Textural-Structural Joint Learning for No-Reference Super-Resolution Image Quality Assessment

Yuqing Liu, Qi Jia, Shanshe Wang, Siwei Ma Senior Member, IEEE, Wen Gao Fellow, IEEE,

Abstract—Image super-resolution (SR) has been widely investigated in recent years. However, it is challenging to fairly estimate the performances of various SR methods, as the lack of reliable and accurate criteria for perceptual quality. Existing SR image quality assessment (IQA) metrics usually concentrate on the specific kind of degradation without distinguishing the visual sensitive areas, which have no adaptive ability to describe the diverse SR degeneration situations. In this paper, we focus on the textural and structural degradation of image SR which acts as a critical role for visual perception, and design a dual stream network to jointly explore the textural and structural information for quality prediction, dubbed TSNet. By mimicking the human vision system (HVS) that pays more attention to the significant areas of the image, we develop the spatial attention mechanism to make the visual-sensitive areas more distinguishable, which improves the prediction accuracy. Feature normalization (F-Norm) is also developed to investigate the inherent spatial correlation of SR features and boost the network representation capacity. Experimental results show the proposed TSNet predicts the visual quality more accurate than the state-of-the-art IQA methods, and demonstrates better consistency with the human’s perspective. The source code will be made available at http://github.com/yuqing-liu-dut/NRIQA_SR.

Index Terms—No-reference image quality assessment, convolutional neural network, image super-resolution, attention mechanism, feature normalization, human vision system.

I. INTRODUCTION

W ith the rapid development of high-definition display technologies, image super-resolution (SR) has been widely investigated in advanced applications, which aims to generate high-resolution (HR) images from given low-resolution (LR) instances. Although there are numerous image SR works in the past years, how to estimate the quality of super-resolved images still remains challenging.

There are special textural and structural degradation situations in SR processing, making it hard to accurately predict the perceptual quality [1]. According to the human vision system (HVS), the textural and structural degeneration is essential to the visual experience [2]. General image quality assessment (IQA) metrics are usually developed for the simulated signal degradation and random noise, which are in low correlation with the subjective perspective of SR images. In Figure 1, we can find that the SR image removes the noise and blur and gets higher PSNR and SSIM scores. However, the over-smoothed and sharpened SR image loses the textural and structural information and decreases the visual quality. In this point of view, SR IQA metrics are more suitable for visual quality prediction [5–8].

Most of the recent no-reference (NR) IQA methods for SR images usually focus on only one kind of the degradation, which limits the representation of information loss. There are different hand-crafted extractors to describe the textural features, such as the local/global frequency features [9], spatial principal component analysis (PCA) features [10], and the mean subtracted contrast normalized (MSCN) coefficients [11]. The textural features are used for evaluating the difference between the SR images and the natural images by natural scene statistics (NSS). However, it is difficult to
describe the structural degradation by the pixel-wise statistical analysis. There is also work utilizing the CNN-based feature extractor to explore the perceptual information for mean opinion score (MOS) regression [12]. The pre-trained deep CNN holds effective capacity for structural feature exploration, but lacks to describe the textural degradation by the high-level feature representation [13]. Recently, KLTSRQA [14] utilizes Karhunen-Loève Transform (KLT) to separate the structural and textural information and regresses the MOS. However, the hand-crafted extractors have no adaptive ability to learn the diverse SR features and limits the accuracy.

The key issue of NR-IQA is to build a metric that in consistence with the human vision system (HVS). According to the HVS, different areas of the images hold different importance for visual perception [15]. However, recent NR-IQA methods usually neglect to distinguish the visual sensitive information in the image, which restricts the effectiveness of prediction. The NSS-based SR metrics usually take the image as a whole without the consideration of saliency detection. Recent CNN-based NR-IQA metrics for image SR treat different areas of the image equally [16]–[18], which have no ability to highlight the visual sensitive information for MOS prediction.

In this paper, we design an end-to-end dual stream network to jointly explore the textural and structural features from the image, dubbed TSNet, for NR-IQA. One VGG-based branch is designed for perceptual structural information extraction [13]. Another CNN branch is developed to explore the pixel-wise textural information with shallow layer extractors [13]. By mimicking the HVS [15] that pays more attention to the significant information, spatial attention mechanism is introduced to make the visual sensitive areas more distinguishable. Furthermore, feature normalization (F-Norm) is also developed to investigate the inherent spatial correlation of SR features [20], [21]. Experimental results show the proposed TSNet predicts the perceptual score more accurately than state-of-the-art IQA methods, and demonstrates better consistency with the human’s perspective, as shown in Figure 1.

