SGDNet: An End-to-End Saliency-Guided Deep Neural Network for No-Reference Image Quality Assessment

Sheng Yang
Nanyang Technological University
Singapore
syang014@e.ntu.edu.sg

Qiuping Jiang
Ningbo University
Ningbo, China
jiangqiuping@nbu.edu.cn

Weisi Lin*
Nanyang Technological University
Singapore
wslin@ntu.edu.sg

Yongtao Wang
Peking University
Beijing, China
wyt@pku.edu.cn

ABSTRACT

We propose an end-to-end saliency-guided deep neural network (SGDNet) for no-reference image quality assessment (NR-IQA). Our SGDNet is built on an end-to-end multi-task learning framework in which two sub-tasks including visual saliency prediction and image quality prediction are jointly optimized with a shared feature extractor. The existing multi-task CNN-based NR-IQA methods which usually consider distortion identification as the auxiliary sub-task cannot accurately identify the complex mixtures of distortions exist in authentically distorted images. By contrast, our saliency prediction sub-task is more universal because visual attention always exists when viewing every image, regardless of its distortion type. More importantly, related works have reported that saliency information is highly correlated with image quality while this property is fully utilized in our proposed SGNet by training the model with more informative labels including saliency maps and quality scores simultaneously. In addition, the outputs of the saliency prediction sub-task are transparent to the primary quality regression sub-task by providing a kind of spatial attention masks for a more perceptually-consistent feature fusion. By training the whole network with the two sub-tasks together, more discriminant features can be learned and a more accurate mapping from feature representations to quality scores can be established. Experimental results on both authentically and synthetically distorted IQA datasets demonstrate the superiority of our SGDNet, as compared to the state-of-the-art approaches.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Image processing**;

*Corresponding Author.

KEYWORDS

Image Quality Assessment; No-Reference; Visual Saliency Prediction; Multi-Task Learning; Authentic Distortions

ACM Reference Format:
Sheng Yang, Qiuping Jiang, Weisi Lin, and Yongtao Wang. 2019. 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 (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350990

1 INTRODUCTION

Image quality assessment (IQA) aims to predict the perceptual quality of digital images in a manner that is consistent with human subjective opinions. It is a fundamental problem in many perceptual-based visual media applications, such as image compression [23], image deblurring [19, 22], image super-resolution [28], and more. According to the accessibility of the pristine reference images, IQA models can be classified into full-reference (FR) [26, 40, 47, 49], reduced-reference (RR) [30, 42, 46], and no-reference (NR) [8, 12, 13, 21, 43] IQA types. Among them, NR-IQA has a board range of application scenarios since reference images are not accessible in most practical applications, especially for predicting the quality of real-world images with authentic distortions.

Traditional NR-IQA methods generally follow a two-stage processing framework including feature extraction and quality regression. Related works have shown that the performance of these NR-IQA models heavily depends on their carefully designed quality-aware features based on the domain knowledge of natural scene statistics (NSS) [31, 32, 36] and human visual properties [8, 24]. Lately, with the advent of deep convolutional neural network (CNN), these hand-crafted feature-based NR-IQA models are surpassed by deep CNN-based models due to the powerful capacity of deep CNN architectures in jointly optimizing the feature extraction and quality regression modules in a data-driven manner.

To learn a better feature representation, some recently proposed deep CNN-based NR-IQA models seek to use a multi-task learning strategy where an auxiliary sub-task and a primary sub-task (i.e., quality prediction) are jointly optimized in an end-to-end manner. With the help of such an auxiliary yet closely related sub-task, these models can learn more discriminant feature representations from the input raw data to improve the quality prediction performance effectively. For example, the recent MEON model in [29] considered

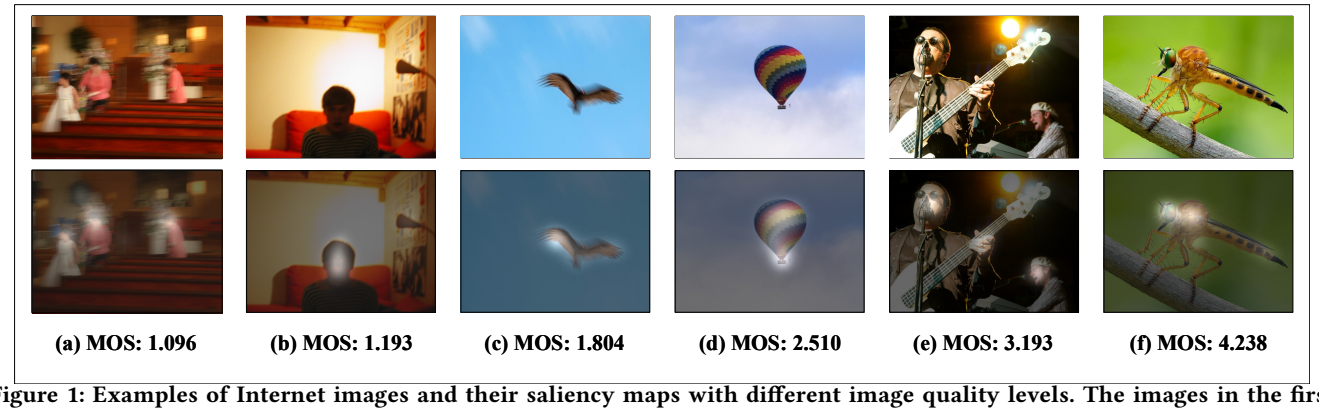

Figure 1: Examples of Internet images and their saliency maps with different image quality levels. The images in the first row are from KonIQ-10k dataset [25], and larger MOS (mean opinion score) shown in the bottom indicates better subjective perceptual quality. Their saliency maps in the second row are generated by a state-of-the-art saliency model [44] and fused with the original images where a pixel with brighter intensity indicates a higher probability of attracting human visual attention. Images (a)-(c) are obviously with low visual quality due to the severe distortions/blurs on the salient regions. Image (d) is slightly better yet still have distortions on its attended areas. The quality of image (e) is not as good as image (f) because the second most salient object (i.e., the man in the rear) is not that clear.

distortion identification as the auxiliary sub-task. This sub-task intends to make use of the distortion category information which is available in the legacy IQA datasets with some common synthetic distortions [20, 34, 37]. However, the massive quantity of Internet images captured by real cameras are usually afflicted by complex mixtures of multiple authentic distortions [6, 25], which cannot be well-simulated by the limited algorithm-generated distortions in the legacy IQA datasets. As a result, such a distortion identification sub-task cannot accurately identify the complex mixtures of distortions existing in authentically distorted images and may lead to performance degradation when applying to evaluate the real-world images with diverse authentic distortions.

