Content-dependency reduction with multi-task learning in blind stitched panoramic image quality assessment

Hou, Jingwen; Lin, Weisi; Zhao, Baoquan

2020

Hou, J., Lin, W., & Zhao, B. (2020). Content-dependency reduction with multi-task learning in blind stitched panoramic image quality assessment. Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP). doi:10.1109/ICIP40778.2020.9191241

https://hdl.handle.net/10356/144376

https://doi.org/10.1109/ICIP40778.2020.9191241

CONTENT-DEPENDENCY REDUCTION WITH MULTI-TASK LEARNING IN BLIND STITCHED PANORAMIC IMAGE QUALITY ASSESSMENT

Jingwen Hou, Weisi Lin[†] *and Baoquan Zhao*

Nanyang Technological University, Singapore
jingwen003@e.ntu.edu.sg, {wslin, bqzhao}@ntu.edu.sg

ABSTRACT

In this work, we investigate deep learning based solutions to blind quality assessment of stitched panoramic images (SPI). The main problem to tackle is that the ground truth data is usually insufficient. As a result, the learned model can easily overfit data with specific content. Because most distortions of SPIs lie within local regions, the problem cannot be alleviated by commonly-used patch-wise training, which assumes local quality equals global quality. We propose a multi-task learning strategy which encourages learned representation to be less dependent on image content. A siamese network with two weight-shared CNN branches is trained to simultaneously compare the quality of two images of the same scene and predict the quality score of each image. Since two images of the same scene are processed by the same CNN, the CNN tends to find their quality differences instead of content differences under the constraint of the quality ranking objective. Because two tasks share the same representations learned by the CNN, the regression task can be further benefited from the quality-sensitive representations. Extensive experiments demonstrate the effectiveness of the proposed model and its superiority over existing SPI quality assessment methods.

Index Terms— blind image quality assessment, stitched panoramic image, multi-task learning, virtual reality

1. INTRODUCTION

Virtual reality (VR) targets at providing users with an immersive and realistic experience in a virtual environment via a wide field-of-view perceived on head mounted displays. Image stitching is one of the key technologies to generate 360-degree panoramic images for VR applications. Typically, a panoramic image is obtained by stitching together a sequence of images with overlapping areas into an image with a large field-of-view. Since humans are generally the ultimate users of such stitched panoramic images (SPI), perceptual visual quality assessment tool becomes a necessity in aiding the evaluation of user experience [1]. And specifically, as there is no reference image for SPI, blind image quality assessment (BIQA) metrics are highly desired.

To deal with the blind stitched panoramic image quality assessment (SPIQA) problem, a naive solution is to directly employ existing generic BIQA techniques, which can be broadly classified into two branches: NSS-based methods [2, 3, 4, 5] and deep learning based methods [6, 7, 8, 9, 10]. For the first branch, it assumes that when an image suffers from distortions, perturbations of statistical naturalness can be observed, measured and used to predict perceptual visual quality. For the second branch, one key issue is to resolve the problem of insufficient ground truth. Therefore, patch-wise training is widely used in previous deep learning methods for data augmentation [6, 7, 8, 9]. Unfortunately, SPI distortions such as ghosting, color and geometric distortion [11] are usually presented in local regions rather than the whole image, which makes both NSS-based methods and deep learning based methods fail in the case of SPIs. NSS-based methods' failure mainly stems from its inability in capturing local perturbations. Additionally, hand-crafted features from previous NSS-based methods may also not suitable for new types of distortions in SPIs. For deep learning based methods, patch-wise training is also not applicable because it labels quality of each image patch as its source image, based on the assumption that image quality is evenly distributed.

The limitations of generic BIQA techniques promote researchers to develop tailored quality metrics for SPI [11, 12, 13]. Ling et al. [12] uses convolutional sparse coding for feature extraction, while its feature selection relies on a rule-based process instead of a learnable process as CNN. Gao et al. [13] proposed a quality evaluator that locates distortions on SPIs and quantifies image quality based on detected distortions. But the proposed metric is not perceptual-driven. Recently, the first public-available SPIQA dataset ISIQA and a reduced-reference objective quality assessment method was introduced [11]. However, it requires to compare an SPI with its constituent images.

