ENCODING DISTORTIONS FOR MULTI-TASK FULL-REFERENCE IMAGE QUALITY ASSESSMENT

Chen Huang[1,2,4]*, Tingting Jiang*[2,3]*, Ming Jiang*[1,2]

[1] LMAM, School of Mathematical Sciences & BICMR
[2] Cooperative Medianet Innovation Center, Peking University, China
[3] NELVT, School of EECS, Peking University, China, [4] Baidu Inc.

ABSTRACT

Most existing image quality assessment models focus on evaluating the image quality score, however, the quality score alone is not enough to characterize the degeneration. In this paper, we propose a full reference framework named Mask Gated Convolutional Network (MGCN) for evaluating the image quality score and identifying distortions simultaneously. Observing the fact that the reference images are distorted by various distortions in pixel space, we design an encoder module to capture the transformation between reference images and distorted images as low level features. Further higher level features are extracted from the low level features and shared by both the regression and the classification tasks. Instead of simply cropping patches to augment data, we mask the high level feature map in the spatial domain to randomly sample patches from the image and learn to assign the image quality score to the patch set. The proposed method achieves the state-of-the-art performance on LIVE2, TID2008 and TID2013 datasets.

Index Terms— image quality assessment, gated convolutional network

1. INTRODUCTION

Generally, objective image quality assessment (IQA) can be divided into three categories according to the availability of reference images: full-reference (FR), reduced-reference (RR) and no-reference (NR). This paper focuses on FR IQA which has full access to the whole reference image.

In all existing FR databases such as [1, 2, 3], the reference images are usually distorted by different distortions at different levels. Image quality is closely related to the distortion type. The degeneration arises from the specific distortion and the image quality depends on the distortion type and level. However, neither the image quality score nor the distortion type alone is enough to describe the degeneration from the reference images to the distorted images. On the one hand, images with different types of distortions may have similar image quality scores; on the other hand, images with the same distortion may have different quality scores. So it's more appropriate to characterize the distorted images in terms of both the distortion type and the image quality score. Unfortunately, all existing FR methods only evaluate the image quality ignoring the distortion type as far as we know. Only a few NR methods try to evaluate the image quality considering the information of the distortion type.

In this work, our goal is to design a FR deep network for evaluating the image quality score and identifying distortions simultaneously. Different from previous works which take image patches as inputs, this network is designed to take both the reference image and distorted image as inputs directly. To achieve this goal, there are two questions to answer: 1) How to extract features from reference images and distorted images, which are correlated to both image quality and distortion type? 2) How to deal with the insufficiency of training data because the size of current FR-IQA datasets is limited.

For the first question, some prior works use Siamese network [4] with two branches sharing parameters. It can be viewed as hierarchical feature extraction followed by similarity measures. Different strategies have been applied to merge two information flows on the top of the Siamese network for IQA. Then the merged feature is passed on to the subsequent procedure. The simplest way of merging is to concatenate the output features of the two branches directly where it allows great flexibility but requires lots of parameters, e.g. [5]. Another way of merging is to bind the output features of the two branches together assuming the alignment between the two features is known implicitly. For example, taking the difference of two features is one special case under such assumption. In this work, we propose to fuse the information of reference images and the distorted images earlier. Instead of extracting image features for each image independently followed by merging, we capture the transformation between one reference image and its distorted image, and then use the transformation feature to predict the image quality. Because reference images are transformed in raw pixel space by distortions such as Gaussian blur, the transformations between reference images and distorted images relate to both the distortion type and the image quality score closely. This motivates us to design the dedicated module to capture the transforma-