To address the above-mentioned limitation of the existing multi-task deep CNN-based NR-IQA models, we propose to use visual saliency prediction as the auxiliary sub-task for providing more universal yet closely related perceptual information to facilitate quality prediction. Compared with the distortion identification sub-task, visual saliency always exists when viewing every image, regardless of its distortion type. More importantly, visual saliency has been reported to be closely related with image quality [47]. The rationale is that human beings tend to focus on visually salient areas while assessing image quality [7, 9, 39]. This inspires us to incorporate the visual saliency prediction as the auxiliary sub-task to learn a powerful multi-task end-to-end optimized deep CNN model for NR-IQA of authentically distorted images. Fig. 1 presents some Internet images along with the estimated saliency maps using a state-of-the-art saliency model [44].

The proposed saliency-guided deep neural network (SGDNet) inherits the same idea of other saliency-based NR-IQA methods [1, 9] by estimating saliency maps as a kind of local weighting functions to facilitate quality prediction, but it differs greatly from these two related works in two main aspects. On the one hand, their immediate visual saliency targets are optimized only with the single global quality scores as supervisions while these targets in our SGDNet have the direct supervisions, provided by a state-of-the-art saliency model [44]. On the other hand, the outputs of our saliency prediction sub-task are transparent to the quality prediction task by working as the spatial attention masks on the extracted features from the whole image for feature fusion. As a result, our method is an end-to-end image-based approach which avoids using the problematic local patch quality scores as labels in the training process.

The performance of our SGNet is evaluated on several public available benchmark datasets. The peer comparison results indicate that our SGDNet can achieve state-of-the-art performance on both authentically and synthetically distorted IQA datasets. Meanwhile, the ablation study shows that the quality prediction performance is indeed boosted by incorporating a saliency prediction sub-task and our multi-task learning framework can further improve the performance due to its learned adaptive spatial attention masks for feature fusion. Our contributions can be summarized as follows:

(1) We propose an end-to-end optimized SGDNet to solve the challenging NR-IQA task. To our best knowledge, it is the first attempt to optimize the saliency prediction and quality prediction sub-tasks together in an end-to-end multi-task learning framework. The proposed SGDNet is particularly suitable to blindly evaluate the visual quality of real-world images with authentic distortions.

(2) Our SGDNet is trained with more informative labels including saliency maps and global quality scores simultaneously for better quality prediction. More importantly, the outputs of the saliency prediction sub-task are transparent to the primary quality regression sub-task for a more perceptually-consistent feature fusion.

2 RELATED WORK

2.1 NR-IQA

According to the type of extracted features, NR-IQA methods can be roughly classified into three groups: NSS-based methods, feature learning-based methods, and CNN-based methods. NSS-based approaches are based on the assumption that distortion-free natural images have inherent statistical regularities and the presence of distortions in natural images will change such regularities. Complex statistical models for wavelet coefficients [32] or discrete cosine

transform coefficients [36] or locally normalized luminance coefficients [31] were developed for NSS modeling to extract quality-aware features. However, the NSS-based NR-IQA methods heavily depend on the domain knowledge on NSS modeling which is too complex to achieve a sufficient understanding.

Different from those NSS-based NR-IQA methods, the quality-aware features are automatically learned from data in the feature learning-based NR-IQA methods. In CORNIA [45] and HOSA [43], a codebook pre-learned in an unsupervised manner (i.e., K-means clustering) was used for feature encoding to generate quality-aware features directly from local normalized image patches. Experimental results have demonstrated the unique advantages of the feature learning-based NR-IQA methods as compared to the previous approaches using hand-crafted NSS features. However, with the advent of deep CNN, these feature learning-based methods were further surpassed by deep CNN-based solutions for NR-IQA.

Kang et al. [14] firstly implemented a shallow CNN model to extract features on small image patches for NR-IQA. The final quality score of an input image was computed by averaging the predictions of all patches cropped from it. However, this method and its successors suffer from the label noise problem caused by assigning the global MOS for all the patches cropped from the same input image. To partially represent the region-wise perceptual quality variation, Kim and Lee [16] proposed to use one of the classical FR-IQA methods to generate local quality scores for image patches as the local ground truth targets. However, such a strategy is not applicable to the real-world authentically distorted images which usually do not have the corresponding reference versions.

To learn a better feature representation, recent CNN-based NR-IQA methods proposed to use a multi-task learning framework. Kang et al. [15] further extended their work [14] to estimate image quality and distortion type simultaneously via a traditional multi-task CNN. Although these two sub-tasks are jointly optimized, there is no interaction between these two sub-tasks in their designed network. To address this problem, Ma et al. [29] proposed a new multi-task end-to-end optimized deep CNN network for NR-IQA. In their framework, the quality prediction sub-task depends on the outputs of the distortion identification sub-task. As such, the distortion type information was transparent to the primary quality prediction sub-network for better quality prediction. Our model also follows this pipeline where two sub-tasks are jointly optimized and have certain dependencies on each other. The main difference between ours and Ma's method is that we adopt saliency prediction to replace distortion identification as the complementary sub-task for providing more universal yet closely related perceptual information to support the primary quality prediction sub-task.

2.2 Visual Attention for NR-IQA

It is known that visual attention for different regions in an image is non-uniform, thereby a local visual importance measure has been considered in NR-IQA methods to capture such spatial attention variations for better quality prediction [1, 9]. Zhang et al. [50] utilized the classical visual saliency models to obtain the local weights (i.e., saliency map) of an image. These local weights were integrated into one of existing traditional IQA metrics to assess the quality of this image via the obtained weighted local quality map. With selecting an appropriate saliency model, these saliency-based IQA metrics outperform their original version significantly. However, even these saliency-based version of NR-IQA methods were still not comparable with those aforementioned CNN-based approaches.