In this work, we present a novel deep learning based blind SPIQA method. Like the aforementioned deep learning based solutions, similar problem also arises from lack of ground truth data. Generally, the shortage of training data can be relieved by data augmentation which essentially increases the size of expressive space, or introducing extra constraints to exclude low-quality solutions in the solution space [14, 15]. Data augmentation can be achieved by collecting more ground-truth or patch-wise training, while both methods are not practical in the case of SPIQA. Our idea is to optimize the solution space by introducing quality ranking as an auxiliary learning task. Intuitively, a model works well on quality scoring should also be good at quality ranking. Therefore, using quality ranking as an extra constraint can make the model more likely to do well in both tasks than model merely trained subject to quality scoring.

One problem caused by data insufficiency is overfitting data with specific content. For example, when training with holistic images, if most low-quality images in training set have trees, the model may learn that an image is of low quality if it has trees inside. Training with patches instead of holistic images can resolve this issue, while it cannot be used in blind SPIQA. Alternatively, we require each training image pair to have the same scene. In this setting, features of two images with similar content are learned by the same CNN and used for both ranking and scoring. Because two images have similar content, the CNN becomes less sensitive to image content in order to tell the difference between the two images for the ranking purpose. Because two tasks share the same representations learned by CNN, the scoring task can be further benefited from the representations which are less dependent on image content. A work similar to ours is RankIQA [10]. Though it enables neural network to initialize with appropriate parameters, the solution is very likely to fall into a bad local optimum because the solution space is still unconstrained to an optimized region during the finetuning stage.

In summary, this work makes three main contributions: Firstly, we introduce multi-task learning to the task of blind SPIQA and develop an effective siamese network to address its particular problem of ground truth insufficiency; Secondly, in view of the characteristic of SPI and the limitations of existing patch-wise training schemes, a joint training strategy of image pairs is proposed to reduce image content dependency and prevent overfitting; Last but not least, the proposed model markedly outperforms both existing NSS-based and deep-learning based methods on the test SPI dataset.

2. PROPOSED METHOD

2.1. Problem formulation

The BIQA problem can be expressed as: given an input image $\mathbf{I}$, the model $M(\cdot)$ should predict its mean opinion score (MOS). Therefore, the problem can be expressed as $q = M(\mathbf{I})$, where q represents quality score. We decompose a deep learning based model $M(\cdot)$ into a feature extractor $f(\cdot)$ which is usually a CNN, and a predictor $h(\cdot)$ which is usually one or multiple stacking fully-connected (FC) layers. Given image $\mathbf{I}$, the feature extractor extracts feature $\mathbf{x} \in \mathbb{R}^n$. Then the features are fed into the predictor to predict MOS, which can be expressed as $q = h(\mathbf{x})$. Therefore, the inference process of a neural network can be formulated as:

$$q = h_{\theta_2} \circ f_{\theta_1}(\mathbf{I}) \tag{1}$$

where θ_1 and θ_2 are parameters of the feature extractor and the predictor, respectively.

In the proposed multi-task learning method, learning-to-rank is introduced as the auxiliary task. During the training stage, given a pair of images, the model is expected to predict the quality score of each image in the pair and the ranking vector. Suppose two SPIs of the same scene $\mathbf{I_1}, \mathbf{I_2}$ are given and $\hat{q}_1, \hat{q}_2$ are their ground truth quality score, ground truth of ranking is given as a 2-dimensional one-hot encoded vector $\hat{\mathbf{r}}$. When $\hat{q}_1 > \hat{q}_2$, $\hat{\mathbf{r}} = [1\ 0]^T$. When $\hat{q}_1 < \hat{q}_2$, $\hat{\mathbf{r}} = [0\ 1]^T$.

In practice, we use same feature extractor $f_{\theta_1}(\cdot)$ and predictor $h_{\theta_2}(\cdot)$ to extract features from both images and predict quality score of each image. And a separate predictor $h_{\theta_3}(\cdot)$ is used to predict a ranking vector $\mathbf{r}$ with fused feature. The inference process of the training stage is expressed as:

$$q_1 = h_{\theta_2} \circ f_{\theta_1}(\mathbf{I_1}) \tag{2}$$

$$q_2 = h_{\theta_2} \circ f_{\theta_1}(\mathbf{I_2}) \tag{3}$$

$$\mathbf{r} = \sigma \circ h_{\theta_3} \circ g(f_{\theta_1}(\mathbf{I_1}), f_{\theta_1}(\mathbf{I_2})) \tag{4}$$