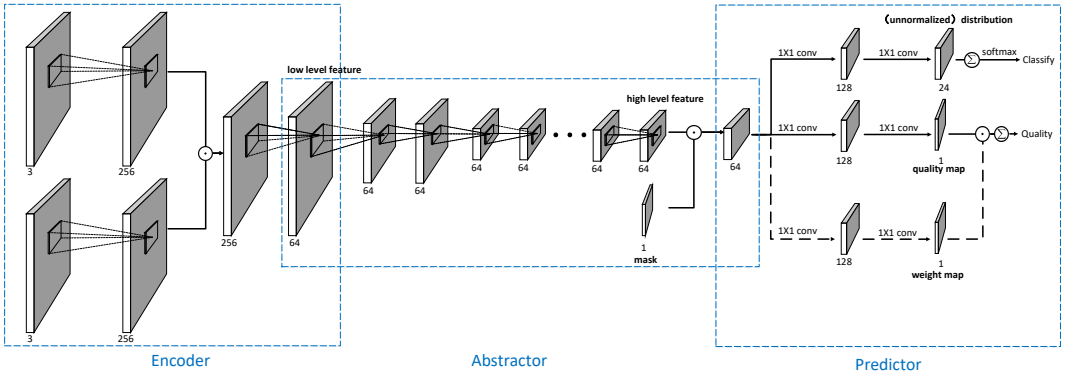

Fig. 1. Architecture of MGCN. $\odot$ indicates the Hadamard product. The model takes a distorted image and a reference image as inputs. The encoder module learns to encode the transformation between images as the low level feature. The abstractor module further extracts the high level feature. The predictor module evaluates the image quality score and identifies distortions simultaneously. The image quality score equals to the average or the weighted average of patch quality scores.

tion. In [6, 7], the gated auto-encoder (GAE) demonstrates the ability to delineate not only the combination of multiple affine transformation such as translation, rotation and scaling, but also the parent-offspring relationship. We conjecture that GAE can work with not only affine transformations but also complex transformations. Inspired by the concept of the GAE, we design a gated encoder using multiplicative interactions between pixels to encode the transformation between images. The encoder is pre-trained by unsupervised learning within the GAE, which alleviates over-fitting and shares the ability to represent transformation. To our best knowledge, this is the first time that GAE is introduced to IQA.

For the second question, most IQA methods augment datasets by randomly cropping patches from the image, and assigning the image quality score to the patches. However, this limits the performance because of: 1) Inconsistency. It is questionable to equate the patch quality score with the image quality score, especially when the patch size is much less than the image size or the distortion is inhomogeneous. 2) Redundancy. Multiple patches with overlap from one image lead to redundant computation. To address the issue of inconsistency, we can treat an image as a set of patches rather than a single patch, and learn to assign the image quality score to the patch set. As for the redundancy, inspired by dropout [8], we extract high level features from the image in a convolutional way, and mask the high level feature map in the spatial domain. This process is equivalent to randomly sampling multiple patches and sharing feature computation among multiple patches. Then we can combine the quality scores of multiple patches to predict the quality score of the image.

Fig. 1 shows our network named as Mask Gated Convolutional Network (MGCN). Our contributions are summarized as follows. (1) This is the first time that GAE is introduced to IQA. We design the dedicated encoder module to capture the distortion feature between reference images and distorted images. (2) A deep model is proposed for multi-task learning in FR IQA to fully characterize the distorted images. As far as we know, this is the first paper which predicts the image quality score and distortion type simultaneously in FR. (3) We resolve the inconsistency between the patch quality score and the image quality score, and improve efficiency by masking the high level feature map.

2. RELATED WORK

DCNN [5] is an early work using deep learning in FR IQA. It uses images which have similar scene with distorted images as reference to evaluate the image quality score. FR IQA is the special case where "similar images" are distortion free images. They crop 224×224 patches to augment the dataset. Within the framework of Siamese network, they concatenate features extracted from two patches and predict the patch quality score. The subsequent models [9, 10] work in a similar way in feature fusion between reference images and distorted images, but differ in the weighting phase. DeepQA [9] crops patches and learns a weight map to weight the MSE map between the reference patch and the distorted patch. WaDIQaM-FR [10] crops 32 patches , then predicts the patch quality score and the patch weight. The image quality score is the weighted sum of the quality scores of multiple patches.

Multi-task learning is widely used in deep learning, however only a few NR methods consider the distortion. [11] first pre-trains a sub-network for distortion identification, then end-to-end fine-tunes the quality prediction sub-network. [12] uses a convolution network to evaluate the image quality score and classify the distortion simultaneously.

3. MASK GATED CONVOLUTIONAL NETWORK

The proposed MGCN contains three modules as shown in Fig. 1: the encoder module, the abstractor module and the predictor module. The network takes the distorted image and its reference image as inputs, then predicts the distortion and the image quality score. The loss is constructed to contain

both classification loss and regression loss:

$$L_{\text{MGCN}} = L_{\text{quality}}(q, \hat{q}) + \lambda L_{\text{classify}}(c, \hat{c}) \tag{1}$$

where L_{quality} is the mean square error between the predicted quality score $\hat{q}$ and the ground-truth score q of the image, and L_{classify} is the cross entropy between the predicted distortion type $\hat{c}$ and the ground-truth type c of the image.