Most of the CNN-based NR-IQA methods trained on local image patches simply assign the subjective quality score of an image to all the local patches cropped from it as their local quality label for training the network. This is problematic because local perceptual quality is not well-defined and not always consistent with the global quality score [29]. For this consideration, VIDGIQA [9] and WaDIQaM-NR [1] proposed to train the deep CNN models by jointly learning the visual importance and quality score of each local patch and then using the visual importance-weighted average local quality scores to estimate the global quality. However, these two local immediate regression targets are jointly optimized only with the single global quality scores as supervision. We argue that their newly introduced local visual importance weights are still not well-defined, as evidenced by there is no direct supervision for training these local immediate components.

Unlike the above methods, our SGDNet model is an image-based approach which generates feature maps from the whole input images instead of their local patches to avoid the potential problem of label noises. Moreover, we utilize the proxy saliency maps produced by a teacher saliency model [44] pretrained on large-scale saliency datasets to serve as the direct supervision of our saliency prediction sub-task.

3 PROPOSED SGDNET MODEL

3.1 Overview

We propose an end-to-end multi-task saliency-guided CNN model for NR-IQA. It consists of two sub-tasks including visual saliency prediction and image quality prediction which are jointly optimized with a shared feature extractor. In the literature, there are some IQA datasets [27, 35, 51] which provide saliency maps, in addition to the MOS (mean opinion scores), to investigate the interaction of visual attention and quality perception. However, the scale of these datasets is too small for training a powerful saliency prediction model. Considering that, our saliency prediction sub-network is trained with proxy saliency maps produced by a teacher saliency model [44], which is pre-trained on large-scale saliency datasets. To verify the effectiveness of these proxy saliency maps, we implement another saliency-guided CNN model, called Direct SGDNet, which directly use the proxy saliency maps as the additional model input to provide a kind of saliency guidance for boosting the accuracy of quality prediction. The architectures of these two models are depicted in Fig. 2. For these two models, either the proxy or learned saliency maps are served as the spatial attention masks to fuse the extracted features from the whole input images. As such, the visual importance of local regions over the whole image is modeled and a perceptually-consistent feature fusion is achieved.

3.2 Problem Formulation and Modelling

For an input image I, a NR-IQA model $\mathcal{M}$ is used to estimate the perceptual quality of this image Q_{est}:

$$Q_{est} = \mathcal{M}(I; \theta), \tag{1}$$

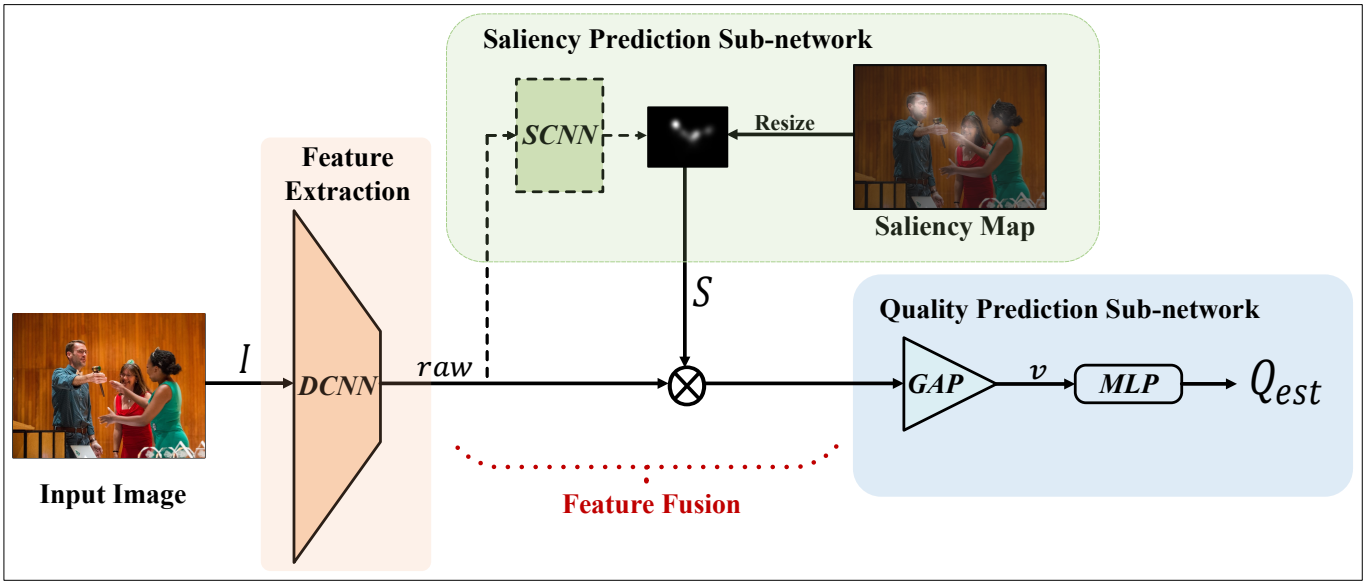

Figure 2: Architectures of two variants of proposed saliency-guided deep CNN models. (1) Direct SGDNet (without the saliency prediction sub-network indicated by dashed lines): directly use the saliency map, obtained by a saliency model [44], together with the input image to estimate its subjective quality. (2) (Multi-task) SGDNet: use a saliency prediction sub-network to predict saliency mask under the supervision of obtained one and then incorporate this learned mask with the extracted features to predict the image quality. Definitions of notations used in this figure are described in Sections 3.2 and 3.3.

where θ indicates all of the parameters of this model. Denote the ground truth quality of this image as Q_{gt}, the training target of this model $\mathcal{M}$ is to find the optimal parameter setting $\hat{\theta}$ so that the quality prediction loss $\mathcal{L}_q$ between the Q_{est} and Q_{gt} of all test images in the evaluated dataset is in its minimum. According to our validation experiments, we consider the $\ell 1$-norm instead of widely used $\ell 2$-norm as our $\mathcal{L}_q$:

$$\mathcal{L}_q = \frac{1}{N}\sum_{i=1}^{N} \| Q_{est,i} - Q_{gt,i} \|_1, \tag{2}$$

where the subscript i of $Q_{est,i}$ and $Q_{gt,i}$ represent the estimated quality and group truth quality of i-th image, respectively. Without loss of generality, we ignore this subscript in the following statements for simplification.