where $\mathbf{r}$ represents predicted ranking vector for the image pair, $g(\cdot)$ represents a feature fusion function and $\sigma(\mathbf{x_{(i)}}) = e^{\mathbf{x}_{(i)}} / \sum_j e^{\mathbf{x}_{(j)}}$, which is the softmax function. Thus, the model is not only trained subject to a scoring loss, but also subject to a ranking loss. The overall loss function can be formulated as:

$$\begin{aligned} &L(q_1, q_2, \mathbf{r}, \hat{q}_1, \hat{q}_2, \hat{\mathbf{r}}) \\ &= \lambda_1 L_1(q_1, \hat{q}_1) + \lambda_1 L_1(q_2, \hat{q}_2) + \lambda_2 L_2(\mathbf{r}, \hat{\mathbf{r}}) \end{aligned} \tag{5}$$

where λ_1, λ_2 are weights for each task and $L_1(\cdot), L_2(\cdot)$ are loss functions for scoring and ranking, respectively. Noted that during testing stage, only the scoring branch is taken.

2.2. Network architecture

As shown in Fig.1, network architecture is designed according to the formulation in Sec.2.1. ResNet-18 [16] has been used as feature extractor. Noted that its last output layer has been removed. Therefore, the output of the ResNet-18 feature extractor from its last global average pooling layer [17] is a 512-dimensional feature vector, which is fed into following predictors. The design of predictors is also inspired by [16], whose feature extractor is heavy while the predictor is light. For the predictor of scoring task, instead of using multiple stacking FC layers, only one single-unit FC layer is attached to the feature extractor. The same design methodology has been used for building the predictor of quality ranking. A problem left in Sec.2.1 is how to fuse features of two images. One desired result for using a feature fusion function is that, when we swap the order of two inputs, the ranking results

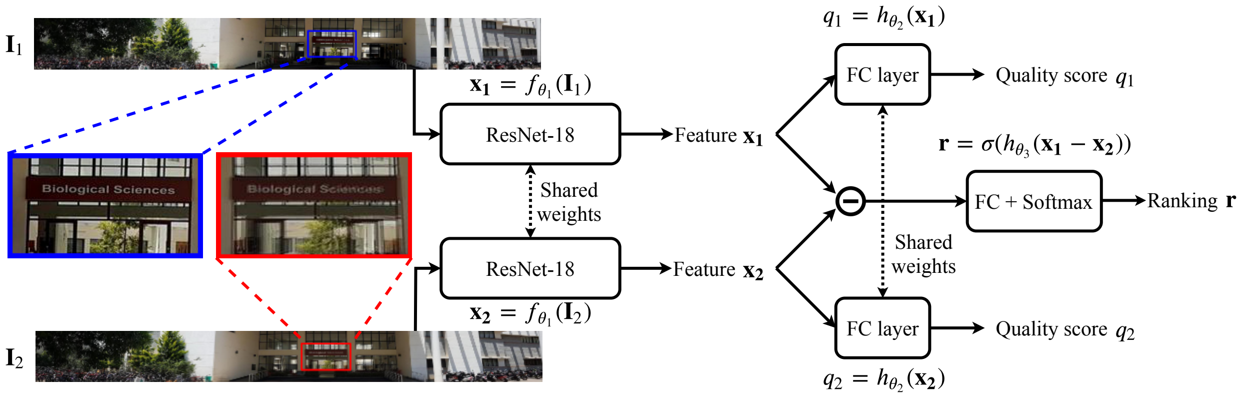

Fig. 1: Overview of the proposed method for blind SPIQA. Given a pair of SPIs of the same scene $\mathbf{I_1}, \mathbf{I_2}$, the model predicts their quality score $\mathbf{q_1}, \mathbf{q_2}$ and one-hot ranking vector $\mathbf{r}$. Zoomed-in part of $\mathbf{I}_1$ is undistorted while that of $\mathbf{I}_2$ has ghosting artifact. The intermediate features from ResNet-18 backbone $\mathbf{x_1}, \mathbf{x_2} \in \mathbb{R}^{512}$ are fused by element-wise subtraction.

should also be swapped. We prove that using element-wise subtraction as the feature fusion function $g(\cdot)$ in our setting, Eq.4 satisfies this condition. The proof is given as follows.