Our contributions can be concluded as follows:

- We design an end-to-end dual stream network named TSNet for NR-IQA on SR images, which jointly explores the textural and structural information for visual perception.
- We utilize spatial attention mechanism to emphasize the significant information, resulting the visual sensitive areas more distinguishable.
- Experimental results show the proposed TSNet is more consist with the human’s perspective than state-of-the-art IQA methods, rendering the accurate prediction of the subjective quality.

II. RELATED WORKS

A. Image Super-Resolution

Image super-resolution (SR) is a classical topic in computer vision area. The task of image SR is to generate a high-resolution (HR) image from the given low-resolution (LR) instance. Traditional image SR methods usually upscale the image by interpolation [22] or dictionary learning [23]. The generated textures vary a lot due to the different technologies and training datasets [24]. As such, it is difficult to model the degeneration of SR step by simply combining the hand-crafted signal losses. Recently, deep learning has demonstrated its amazing performance on image SR. SRGAN [28] is the first convolutional neural network (CNN) based method for image SR. After that, VDSR [26], EDSR [27], and other works achieve great success on reconstructing the HR images. These works aim to build a reliable result and choose mean average error (MAE) or mean square error (MSE) as the loss function, which product unnatural results with higher PSNR/SSIM scores and lower perceptual qualities [28]. Generative adversarial networks (GANs) have also been considered in image SR area for generating images with higher visual quality. SRGAN [28], ESRGAN [29] and other works have shown superior performances on restoring satisfying results but the objective scores are lower. Due to the diversity of SR technologies, it is challenging to find a reliable and accurate metric for estimating the restored images.

B. General Image Quality Assessment

Image quality assessment (IQA) has been widely investigated in recent years. The task of IQA is to design a metric consist with the human’s perspective and accurately estimate the image quality. General IQA aims to measure specific types of degradation or their combinations, such as ringing effect, blocking artifacts, blur and noise [30]. IQA can be generally divided into three categories: full-reference (FR) IQA, reduced-reference (RR) IQA and no-reference (NR) IQA. PSNR [3] and SSIM [4] are two most famous FR-IQA metrics that widely used in different applications. Although they can successfully describe the difference between original and distorted images, the objective metrics are not highly correlated with the subjective opinion. Based on PSNR/SSIM, there are variants to better illustrate the image quality, such as PSNR-HVS [31], MS-SSIM [32] and CW-SSIM [33]. GMSD provided a new perspective to evaluate the image quality based on the gradient magnitude similarity deviation [34]. VIF studied the human vision system (HVS) and estimated the image quality by visual information fidelity [35]. Recently there are CNN-based FR-IQA methods achieving accurate performance. PieAPP [36] developed a deep CNN-based network to assess the perceptual image error between the referenced and distorted images. LPIPS [37] utilized a VGG-based encoder to extract the image feature and calculated the distance between different images. DISTS [38] observed the influence of structure and texture similarity on HVS, and proposed a CNN-based image quality metric. Although these works can good describe the subjective difference between the referenced and distorted images, the FR-IQA metrics require original high quality images which may not be accessible in practical situations.

NR-IQA gets more and more attentions in recent years because of the flexibly. NR-IQA metrics usually rely on the hand-crafted extractors or CNN-based architectures to explore the image features and predict the human’s perspective [39]. NIQE [40] built a space domain natural scene statistic model...
and provided an opinion-free indicator to estimate the subjective quality. BRISQUE [41] assessed the naturalness of images that consists with the visual perception. Recently, CNN-based metrics also demonstrate good performances on NR-IQA. Kang et al. used convolutional and pooling layers to explore the features and fit the perceptual score [42]. NIMA [43] developed a network with the help of classification backbones and regressed the quality score. Su et al. proposed a self-adaptive hyper network (HyperIQA) for blind IQA with good performance [44]. DBCNN [45] provided a dual bilinear network for NR-IQA. MUSIQ [46] also developed a transformer-based NR-IQA metric for the multi-scale information. Despite there are numerous NR-IQA methods with well-designed extractors and regressors, they almost neglect to investigate the special textural and structural degradation caused by image SR.