To be specific, we divide the pipeline of our end-to-end deep CNN-based NR-IQA model into several stages according to the change in feature dimension. Firstly, we use a Deep CNN (DCNN) as the feature extractor to get the raw CNN features f_{raw} directly from the input image I with a size of $h \times w \times 3$. In our implementation, we use one of two commonly used backbone networks, VGG16 [38] and ResNet50 [10]. Within these two backbone networks, their fully connected layers are discarded since we need the feature maps with spatial information for further feature fusion. The spatial dimension of these feature maps generated from the last layer of any of these two backbones is $\frac{h}{32} \times \frac{w}{32}$. To ensure the output channels of these feature maps are also the same, an extra $1 \times 1 \times 512$ convolutional layer is added, which can also perform the feature adaption, at the end of the backbone network. We treat this modified network as our feature extractor and use the $DCNN(\cdot)$ to denote this mapping: $\mathbb{R}^{h\times w\times 3} \rightarrow \mathbb{R}^{\frac{h}{32}\times\frac{w}{32}\times 512}$. Therefore, this feature extraction processing can be simply represented by:

$$\textbf{Feature Extraction:} \quad f_{raw} = DCNN(I;\theta_1), \tag{3}$$

where θ_1 indicates the weights of the convolutional layers within this DCNN. These raw features f_{raw} can be further processed to become more discriminative feature maps f_m by fusing the saliency information with them as our models, which will be detailed in the next section. As a comparison, we build our baseline model by direct feeding these raw features to the remaining quality prediction sub-network. It means $f_m = f_{raw}$ in this baseline model. End-to-end CNN-based NR-IQA models usually adopt multi-layer perceptron (MLP) to regress the image/patch features into a subjective quality score. However, feature maps f_m cannot be directly fed into MLP. A conversion from feature maps f_m to the image feature vector f_v is required. Here, we empirically choose the global average pooling (GAP) to perform this conversion: $\mathbb{R}^{\frac{h}{32}\times\frac{w}{32}\times 512} \rightarrow \mathbb{R}^{512}$. Our quality prediction sub-network can be written in the following simultaneous equations, where θ_2 indicates the parameters inside the MLP.

$$\textbf{Quality Prediction:} \quad \begin{cases} f_v = GAP(f_m), \\ Q_{est} = MLP(f_v;\theta_2). \end{cases} \tag{4}$$

3.3 Multi-task SGDNet

In the feature fusion stage, we propose to incorporate the saliency information into the raw extracted features for getting perceptually-consistent feature maps f_m. Considering that a saliency map is used to measure the visual importance of local regions within its corresponding input image, we implement this property by using the element-wise multiplication $\otimes$ to combine the raw features f_{raw} and saliency map S_m together, as shown in Eq. 5. As such, this saliency map will be treated as a spatial attention mask, where

the value of each pixel belongs to [0,1], to re-weight the raw features of the corresponding input image in its spatial domain for a perceptually-consistent feature fusion. Note that the spatial dimension of raw feature maps is not matched with the original size of generated saliency map, we need to resize the original saliency map with a down-sampling rate of 32 to make this multiplication meaningful.

$$\textbf{Feature Fusion:} \quad f_m = f_{raw} \otimes S_m. \tag{5}$$

The roles of these obtained saliency maps in our two variants of SGDNets are different. In our Direct SGDNet, saliency maps are directly used as one of the model inputs for providing additional guidance. By contrast, our Multi-task SGDNet takes these saliency maps as the immediate regression targets by using a shallow CNN (SCNN) as the saliency prediction sub-network to predict them, as shown in Fig. 2. In other words, both of these two variants of SGDNets are all using the S_m as the spatial attention mask. The difference between them is that S_m is learned before the feature fusion in the multi-task version or directly provided in the direct one. The processing of our saliency prediction sub-network can be represented by:

$$\textbf{Saliency Prediction:} \quad S_m = SCNN(f_{raw}; \theta_3), \tag{6}$$

where θ_3 indicates the parameters of this SCNN. For predicting visual saliency within this SCNN, a saliency prediction loss $\mathcal{L}_s$ should be taken into consideration to measure the gap between the predicted saliency mask and its corresponding proxy ground-truth. By viewing the saliency map as a kind of probability distribution [44], we adopt the total variation distance as this $\mathcal{L}_s$:

$$\mathcal{L}_s(x^p, x^g) = \frac{1}{2}\sum_{i=1}^{N} \left| \frac{x_i^p}{\sum_{i=1}^{N} x_i^p} - \frac{x_i^g}{\sum_{i=1}^{N} x_i^g} \right|, \tag{7}$$

where $x = (x_1, ..., x_i, ..., x_N)$ is the set of raw saliency response values for either the predicted saliency map (x^p) and the ground-truth saliency map (x^g). We have also tried other probability distribution distance metric, such as Kullback-Leibler divergence, as this saliency prediction loss and observed a similar performance. For jointly optimizing this multi-task SGDNet, an overall loss function should be defined. Here we simply use a linear combination of $\mathcal{L}_p$ and $\mathcal{L}_s$ to represent our optimization target of this SGDNet:

$$\hat{\theta} = \arg\min_{\theta}(\mathcal{L}_q + \alpha \mathcal{L}_s), \tag{8}$$

where $\theta = (\theta_1; \theta_2; \theta_3)$ in this model and α is a non-negative parameter to control the relative importance of the saliency prediction sub-task.

3.4 Spatial Attention and Channel-wise Attention

As introduced by [11], CNN extract features by fusing spatial and channel-wise information together, which means there are some inter-dependencies between the channels of CNN features. They explicitly model the channel-wise attention (CA) mechanism by their proposed Squeeze-and-Excitation (SE) blocks to selectively emphasize informative feature channels and suppress less useful ones. Since spatial attention and channel-wise attention are intended to re-calibrate the CNN features in the spatial and channel dimensions separately, we can further extend our model with the CA mechanism for a more comprehensive feature representation. Here, one SE block [11] is directly plugged into our framework for modeling CA. In this case, the previous feature fusion stage should be modified by:

$$\textbf{Feature Fusion with CA:} \quad \begin{cases} f_{CA} = SE(f_{raw}; \theta_4), \\ f_m = f_{CA} \otimes S_m. \end{cases} \tag{9}$$

where θ_4 represents the parameters within this SE block. To avoid unnecessary interactions between the spatial attention and channel-wise attention in our architecture, the raw features from the feature extractor are processed with these two attention mechanisms separately in a parallel manner. Our experimental results show that incorporating CA can improve the performance of our baseline model, but not as good as our proposed spatial attention.