Proposition 2.1. *Suppose* $h_{\theta_3}(\mathbf{x}) = \theta_3^T\mathbf{x}$*,* $\theta_3 \in \mathbb{R}^{n\times 2}, \mathbf{x} \in \mathbb{R}^n$*, feature fusion function* $g(\mathbf{x_1}, \mathbf{x_2}) = \mathbf{x_1} - \mathbf{x_2}$*,* $\mathbf{x_1}, \mathbf{x_2} \in \mathbb{R}^n$ *and* $\sigma(\mathbf{x_{(i)}}) = e^{\mathbf{x}_{(i)}}/\sum_j e^{\mathbf{x}_{(j)}}$*,* $\mathbf{x} \in \mathbb{R}^n$*. If* $\sigma \circ h_{\theta_3} \circ g(\mathbf{x_1}, \mathbf{x_2}) = [a\ b]^T$*, then* $\sigma \circ h_{\theta_3} \circ g(\mathbf{x_2}, \mathbf{x_1}) = [b\ a]^T$ *for* $a, b \in \mathbb{R}$*.*

Proof. Let $\mathbf{x} = \mathbf{x_1} - \mathbf{x_2}$, and $\theta_3^T\mathbf{x} = [\mathbf{r}_{(1)}, \mathbf{r}_{(2)}]^T$, then
$\because\ \sigma \circ h_{\theta_3} \circ g(\mathbf{x}_1, \mathbf{x}_2) = \sigma(\theta_3^T\mathbf{x}) = [\frac{e^{\mathbf{r}(\mathbf{1})}}{e^{\mathbf{r}(\mathbf{1})}+e^{\mathbf{r}(\mathbf{2})}}\ \frac{e^{\mathbf{r}(\mathbf{2})}}{e^{\mathbf{r}(\mathbf{1})}+e^{\mathbf{r}(\mathbf{2})}}]^T$
$\therefore \sigma \circ h_{\theta_3} \circ g(\mathbf{x}_2, \mathbf{x}_1) = \sigma(-\theta_3^T\mathbf{x})$
$= [\frac{e^{-\mathbf{r}(\mathbf{1})}}{e^{-\mathbf{r}(\mathbf{1})}+e^{-\mathbf{r}(\mathbf{2})}}\ \frac{e^{-\mathbf{r}(\mathbf{2})}}{e^{-\mathbf{r}(\mathbf{1})}+e^{-\mathbf{r}(\mathbf{2})}}]^T = [\frac{e^{\mathbf{r}(\mathbf{2})}}{e^{\mathbf{r}(\mathbf{1})}+e^{\mathbf{r}(\mathbf{2})}}\ \frac{e^{\mathbf{r}(\mathbf{1})}}{e^{\mathbf{r}(\mathbf{1})}+e^{\mathbf{r}(\mathbf{2})}}]^T$ □

Thus, the inference process of ranking can be given as:

$$\mathbf{r} = \sigma \circ \mathbf{h}_{\theta_3}(\mathbf{f}_{\theta_1}(\mathbf{I_1}) - \mathbf{f}_{\theta_1}(\mathbf{I_2})) \tag{6}$$

Taking the feature difference $\mathbf{x_1} - \mathbf{x_2}$ as input, a two-unit FC layer followed by a softmax function is used to predict the quality ranking vector $\mathbf{r}$. To ensure the two branches of the siamese network are simultaneously updated, the feature extractors and the predictors of quality score inference branches share the same parameters.

2.3. Loss function

As formulated in Eq.5, the loss function for training the siamese network is a combination of scoring loss and ranking loss weighted by parameters λ_1, λ_2. For quality score regression, we use mean squared error (MSE) as the loss function. The choice of quality ranking loss mainly follows suggestions from [18, 19] and therefore we use cross-entropy (CE) loss as the ranking loss. Since we treat weight parameters λ_1, λ_2 as hyper-parameters, to reduce the number of hyper-parameters, we only add a weight parameter to the scoring loss. Therefore, the multi-task loss can be expressed as:

$$L(q_1, q_2, \mathbf{r}, \hat{q}_1, \hat{q}_2, \hat{\mathbf{r}}) =$$

$$\lambda[MSE(q_1, \hat{q}_1) + MSE(q_2, \hat{q}_2)] + CE(\mathbf{r}, \hat{\mathbf{r}}) \tag{7}$$

3. EXPERIMENTS

3.1. Implementation details

Our network is firstly pretrained on KADID-10K [20] and then finetuned and evaluated on the target dataset ISIQA [11].