C. Image Quality Assessment for Image Super-Resolution

Different from General IQA, there are special textural and structural degradation situations in SR, making it hard to accurately predict the image quality. Ma et al. devised a NR-IQA method to regress the perceptual quality of SR images [9]. Jiang et al. provided a new perceptive on SR metric by splitting the structural and textural information with Karhunen-Loève Transformation [14]. Zhang et al. integrated AdaBoost decision tree regression and ridge regression to predict the quality score [10]. Zhou et al. provided a hand-crafted similarity estimator for FR-IQA on SR images [5]. Berón et al. built an opinion-free NR-IQA metric with the help of optimally extracted perceptual features [47]. These works highly rely on the hand-crafted extractors, which limit the representation capacity of features.

There are also CNN-based SR IQA metrics. Ahn et al. observed the influence of different distortion models, and developed a deep learning-based distortion sensitivity FR-IQA network for SR images [48]. Zhao et al. devised a dual stream network to predict the image quality with the help of LR images [1]. Zhang et al. also provided a NR-IQA metric with the help of VGG feature extractor [12]. However, these works almost neglect to distinguish the special degradation of SR images, and just utilize general CNN architectures to predict the opinion score.

III. METHODOLOGY

In this section, we introduce the proposed textural-structural joint learning network (TSNet) in the following manner. We introduce the prediction pipeline firstly. Then, we discuss the block design of the network with feature normalization (F-Norm) and the spatial attention (SA) mechanism, which are specially designed for SR features. Finally, the implementation details are described particularly.

A. Prediction Pipeline

Given a SR image $I^{SR}$, the task of NR-IQA is to predict the perceptual quality score $Q_{\text{score}}$ by a network such that

$$Q_{\text{score}} = TSNet(I^{SR}),$$

where $TSNet(\cdot)$ denotes the proposed TSNet.

Figure 2 shows the design of TSNet. The network is composed of the extractor and the regressor. The extractor explore the textural and structural features by two dual branches. After exploration, the regressor predicts the quality score by the non-linear mapping design. There are two branches in the extractor. The structural branch extracts the high-level semantic information by a pretrained VGG-19 extractor [13]. [19]. Let $F^S$ be the structural features, then there is

$$\{F^S_i\}_{i=1}^5 = VGG(I^{SR}),$$

where $F^S_i$ is the i-th structural feature explored by the VGG-19 extractor. The channel numbers of extracted features are with $c = 64, 128, 256, 512$ and 512 separately, and the resolutions of features are halved progressively.

Correspondingly, there are stages in the textural branch to explore the low-level textural information and mix the structural features by the designed residual SR block. Let $F^T$
be the explored textural feature, then for the \( i \)-th stage in the textural branch, there is

\[
F_i^T = RSRB([F_{i-1}^S, F_{i-1}^T]),
\]

where \( RSRB(\cdot) \) is the designed residual SR block, and \([\cdot]\) denotes the channel concatenation operation. To keep the same resolution as \( F^S \), there is a max-pooling operation on \( F^T \) after each stage.

For the first stage of \( F^T \), we utilize one convolutional layer and one residual SR block to explore the features from \( I^{SR} \), that is

\[
F_1^T = RSRB(Conv(I^{SR})).
\]

After exploration, the regressor predicts the quality score from the extracted features. There are 6 stages in the textural branch, then the quality score is predicted as

\[
Q_{score} = Reg(F_6^T),
\]

where \( Reg(\cdot) \) is the regressor.

### B. Residual SR Block

As shown in Figure 2, the residual SR block is composed of two convolutional layers, one ReLU activation, one SA layer and one F-Norm. The residual SR block follows the design in recent SR works [20], [27], [29] and removes the batch normalization. The SA layer and the F-Norm are developed at the end of residual SR block, following the recent network designs [20], [49], [50].

In the block, SA layer is utilized to make the important information more distinguishable by mimicking the human vision system (HVS) [15], which is composed of two group convolutional layers, one ReLU activation and one Sigmoid activation. Figure 3 shows the design of SA layer. One group convolution processes the input feature maps with group number as \( c_{sa} \)/4, where \( c_{sa} \) is the channel number of the input feature of SA. There are \( c_{sa}/4 \) filters in the group convolution. After that, one ReLU activation processes the feature to introduce the non-linearity. One symmetrical group convolution restores the shape of feature with filter number as \( c_{sa} \) and group number as \( c_{sa}/4 \). A Sigmoid activation is used to make the attention no-negative.