4 EXPERIMENTS

4.1 Experimental Setups

4.1.1 IQA Datasets. We perform experiments on two types of image quality datasets covering the synthetic and authentic distortions, respectively. In the former type of IQA datasets, including LIVE [37], CSIQ [20], and TID2013 [34], the distorted images are generated by simulating a single type of synthetic distortion, such as JPEG compression, Gaussian blur, or white noise, with several distortion levels on the pristine images. However, numerous real-world images suffer from a mixture of diverse and complex authentic distortions which cannot be well-simulated by the above synthetic distortions. Therefore, newly emerged IQA datasets, including CLIVE [6] and KonIQ-10k [25], start to investigate the images with authentic distortions. In CLIVE, images are captured by a wide variety of mobile camera devices under highly diverse conditions. While in KonIQ-10k, images are sampled from the massive quantity of Internet images, and then filtered to ensure the content diversity and distortion authenticity. Both of the CLIVE and KonIQ-10k datasets have no reference images, and thus only NR-IQA methods can be used to evaluate them. For better readability, we provide a detailed information summary of the above IQA datasets in Table. 1.

4.1.2 Evaluation Criteria. Two commonly used evaluation metrics, Spearman rank order correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC), are adopted to benchmark the IQA models. Both metrics measure the correlation between a set of objective quality scores Q_{est} predicted by IQA algorithms and a set of subjective quality scores Q_{gt} provided by subjective experiments. Generally, SRCC measures the prediction monotonicity and PLCC measures the prediction accuracy. For both metrics, a higher value close to 1 indicates the higher performance of a specific IQA method.

4.1.3 Implementation Details. Our SGDNet model is implemented with Keras [3]. The backbone network of SGDNet is ResNet50. In our saliency prediction sub-network, the SCNN module is simply implemented by one 1×1 convolutional layer which can achieve the mapping from the raw features to one spatial attention mask. Moreover, the MLP applied in the remaining quality prediction sub-network consists of 3 hidden layers with neuron sizes 1024,

Table 1: Information Summary of IQA Datasets

Dataset	Year	# Reference images	# Distorted images	Distortion Type	# Distortion Types	Score type	Score range	Image Resolution
LIVE [37]	2006	29	779	synthetic	5	DMOS	[0,100]	mostly 512×768
CSIQ [20]	2009	30	866	synthetic	6	DMOS	[0,1]	512×512
TID2013 [34]	2013	25	3000	synthetic	24	MOS	[0,9]	384×512
CLIVE [6]	2016	N/A	1162	authentic	N/A	MOS	[0,100]	mostly 500×500
KonIQ-10k [25]	2017	N/A	10073	authentic	N/A	MOS	[1,5]	768×1024

1024, and 1. The α used in computing the overall loss is set to 0.25 for highlighting the importance of the quality prediction task.

During training, the weights in the backbone network are initialized from its Imagenet [5] pre-trained model. The weights of the remaining layers are initialized by the default setting of Keras. All of the models in our experiment are trained with the widely used Adam optimizer with an initial learning rate of 10^{-4}. This learning rate will be scaled down by a factor of 0.1 after every five epochs without validation loss decreasing. For each dataset, except KonIQ-10k, all of the images are resized to the dominant image resolution of this dataset for mini-batch training. Because of the limited GPU memory, the batch size varies according to the resolution of input images. It is worthy to mention that our SGDNet takes only 0.0202s for predicting the image quality of one input image of size 384×512 by using one single GTX 1080 Ti GPU. The source code of our SGDNet and its pre-trained models are publicly available[1].

4.2 Performance Comparison

4.2.1 Individual Dataset Evaluation. In this part, we conduct individual dataset evaluation on four IQA datasets, including LIVE [37], CSIQ [20], TID2013 [34], and CLIVE [6]. Among them, the first three datasets are for synthetic distortions while CLIVE is for authentic distortions. The KonIQ-10k dataset is not adopted in this experiment as it is not used by most of the compared methods. Following the experimental setting of [18], for each individual dataset, we randomly divide it into two subsets according to the reference images, 80% of the data for training and the remaining for testing. Then, the corresponding distorted images can be divided into two subsets with non-overlapping image contents. The SRCC and PLCC results of our method are averaged after ten repetitions of this random process.

In Table 2, we compare the performance of our SGDNet with six traditional NR-IQA methods, shown in the first six rows, and eight CNN-based NR-IQA methods, starting from the Kang's CNN [14] to the up-to-date DIQA [17]. For SRCC and PLCC scores on each dataset, the best and second-best models among these NR-IQA methods are highlighted in **boldface** and ***boldface italic***, respectively. From this table, we can conclude that our method achieves the state-of-the-art performance on both authentically and synthetically distorted IQA datasets.

From Table 2, we can observe that the CNN-based methods are generally superior to the traditional methods. Among those existing NR-IQA methods, except several ResNet50-based models, the performance results on the CLIVE dataset is much lower than on the other three synthetically distorted datasets. We interpret this phenomenon as follows. The real-world images with diverse authentic distortions in the CLIVE dataset have a much wider range of image contents which cannot be well handled by those traditional NR-IQA methods without using the deep CNN features. With the help of the ResNet50 backbone network for feature extraction and the incorporation of saliency information learned by our multi-task framework, our SGDNet largely outperforms all of the compared methods in this authentically distorted IQA dataset. Apart from the CLIVE dataset, our model also achieves competitive performance on the other three synthetically distorted IQA datasets, especially on TID2013. Our model is not as good as DIQA [17] on LIVE and CSIQ. The patch-based training strategy used in DIQA is more suitable for the evaluation on these two relatively small datasets with limited synthetic distortions but is not reliable for the evaluation on CLIVE.