1) Construction of pretraining set. Distorted images of each scene are randomly combined into 2,000 image pairs. There are 81 scenes in the KADID-10K, where 64 scenes are chosen for pretraining and the rests are used for validation. Thus, 128,000 image pairs are generated for pretraining.

2) Construction of training set and validation set. After network pretraining, the network is finetuned and evaluated on the ISIQA dataset. The evaluation process mainly follows [11]. Since our GPU is not able to support training on images with a full resolution from ISIQA, all images are downsampled to 2000 × 400. 80% of the data are used for training and the rest 20% are used for validation. To ensure no overlapping image content between training set and validation set, the SPIs are divided according to scenes of their constituent images. Specifically, there are 26 scenes in total, where 21 scenes are randomly chosen for training and the rest 5 scenes are used for validation. We repeat this process 10 times and report the medians and standard deviations of evaluation results. Spearman rank order coefficient (SRCC) and Pearson linear correlation coefficient (PLCC) are used for evaluation.

Table 1: Comparison of different IQA models on ISIQA across 10 sessions regarding SRCC and PLCC.

	SRCC		PLCC	
	MED [†]	STD [§]	MED [†]	STD [§]
BRISQUE [3]	0.6225	0.0828	0.5954	0.0747
NIQE [2]	0.1524	0.1887	0.1051	0.1753
DIIVINE [21]	0.5683	0.1398	0.5880	0.1314
ResNet-18 [16]	0.6679	0.1084	0.6714	0.1477
RankIQA [10]	0.6140	0.1142	0.6144	0.1289
BSPIQA (proposed)	**0.7593**	**0.0621**	**0.8022**	**0.0546**

[†] Median value of SRCC and PLCC.

[§] Standard deviation of SRCC and PLCC.

Note: Bold values in the table indicate the best results.

3) Hyperparameters. We use one GeForce GTX 1080Ti for training. The ResNet-18 backbone is initialized with ImageNet pretrained weights. For both pretraining and finetuning stage, we adopt Adam optimizer with learning rate 3×10^{-5}, and we find $\lambda = 0.01$ in Eq.7 is the best weight for the scoring task. For the pretraining stage, the model is trained for 1 epoch with batch size 32, and reaches 0.788 in SRCC. And for the finetuning stage, the model is trained for 20 epochs with batch size 8.

4) Compared models. For NSS-based methods, we choose BRSIQUE [3], NIQE [2] and DIIVINE [21] as [11] for comparison. Noted that we have not compared with the model proposed in [11], since the model requires to compare SPIs with their constituent images while our model is a blind IQA model. We refer to and validate the reported results in [11]. For deep learning based methods, we choose ResNet-18 and RankIQA for comparison. Since their original settings cannot be directly applied to the case of blind SPIQA, we reimplement and adjust their settings for fair comparison:

ResNet-18 baseline [16]. Because our siamese network takes ResNet-18 as the backbone feature extractor, we also need to compare our method with a ResNet-18. We alter the last layer of ImageNet pretrained ResNet-18 to a single unit FC layer as Sec.2.2. Then the model is further pretrained on KADID-10K and evaluated on the target set ISIQA.

RankIQA [10]. RankIQA uses pairwise ranking as a proxy task for pretraining the network while we propose to learn ranking and scoring simultaneously. To compare with it, we also use pretraining set generated from KADID-10K to pretrain the siamese network merely on ranking task subject to CE loss. Then a single branch is taken to finetune on target dataset ISIQA on MOS regression subject to MSE loss.

3.2. Experiment results

The performance of different image quality assessment models on the test dataset is shown in Table 1. As seen from the table, the proposed model, namely BSPIQA, achieves the best performance among all methods regarding SRCC and PLCC.

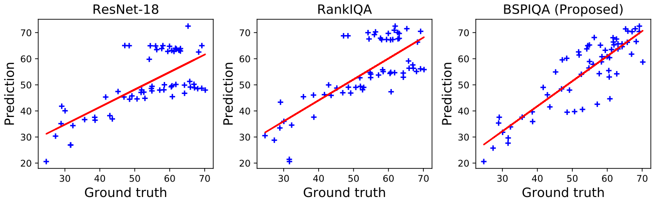

Fig. 2: Scatter plots and linear regression of the predicted results for ResNet-18, RankIQA, and proposed BSPIQA.