Besides the SA layer, F-Norm [20] is also developed in the residual SR block to substitute the batch normalization.

#### Table I

| F-Norm | SA | PLCC ↑ | SRCC ↑ |
|--------|----|--------|--------|
| w/o    | w/o| 0.9673 | 0.9649 |
| w      | w  | 0.9690 | 0.9662 |
| w      | w/o| 0.9711 | 0.9689 |
| w      | w  | 0.9720 | 0.9702 |

The upper right of Figure 2 shows the design of F-Norm. The F-Norm is composed of one depth-wise convolutional layer and one residual connection. Different from the batch normalization that widely used in different works [28], [45], F-Norm is more suitable for SR features since it can avoid the texture confusion and save the memory cost [20].

### C. Implementation Details

The structural branch of the TSNet is implemented by a pretrained VGG-19 network architecture. The five features are from layers with number \( l = 2, 7, 12, 21, \) and 30. The textural branch of the TSNet is composed of six residual SR blocks. All convolutional layers in the residual SR blocks are with filter number as \( f = 64 \) and the kernel size as \( 3 \times 3 \), except for the SA layer.

The regressor uses feature pooling to embed the features and utilizes convolutional layers to regress the quality score. Adaptive max pooling and adaptive average pooling methods compress the feature maps with size \( 4 \times 4 \). Then, the convolutional layers in the regressor process the compressed features with filter number as \( f = 256, 64 \) and 1 separately. There is no padding in the convolutional layers, such that the regressor can generate the quality score from features with any resolution.

### IV. Experiment

#### A. Settings

We choose two widely used SR-IQA datasets (CVIU-17 [9] and QADS [3]) for training and testing our TSNet. CVIU-17 [9] proposed by Ma et al. is one of the famous NR-IQA dataset for SR images, which contains 1620 images generated by nine traditional and CNN-based methods from six scaling factors. We randomly choose 60% images for training, 20% for validation and 20% for testing. QADS is also a famous FR-IQA dataset with 980 SR images, which specially contains the results from the GAN-based method. We use the same strategy as CVIU-17 to split the dataset for training and testing. We update the TSNet for 100 epochs by Adam optimizer [51] with learning rate as \( lr = 10^{-4} \). The network is implemented by the PyTorch [52] platform, and trained on one NVIDIA GTX 3080-Ti GPU. The performances of different methods are estimated by Pearson’s linear correlation coefficient (PLCC) and Spearman’s rank correlation coefficient (SRCC). The loss function is chosen as \( \ell_1 \) loss between the prediction result and the mean opinion score (MOS).

#### B. Model Analysis

1) Investigation on F-Norm and SA: To investigate the effectiveness of F-Norm and SA, we compare the PLCC and
SRCC on the QADS dataset. Table II shows the performance comparisons between F-Norm and SA on QADS dataset. In the table, we can find that the model with both F-Norm and SA achieves the highest PLCC and SRCC results than other methods. According to the results with and without SA (first and second lines), the SA brings 0.002 improvement on both PLCC and SRCC. From the results with and without F-Norm (first and third lines), the F-Norm leads to near 0.004 improvement on PLCC and 0.004 on SRCC. Specially, we can find from the results that F-Norm is more effective than SA with better PLCC/SRCC result. In this point of view, the F-Norm and SA boost the network performance and make the prediction more consist with the human’s perspective.

The SA is designed to make the visual sensitive features more distinguishable. To address this point, we illustrate and demonstrate the learned attention. Figure 4 shows the visualized attention maps of different images. The (a) column denotes the original input image, and the (b)-(e) columns are the learned attention maps from stage $i = 3$ to 6, which are normalized in range 0 to 1. The red area means the higher value, and the blue area means the lower value. From the upper images of the figure, we can find that the complex textures become noticeable at the second stage, such as the hairs, mouths, and canthus. With the increase of stages, the attention concentrates more on the hairs and faces. We can find in the attention map of stage 6 that the eyes and haris become noticeable with a significant higher attention value. This is in accordance with the human vision system (HVS) that people usually pay attention to the faces in the picture. The similar situation can be observed in the lower images of the figure. We can find that at the stage 3, some complex textures are observed with different attention values. With the increase of stages, the face and the flower become more distinguishable, which are more sensitive to the HVS.