4.2.2 Cross Dataset Evaluation. To test the generalization ability of our method, our proposed model is trained on the whole LIVE dataset and evaluated on the whole TID2013, CSIQ, and CLIVE datasets. Note that the score ranges and score types are not unified in these four datasets as shown in Table 1, we choose the settings of CSIQ as our standard in these experiments. Therefore, subjective scores on the other three datasets are linearly scaled to the range of [0,1]. For the MOS values in TID2013 and CLIVE, they are further reversed as $1 - MOS$ to meet this standard. Although our model can process input images with arbitrary sizes, we find that the cross dataset performance of our model is improved by resizing the images to 384×512. The results of the cross dataset evaluation are reported in Table 3 where we can observe the better generalization ability of our proposed SGDNet model.

4.3 Ablation Study

To evaluate the contribution of each component in our SGDNet, we conduct a series of ablation experiments on the KonIQ-10k IQA dataset [25]. KonIQ-10k is the current largest IQA dataset containing authentically distorted images and is suitable for evaluating the performance of different CNN-based NR-IQA models. For this dataset, we randomly pick 8073 out of the total 10073 images as the training set and the remaining images as the test set. All of the images and their original saliency maps are resized to 384×512 for accelerating the training process and using a relatively large batch size (20 in this case) for training. The major ablation results are presented in Table 4. Apart from the introduced SRCC and PLCC, MAE, i.e., $\ell 1$-norm, is also used as another evaluation metric, as it can represent the goodness of fit of these models in this table. Different from the SRCC and PLCC metrics, the lower MAE value indicates a better performance.

4.3.1 Influence of the backbone network. In this ablation experiment, we implemented our model with two different backbone

[1] https://github.com/ysyscool/SGDNet

Table 2: Performance comparison on four individual datasets. In each column, the best and second-best results are highlighted in boldface and boldface italic, respectively.

Methods	LIVE [37]		CSIQ [20]		TID2013 [34]		CLIVE [6]	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BLIINDS-II [36]	0.912	0.916	0.780	0.832	0.536	0.628	0.463	0.507
DIIVINE [32]	0.925	0.923	0.835	0.817	0.549	0.654	0.509	0.558
BRISQUE [31]	0.939	0.942	0.775	0.781	0.572	0.651	0.607	0.645
IL-NIQE [48]	0.902	0.908	0.821	0.865	0.521	0.648	0.594	0.589
CORNIA [45]	0.942	0.943	0.714	0.781	0.549	0.613	0.618	0.662
HOSA [43]	0.948	0.949	0.781	0.841	0.688	0.764	0.659	0.678
Kang's CNN [14]	0.956	0.956	-	-	-	-	-	-
BIECON [16]	0.961	0.962	0.825	0.838	0.721	0.765	0.595	0.613
DIQaM-NR [1]	0.960	0.972	-	-	***0.835***	***0.855***	0.606	0.601
WaDIQaM-NR [1]	0.954	0.963	-	-	0.761	0.787	0.671	0.680
VIDGIQA [9]	***0.969***	***0.973***	-	-	-	-	0.701	-
ResNet50 + fine-tuning [18]	0.950	0.954	0.876	***0.905***	0.712	0.756	***0.819***	***0.849***
Imagewise CNN [18]	0.963	0.964	0.812	0.791	0.800	0.802	0.663	0.705
DIQA [17]	**0.975**	**0.977**	**0.884**	**0.915**	0.825	0.850	0.703	0.704
SGDNet (ours)	***0.969***	0.965	***0.883***	0.903	**0.843**	**0.861**	**0.851**	**0.872**

Table 3: SRCC results in the cross dataset evaluation. In each column, the top result is highlighted in boldface.

Methods	CSIQ [20]	TID2013 [34]	CLIVE [6]
BLIINDS-II [36]	0.577	0.393	0.119
DIIVINE [32]	0.590	0.355	**0.465**
BRISQUE [31]	0.548	0.358	0.313
CORNIA [45]	0.649	0.360	0.443
VIDGIQA [9]	0.641	0.415	0.315
DIQaM-NR [1]	0.681	0.392	-
WaDIQaM-NR [1]	0.704	0.462	-
SGDNet (ours)	**0.719**	**0.532**	0.455

networks including VGG16 [38] and ResNet50 [10]. It is widely accepted that ResNet50 is more powerful than VGG16, as it can be trained very deeply for more comprehensive feature extraction with the help of residual learning. In our case, we compare the models listed in Table 4 with different backbone networks. In our baseline models, some VGG-based models have better performance with regard to SRCC or PLCC. The reasons may be that SRCC and PLCC are not always consistent with our quality prediction loss. In our proposed saliency-guided models, the ResNet-based models perform better than their VGG-based versions on all of the three evaluation metrics.

4.3.2 Effectiveness of saliency information. We incorporate the saliency information in two different ways. In our Direct SGDNet, the obtained saliency maps are served as one of the model inputs. Extracted features from the backbone network are further fused with these fixed saliency maps. As a comparison, our SGDNet is built on an end-to-end multi-task framework and these saliency maps are the target outputs of our proposed saliency prediction sub-network. The predicted saliency maps of our sub-network have the same function as the fixed ones for providing the spatial attention masks. From Table 4, we can observe that the quality prediction

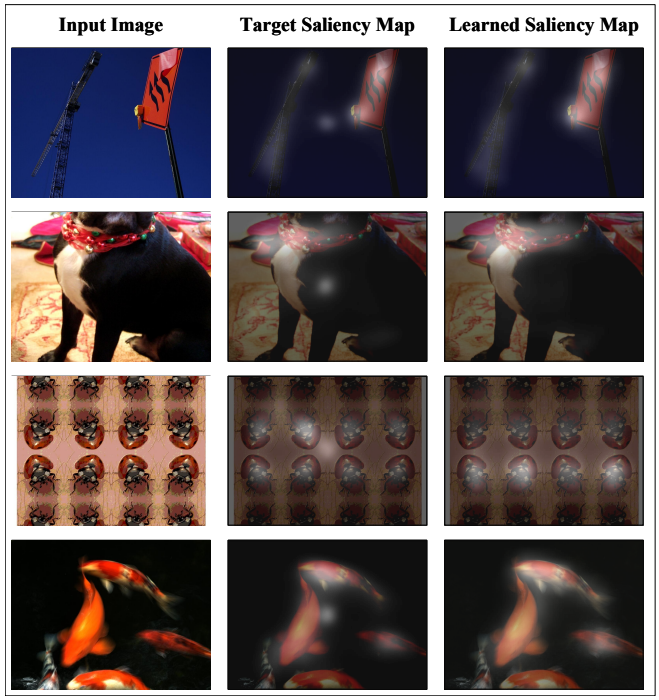

Figure 3: Examples of input images, target saliency maps generated by [44], and predicted saliency maps by our proposed saliency prediction sub-network.

performance indeed boosted by incorporating the saliency information. Moreover, by comparing the Direct SGDNet and SGDNet with the same configuration of backbone and CA, we can find the latter model is always better. It means that our end-to-end multi-task learning framework can further improve the boosted performance achieved by simply using the saliency information as the additional model input.