As for NSS-based IQA models, BRISQUE is a notch above DIIVINE, and both of them are far better than NIQE. However, comparing NSS-based methods to the other three deep learning methods, we can see that even the ResNet-18 baseline can work better than the best one BRISQUE. And the proposed model improves the medians of SRCC and PLCC markedly over the best NSS-based method by around 0.14 and 0.21, respectively. We believe that the performance of deep learning based methods are higher because handcrafted features of previous NSS-based methods are not fully applicable to the current SPIQA problem. While for deep learning based methods, the features are automatically learned by the CNN, and the quality of learned features depends on the design of the network architecture and learning strategy.

For deep learning based methods, our BSPIQA works better than the ResNet-18 baseline, which indicates that by adding the proposed auxiliary task, the quality predication performance can be effectively improved. Besides, by comparing the results of RankIQA and ResNet-18 baseline, the performance of RankIQA becomes even lower than the ResNet-18 baseline. We can also see from Fig.2 that our model has a better correlation with human perception than the other two deep learning methods. These results further support the idea that training merely on the scoring task can lead to overfitting, even if the model is initialized with optimized parameters. By introducing learning-to-ranking as the auxiliary task and training with image pairs of the same scene, the solution space can be well optimized and the content dependency can be effectively reduced.

4. CONCLUSION

This paper introduces a novel deep learning based approach for blind SPIQA. To harvest an effective learning-based BIQA model, a crucial factor is how to cope well with the general problem of ground truth insufficiency. To this end, we adopt the multi-task learning paradigm and develop a siamese network with pair-wise ranking as the second learning task to shrink the solution space and a joint training strategy of image pairs to reduce content-dependency and prevent overfitting. The experimental results demonstrate the effectiveness of the proposed model in addressing the raised particular challenges in blind SPIQA, which cannot be properly handled by conventional patch-wise training techniques.

5. REFERENCES

[1] Weisi Lin and C-C Jay Kuo, “Perceptual visual quality metrics: A survey,” *Journal of visual communication and image representation*, vol. 22, no. 4, pp. 297–312, 2011.

[2] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, “Making a “completely blind” image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.

[3] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[4] Michele A Saad, Alan C Bovik, and Christophe 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, 2012.

[5] Deepti Ghadiyaram and Alan 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, 2017.

[6] Le Kang, Peng Ye, Yi Li, and David Doermann, “Convolutional neural networks for no-reference image quality assessment,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1733–1740.

[7] Yuming Li, Lai-Man Po, Litong Feng, and Fang Yuan, “No-reference image quality assessment with deep convolutional neural networks,” in *2016 IEEE International Conference on Digital Signal Processing (DSP)*. IEEE, 2016, pp. 685–689.

[8] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech 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, 2017.

[9] Jongyoo Kim and Sanghoon Lee, “Fully deep blind image quality predictor,” *IEEE Journal of selected topics in signal processing*, vol. 11, no. 1, pp. 206–220, 2016.

[10] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov, “Rankiqa: Learning from rankings for no-reference image quality assessment,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1040–1049.

[11] Pavan Chennagiri Madhusudana and Rajiv Soundararajan, “Subjective and objective quality assessment of stitched images for virtual reality,” *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5620–5635, 2019.

[12] Suiyi Ling, Gene Cheung, and Patrick Le Callet, “No-reference quality assessment for stitched panoramic images using convolutional sparse coding and compound feature selection,” in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[13] Chenqiang Gao Luyu Yang, Jiang Liu, “An error-activation-guided blind metric for stitched panoramic image quality assessment,” *CCF Chinese Conference on Computer Vision*, vol. 772, pp. 256–268, 2017.

[14] Sheng Yang, Qiuping Jiang, Weisi Lin, and Yongtao Wang, “Sgdnet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment,” in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2019.

[15] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo, “End-to-end blind image quality assessment using deep neural networks,” *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2017.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] Min Lin, Qiang Chen, and Shuicheng Yan, “Network in network,” *arXiv preprint arXiv:1312.4400*, 2013.

[18] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng 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, 2017.

[19] Christopher JC Burges, “From ranknet to lambdarank to lambdamart: An overview,” *Learning*, vol. 11, no. 23-581, pp. 81, 2010.

[20] Hanhe Lin, Vlad Hosu, and Dietmar Saupe, “Kadid-10k: A large-scale artificially distorted iqa database,” in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–3.

[21] Anush Krishna Moorthy and Alan Conrad Bovik, “Blind image quality assessment: From natural scene statistics to perceptual quality,” *IEEE transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.