Dong, K. Wang, W. S. Lee, B. Hooi, H. Xu, and J. Feng, "Understanding and resolving performance degradation in deep graph convolutional networks," in *ACM International Conference on Information & Knowledge Management*, p. 2728–2737, Association for Computing Machinery, Oct. 2021.

[34] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, Mar. 1964.

[35] M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li, "Simple and deep graph convolutional networks," in *International Conference on Machine Learning*, vol. 119, pp. 1725–1735, 2020.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[37] 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, Nov. 2006.

[38] 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, Jan. 2010.

[39] 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, Jan. 2015.

[40] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.

[41] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17, no. 1, pp. 32–32, Jan. 2017.

[42] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, Jan. 2018.