Table 4: Model ablation analysis on KonIQ-10k IQA dataset [25]

Model Type	Backbone Network		Saliency Information		CA	Performance		
	VGG16	ResNet50	saliency input	saliency output		SRCC ↑	PLCC ↑	MAE ↓
Baseline	✓					0.817	0.822	0.2714
		✓				0.808	0.816	0.2532
Baseline + CA	✓				✓	0.827	0.853	0.2623
		✓			✓	0.833	0.851	0.2370
Direct SGDNet	✓		✓			0.843	0.874	0.2017
		✓	✓			0.869	0.890	0.1938
Direct SGDNet + CA	✓		✓		✓	0.851	0.880	0.1963
		✓	✓		✓	0.880	0.899	0.1830
SGDNet	✓			✓		0.870	0.890	0.1881
		✓		✓		0.897	0.917	0.1684
SGDNet + CA	✓			✓	✓	0.878	0.896	0.1846
		✓		✓	✓	0.903	0.920	0.1639

Fig. 3 provides some examples to compare those target and learned saliency maps. Saliency maps generated by [44] have some undesirable responses on the image centers because this saliency model is trained on the saliency prediction datasets where center-bias priors are commonly considered. Our saliency prediction sub-network can skip this trap by jointly training with the quality prediction sub-network while the center regions in these input images are not appealing to the later quality prediction sub-network. Therefore, we conclude that the saliency maps learned in SGDNet are more adaptive and suitable than those fixed ones.

4.3.3 Influence of channel-wise attention. Besides modeling the spatial attention by using the saliency information, our method can be further extended by incorporating the channel-wise attention (CA) module, as described in Section 3.4. From Table 4, in the baseline models, the introduced CA module [11] can improve the performance by a large margin (more than 3.7% on PLCC and 1.2% on SRCC). However, we also observe that the spatial attention mask is more powerful than CA by comparing the models in the second and third type in Table 4. It can be explained that incorporating saliency information can lead to a more perceptually-consistent feature fusion while incorporating CA can only explicitly modeling the inter-dependencies between feature channels which is not much related to the quality prediction task. Furthermore, by fusing the spatial and channel-wise features together, our extended SGDNet (SGDNet + CA) can achieve slightly higher results when comparing with the original one.

4.3.4 Influence of the saliency model. Intuitively, the performance of our SGDNet is also affected by the teacher saliency model, which produces the proxy saliency maps for training our saliency prediction sub-network. To explore the influence of this fact, we test other three state-of-the-art saliency models, including SAM [4], SalGAN [33], and DVA [41], to replace the currently used DINet [44]. Since PLCC is the common evaluation metric on both quality and saliency prediction, we report the PLCC results of different saliency models with our ResNet50-based SGDNet on KonIQ-10k IQA dataset [25] and their own performance on MIT300 saliency benchmark dataset [2] in Table 5. By comparing the results in the second column, we can find that the effectiveness of saliency information is still validated by using different saliency models. Moreover, the performance of our SGDNet has a positive correlation with the performance of the used saliency models in MIT300 saliency benchmark dataset. Specifically, using SAM, SalGAN, or DVA to replace the DINet will lead to the performance degradations by 0.3%, 0.6%, 1.5%, respectively. It can be foreseen that our SGDNet can be further improved with the advent of more powerful saliency models.

Table 5: PLCC results of Saliency models on KonIQ-10k IQA dataset [25] with our SGDNet and MIT300 saliency dataset [2] with themselves.

Saliency models	KonIQ-10k	MIT300
DINet [44]	0.917	0.79
SAM [4]	0.914	0.78
SalGAN [33]	0.911	0.73
DVA [41]	0.903	0.68
(no saliency model)	0.816	-

5 CONCLUSION

In this work, we have proposed a novel saliency-guided deep neural network (SGDNet) for no-reference image quality assessment (NR-IQA). The whole model is built on an end-to-end multi-task learning framework where two sub-tasks including visual saliency prediction and image quality prediction are jointly optimized and have certain dependencies on each other. The effectiveness of incorporating saliency information and our multi-task framework for CNN-based NRI-IQA has been validated by a series of ablation studies. Moreover, our method overcomes the inability of existing multi-task CNN-based NR-IQA methods on predicting the quality of real-world images with authentic distortions. Experimental results on both authentically and synthetically distorted IQA datasets have demonstrated the outstanding performance of our model with respect to other relevant NR-IQA methods. In the future, we will consider a new spatial attention mechanism that can learn the spatial attention maps from the image itself instead of using a proxy saliency map as supervision.

6 ACKNOWLEDGMENT

This research was partially supported by Singapore Ministry of Education Tier-2 Fund MOE2016-T2-2-057(S).

REFERENCES

[1] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. 2018. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing* 27, 1 (2018), 206–219.

[2] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. [n. d.]. MIT Saliency Benchmark.

[3] François Chollet et al. 2015. Keras. https://keras.io.

[4] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing* 27, 10 (2018), 5142–5154.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[6] Deepti Ghadiyaram and Alan C Bovik. 2016. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing* 25, 1 (2016), 372–387.

[7] Ke Gu, Shiqi Wang, Huan Yang, Weisi Lin, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. 2016. Saliency-guided quality assessment of screen content images. *IEEE Transactions on Multimedia* 18, 6 (2016), 1098–1110.

[8] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. 2015. Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia* 17, 1 (2015), 50–63.

[9] Jingwei Guan, Shuai Yi, Xingyu Zeng, Wai-Kuen Cham, and Xiaogang Wang. 2017. Visual importance and distortion guided deep image quality assessment framework. *IEEE Transactions on Multimedia* 19, 11 (2017), 2505–2520.