2) Investigation on the structural and textural extraction: In the network, we devise two branches to explore the structural and textural features. To show the effectiveness of the dual exploration, we compare the performances of models with different branches. Table II shows the PLCC/SRCC performance comparisons between structural and textural branches.
C. Comparison with State-of-the-Art Methods

We compare our model with 10 FR-IQA methods: PSNR [3], SSIM [4], MS-SSIM [32], GMSD [34], FSIM [55], VIF [55], VSI [56], LPIPS [37], PieAPP [36], and DISTS [38]. We also compare our model with 7 NR-IQA methods: CNNIQA [42], HyperNet [44], DBCNN [45], NIQE [40], NIMA [43], BRISQUE [41] and WaDiQaM [57]. Specially, we compare our method with two SR-IQA metric: NRQM [9] and SFSN [5]. NIQE and BRISQUE are calculated by the MATLAB built-in function. NRQM and SFSN are calculated by the official code. We use the provided weight of NRQM for testing without further finetuning. Other implementations follow the GitHub repository. For a fair comparison, we re-train the CNNIQA, HyperNet, WaDiQaM, NIMA and DBCNN under the same protocol according to our method. Specially, the images predicted by HyperNet are resized as 224 × 224 for training and testing, which follows the requirement of the model’s implementation. The FR-IQA methods are not fine-tuned on the datasets for a fair comparison, since we cannot access the HR images during the no-reference assessment.

Table III shows the PLCC/SRCC comparisons on QADS dataset among different IQA methods. The FR-IQA methods are tested with the original model weights. The starred methods are re-trained on the QADS dataset. We can find that our method achieves the best PLCC/SRCC results than other works. Compared with CNNIQA, HyperNet and DBCNN that are specially designed for SR assessment. NRQM is a NR-IQA metric and SFSN is a FR-IQA method. Compared with these works, our method demonstrates a significant superior performance that more consist with the human’s perspective.

In the table, we can also find that the general IQA methods usually perform no better than the SR-IQA methods. This is in accordance with our motivation that the general methods usually focus on the hand-crafted signal degradation and noise, but rarely investigate the special textural and structural degradation in the SR situation.

Besides QADS, we also compare the performance on CVIU-17 dataset. Table IV shows the PLCC/SRCC comparisons on CVIU-17 dataset. We can find that our method achieves the best performance than other works, which means the predicted scores of TSNet are more consist with the human’s perspective. The starred methods are re-trained under the same protocol, our method achieves 0.1 improvement on PLCC and SRCC. Specially, NRQM and SFSN are specially designed for SR assessment. NRQM is a NR-IQA metric and SFSN is a FR-IQA method. Compared with these works, our method demonstrates a significant superior performance that more consist with the human’s perspective.

Table II shows the PLCC/SRCC performance comparisons between structural and textural branches on QADS dataset.

| Structural | Textural | PLCC ↑ | SRCC ↑ |
|------------|----------|--------|--------|
| w/o        | w/o      | 0.8946 | 0.8883 |
| w          | w        | 0.9587 | 0.9568 |
| w          | w        | 0.9704 | 0.9691 |
| w          | w        | 0.9720 | 0.9702 |

*https://github.com/chaofengc/IQA-PyTorch
Fig. 6. Scatter plots of different IQA methods on QADS dataset. The blue points denote the testing instances. The red line is the ideal linear relationship between MOS and the prediction score. All of the values are normalized in range $-5$ to $5$ for better view. The result of TSNet has the highest linear correlation with MOS, which means TSNet is the most consistent with the human’s perspective.