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

[11] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[12] Qiuping Jiang, Feng Shao, Wei Gao, Zhuo Chen, Gangyi Jiang, and Yo-Sung Ho. 2018. Unified No-Reference Quality Assessment of Singly and Multiply Distorted Stereoscopic Images. *IEEE Transactions on Image Processing* 28, 4 (2018), 1866–1881.

[13] Qiuping Jiang, Feng Shao, Weisi Lin, Ke Gu, Gangyi Jiang, and Huifang Sun. 2018. Optimizing multistage discriminative dictionaries for blind image quality assessment. *IEEE Transactions on Multimedia* 20, 8 (2018), 2035–2048.

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

[15] Le Kang, Peng Ye, Yi Li, and David Doermann. 2015. Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In *2015 IEEE international conference on image processing (ICIP)*. IEEE, 2791–2795.

[16] Jongyoo Kim and Sanghoon Lee. 2017. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing* 11, 1 (2017), 206–220.

[17] Jongyoo Kim, Anh-Duc Nguyen, and Sanghoon Lee. 2018. Deep CNN-based blind image quality predictor. *IEEE transactions on neural networks and learning systems* 99 (2018), 1–14.

[18] Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan C Bovik. 2017. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Processing Magazine* 34, 6 (2017), 130–141.

[19] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. 2016. A comparative study for single image blind deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1701–1709.

[20] Eric Cooper Larson and Damon Michael Chandler. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging* 19, 1 (2010), 011006.

[21] Dingquan Li, Tingting Jiang, and Ming Jiang. 2017. Exploiting high-level semantics for no-reference image quality assessment of realistic blur images. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 378–386.

[22] Leida Li, Ya Yan, Yuming Fang, Shiqi Wang, Lu Tang, and Jiansheng Qian. 2016. Perceptual quality evaluation for image defocus deblurring. *Signal Processing: Image Communication* 48 (2016), 81–91.

[23] Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, and David Zhang. 2018. Learning convolutional networks for content-weighted image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3214–3223.

[24] Qiaohong Li, Weisi Lin, Jingtao Xu, and Yuming Fang. 2016. Blind image quality assessment using statistical structural and luminance features. *IEEE Transactions on Multimedia* 18, 12 (2016), 2457–2469.

[25] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. 2018. KonIQ-10K: Towards an ecologically valid and large-scale IQA database. *arXiv preprint arXiv:1803.08489* (2018).

[26] Anmin Liu, Weisi Lin, and Manish Narwaria. 2012. Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing* 21, 4 (2012), 1500–1512.

[27] Hantao Liu and Ingrid Heynderickx. 2009. Studying the added value of visual attention in objective image quality metrics based on eye movement data. In *2009 16th IEEE international conference on image processing (ICIP)*. IEEE, 3097–3100.

[28] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. 2017. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding* 158 (2017), 1–16.

[29] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. 2018. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing* 27, 3 (2018), 1202–1213.

[30] Xiongkuo Min, Ke Gu, Guangtao Zhai, Menghan Hu, and Xiaokang Yang. 2018. Saliency-induced reduced-reference quality index for natural scene and screen content images. *Signal Processing* 145 (2018), 127–136.

[31] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing* 21, 12 (2012), 4695–4708.

[32] Anush Krishna Moorthy and Alan Conrad Bovik. 2011. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing* 20, 12 (2011), 3350–3364.

[33] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. 2017. SalGAN: Visual Saliency Prediction with Generative Adversarial Networks. *arXiv preprint arXiv:1701.01081* (2017).

[34] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication* 30 (2015), 57–77.

[35] Judith Redi, Hantao Liu, Rodolfo Zunino, and Ingrid Heynderickx. 2011. Interactions of visual attention and quality perception. In *Human Vision and Electronic Imaging XVI*, Vol. 7865. International Society for Optics and Photonics, 78650S.

[36] Michele A Saad, Alan C Bovik, and Christophe Charrier. 2012. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE transactions on Image Processing* 21, 8 (2012), 3339–3352.

[37] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing* 15, 11 (2006), 3440–3451.

[38] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[39] Cuong T Vu, Eric C Larson, and Damon M Chandler. 2008. Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience. In *2008 IEEE Southwest Symposium on Image Analysis and Interpretation*. IEEE, 73–76.

[40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[41] Jianbing Shen Wenguan Wang. 2018. Deep Visual Attention Prediction. *IEEE Transactions on Image Processing* (2018).

[42] Jinjian Wu, Weisi Lin, Guangming Shi, and Anmin Liu. 2013. Reduced-reference image quality assessment with visual information fidelity. *IEEE Transactions on Multimedia* 15, 7 (2013), 1700–1705.

[43] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. 2016. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing* 25, 9 (2016), 4444–4457.

[44] Sheng Yang, Guosheng Lin, Qiuping Jiang, and Weisi Lin. 2019. A Dilated Inception Network for Visual Saliency Prediction. *arXiv preprint arXiv:1904.03571* (2019).

[45] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. 2012. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1098–1105.

[46] Guangtao Zhai, Xiaolin Wu, Xiaokang Yang, Weisi Lin, and Wenjun Zhang. 2012. A psychovisual quality metric in free-energy principle. *IEEE Transactions on Image Processing* 21, 1 (2012), 41–52.

[47] Lin Zhang, Ying Shen, and Hongyu Li. 2014. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing* 23, 10 (2014), 4270–4281.

[48] Lin Zhang, Lei Zhang, and Alan C Bovik. 2015. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing* 24, 8 (2015), 2579–2591.

[49] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. 2011. FSIM: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing* 20, 8 (2011), 2378–2386.

[50] Wei Zhang, Ali Borji, Zhou Wang, Patrick Le Callet, and Hantao Liu. 2015. The application of visual saliency models in objective image quality assessment: A statistical evaluation. *IEEE transactions on neural networks and learning systems* 27, 6 (2015), 1266–1278.

[51] Wei Zhang and Hantao Liu. 2017. Toward a reliable collection of eye-tracking data for image quality research: challenges, solutions, and applications. *IEEE Transactions on Image Processing* 26, 5 (2017), 2424–2437.