| Type   | Method       | PLCC   | SRCC   |
|--------|--------------|--------|--------|
| Full Ref | PSNR [3]          | 0.5985 | 0.5659 |
|        | SSIM [4]          | 0.6322 | 0.6249 |
|        | MS-SSIM [32]      | 0.7452 | 0.8089 |
|        | GMSD [34]         | 0.8359 | 0.8580 |
|        | FSIM [55]         | 0.7504 | 0.7678 |
|        | VIF [35]          | 0.8453 | 0.8660 |
|        | VSI [56]          | 0.6969 | 0.7249 |
|        | LPIPS [37]        | 0.8306 | 0.8220 |
|        | PieAPP [36]       | 0.7841 | 0.7832 |
|        | DISTS [58]        | 0.8642 | 0.8643 |
| No Ref  | NIQE [40]         | 0.3150 | 0.3279 |
|        | BRISQUE [41]      | 0.2130 | 0.2277 |
|        | NIMA [43]         | 0.9601 | 0.9585 |
|        | CNNIQA [42]       | 0.9280 | 0.9177 |
|        | HyperNet [43]     | 0.8863 | 0.8836 |
|        | DBCNN [45]        | 0.9659 | 0.9602 |
|        | WaDIQaM [57]      | 0.9254 | 0.9185 |
| SR Metrics | SFSN [5]         | 0.7547 | 0.8612 |
|        | TSNet(Ours)       | 0.9741 | 0.9720 |

Images.

To further investigate the effectiveness of the proposed TSNet, we demonstrate the scatter plots between the MOS and the prediction scores and analyze the correlation. Figure 6 shows the scatter plots of different IQA methods on QADS dataset. The red line shows the ideal linear relationship between MOS and prediction results. All of the values are normalized in range $-5$ to $5$ for better view. The first and second rows are FR-IQA methods. The methods in the last row are retrained CNN-based NR-IQA works and our TSNet. In the figure, we can find that results of TSNet are more in line with the MOS. Compared with CNNIQA and DBCNN, there are fewer outliers in the TSNet. Figure 7 shows the scatter plots on CVIU-17 dataset. We can find that compared with the FR-IQA methods and the finetuned NR-IQA methods, our TSNet has fewer outliers and performs more consistently with the MOS.

In order to evaluate the effectiveness of TSNet on different SR methods, we also compare the PLCC/SRCC performances on three kinds (interpolation-based, dictionary-based, CNN-based) of SR methods in QADS dataset. Table V shows the results of different IQA metrics on different SR methods. In the table, we can find our method achieves the best performances on all SR methods than other IQA metrics. The images of deep learning SR methods are difficult to predict, since the PLCC/SRCC results are lower than other tasks. Even though, TSNet achieves 0.92 on PLCC and 0.94 on SRCC, which mean the metric is in consistence with the human’s perspective.
Fig. 7. Scatter plots of different IQA methods on CVIU-17 dataset. The blue points denote the testing instances. The red line is the ideal linear relationship between MOS and the prediction score. All of the values are normalized in range $-5$ to $5$ for better view. The result of TSNet has the highest linear correlation with MOS, which means TSNet is the most consistent with the human’s perspective.

| Methods     | PSNR [3] | SSIM [4] | MS-SSIM [32] | LPIPS [37] | DISTS [38] | DBCNN [45] | CNNIQA [42] | TSNet (Ours) |
|-------------|----------|----------|--------------|-------------|-------------|------------|-------------|-------------|
| Interpolation |          |          |              |             |             |            |             |             |
| PLCC        | 0.0912   | 0.4710   | 0.5885       | 0.6835      | 0.6812      | 0.9623     | 0.9427      | 0.9730      |
| SRCC        | 0.1930   | 0.4545   | 0.6076       | 0.6284      | 0.6556      | 0.9374     | 0.9154      | 0.9551      |
| Dictionary  |          |          |              |             |             |            |             |             |
| PLCC        | 0.3166   | 0.3268   | 0.7211       | 0.6838      | 0.6723      | 0.9478     | 0.9114      | 0.9698      |
| SRCC        | 0.3195   | 0.5167   | 0.7859       | 0.6728      | 0.6599      | 0.9383     | 0.8812      | 0.9584      |
| Deep learning |          |          |              |             |             |            |             |             |
| PLCC        | 0.2005   | 0.3214   | 0.4077       | 0.5308      | 0.4596      | 0.7980     | 0.7339      | 0.9266      |
| SRCC        | 0.2392   | 0.4143   | 0.6833       | 0.5711      | 0.4832      | 0.8362     | 0.7763      | 0.9455      |
| Total       | 0.3099   | 0.5188   | 0.6586       | 0.6775      | 0.6739      | 0.9477     | 0.9105      | 0.9720      |

V. Conclusion

In this paper, we proposed a CNN-based NR-IQA method named TSNet. Different from existing NR-IQA methods, we noticed that there are special textural and structural information losses in the SR situation, and devised a dual stream network for joint textural and structural feature exploration.
Motivated by the human vision system (HVS), we developed the spatial attention mechanism to make the salient information more distinguishable and improve the accuracy of the prediction. Feature normalization (F-Norm) was also considered in the TSNet to better explore the super-resolved features. Experimental results show the proposed TSNet has better performance on the QADS and CVIU-17 datasets than other state-of-the-art IQA methods.

In future work, we aim to improve the TSNet from three perspectives. Firstly, we decide to design a FR-IQA version of TSNet for better assessment performance. Secondly, we apply the TSNet to image super-resolution tasks, which acts as a loss function for improving the restoration capacity. Finally, we attempt to apply the meta-learning strategy to TSNet, making it more flexible for practical applications.

REFERENCES

[1] T. Zhao, Y. Lin, Y. Xu, W. Chen, and Z. Wang, “Learning-based quality assessment for image super-resolution,” IEEE Transactions on Multimedia, pp. 1–1, 2021.

[2] F. Zhou, R. Yao, B. Liu, and G. Qiu, “Visual quality assessment for super-resolved images: Database and method,” IEEE Transactions on Image Processing, vol. 28, no. 7, pp. 3528–3541, 2019.

[3] A. Horéd and D. Zou, “Image quality metrics: Psnr vs. ssim,” in International Conference on Pattern Recognition (ICPR), 2010, pp. 2366–2369.

[4] 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, 2004.

[5] W. Zhou, Z. Wang, and Z. Chen, “Image super-resolution quality assessment: Structural fidelity versus statistical naturalness,” in International Conference on Quality of Multimedia Experience (QoMEX), 2021, pp. 61–64.

[6] H. Yeganeh, M. Rostami, and Z. Wang, “Objective quality assessment of interpolated natural images,” IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 4651–4663, 2015.

[7] J. Chen, Y. Xu, K. Ma, H. Huang, and T. Zhao, “A hybrid quality metric for non-integer image interpolation,” in 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), 2018, pp. 1–3.

[8] G. Wang, L. Li, Q. Li, K. Gu, Z. Lu, and J. Qian, “Perceptual evaluation of single-image super-resolution reconstruction,” in IEEE International Conference on Image Processing (ICIP), 2017, pp. 3145–3149.

[9] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, “Learning a no-reference quality metric for single-image super-resolution,” Computer Vision and Image Understanding, vol. 158, pp. 1–16, 2017.

[10] K. Zhang, D. Zhu, J. Jing, and X. Gao, “Learning a cascade regression for no-reference super-resolution image quality assessment,” in IEEE International Conference on Image Processing (ICIP), 2019, pp. 450–453.

[11] J. Berson, H. D. Benitez-Restrepo, and A. C. Bovik, “Blind image quality assessment for super resolution via optimal feature selection,” IEEE Access, vol. 8, pp. 143 201–143 218, 2020.

[12] K. Zhang, D. Zhu, J. Li, X. Gao, F. Gao, and J. Lu, “Learning stacking regression for no-reference super-resolution image quality assessment,” Signal Processing, vol. 178, p. 107771, 2021.

[13] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in European Conference on Computer Vision (ECCV), B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., 2016, pp. 694–711.

[14] Q. Jiang, Z. Liu, K. Gu, F. Shao, X. Zhang, H. Liu, and W. Lin, “Single image super-resolution quality assessment: A real-world dataset, subjective studies, and an objective metric,” IEEE Transactions on Image Processing, vol. 31, pp. 2279–2294, 2022.

[15] Q. Lai, S. Khan, Y. Nie, H. Sun, J. Shen, and L. Shao, “Understanding more about human and machine attention in deep neural networks,” IEEE Transactions on Multimedia, vol. 23, pp. 2086–2099, 2021.

[16] B. Bare, K. Li, B. Yan, B. Feng, and C. Yao, “A deep learning based no-reference image quality assessment model for single-image super-resolution,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 1223–1227.

[17] Y. Fang, C. Zhang, W. Yang, J. Liu, and Z. Guo, “Blind visual quality assessment for image super-resolution by convolutional neural network,” Multimedia Tools and Applications, vol. 77, no. 22, pp. 29 829–29 846, 2018.

[18] V. Khrulkov and A. Babenko, “Neural side-by-side: Predicting human preferences for no-reference super-resolution evaluation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021, pp. 4988–4997.