3.1. Encoder

To perform relational feature learning on tri-partite networks, [6, 13] propose the multiplicative interactions between three variables $z_k = \sum_{ij} w_{ijk} x_i y_j$. This can be seen as variable x modulating the parameter $W = (w_{ijk})$ that connects variable z and variable y or vice versa. In the case where x, y correspond to the reference image and the distorted image respectively, we can think of the pixels in one image as gating the parameters of a standard feature learning model applied to another image. As [6] claims, learning the factored gated feature learning model has the ability to find appropriate sets of filter pairs that can detect the rotations in feature space. So the hidden variable z may be viewed as the representation of various distortions and the gated model learns to capture the transformation between reference and distorted images.

The GAE works in a way similar to a standard auto-encoder but the model parameter W is a linear function of x or y. The encoding and decoding are written as follows:

$$w_{ijk} = \sum_{f=1}^{F} w_{if}^{x} w_{jf}^{y} w_{kf}^{z} \tag{2}$$

$$z_k = \sigma\left(\sum_f w_{kf}^z (\sum_i w_{if}^x x_i)(\sum_j w_{jf}^y y_j)\right) \tag{3}$$

$$\hat{x}_j(z, y) = \sum_f w_{jf}^x (\sum_i w_{if}^y y_i)(\sum_k w_{kf}^z z_k) \tag{4}$$

$$\hat{y}_j(z, x) = \sum_f w_{jf}^y (\sum_i w_{if}^x x_i)(\sum_k w_{kf}^z z_k) \tag{5}$$

where F is the number of the units in the factor layer.

During the pre-training, we feed the GAE on 5×5 patch pairs sampled from reference images and distorted images. By minimizing the symmetric reconstruction cost

$$L_{\text{GAE}} = \frac{1}{N} \sum_n^N \sum_j \left[(y_j^{(n)} - \hat{y}_j^{(n)})^2 + (x_j^{(n)} - \hat{x}_j^{(n)})^2 \right] \tag{6}$$

where N is the number of patch pairs. We expect that the GAE learns to capture the transformation between images.

In MGCN, the encoder module is initialized by the encoder in the GAE. Note that, the original gated auto-encoder [6, 13] is designed as fully connected layers operating on vectors. To work on images, we design the encoder module taking the form of convolutional layers. The gated encoder shares the ability of the GAE capturing the transformation and benefits from the pre-train by unsupervised learning. It merges two branches and outputs the transformation feature for subsequent learning process.

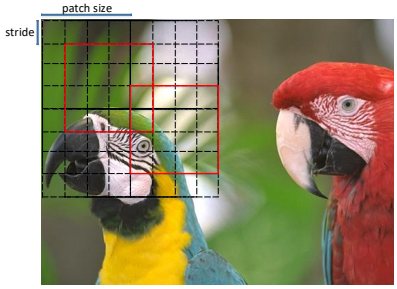

Fig. 2. In the view of sampling, the abstractor module first uniformly samples 196×196 patches, and randomly discards fix ratio of patches. The remained patches are colored in red.

Although motivated by [6], we make great improvement to adapt the GAE with the FR-IQA problem. [6] explores to represent correlation patterns across multiple images using multiplicative interactions. It focuses on the theoretical analysis of the role that multiplicative interactions play in learning to encode relations. It only validates the representation power by illustrating some toy examples and simple affine transformations. In our work, we validate the representation power in complex situations, and illustrate that the GAE can not only learn the distortions between reference and distorted images but also reconstruct the reference and distorted images simultaneously. We further reform the GAE and integrate it into a deep convolution model so that it can be trained end-to-end.

3.2. Abstractor

The abstractor module is designed as a convolutional network, where convolution layers alternate with pooling layers. The abstractor module maps the low level features encoded by the encoder module into high level features shared by the classification and the regression.

Focusing on the high level feature map in Fig. 1, we observe that each unit of size $(1 \times 1 \times \text{channels})$ summaries the information of a patch in inputs. The patch corresponds to the perceptual field in the image. Hence, the convolution and pooling layers work as a sampler which samples patches uniformly in the image. The patch size and the sample interval depend on the pooling size and stride. See Fig. 2 for illustration. By introducing the mask code which takes 1 with probability p otherwise 0, we mask the high level feature map in the spatial domain. This is equivalent to randomly sampling a number of patches without redundant computation on the overlap. Consider n patches with size $s \times s$ sampled from the image with size $r \times c$, we reduce the computation complexity from $O(ns^2)$ to $O(rc)$ by feature sharing. In our experiment, the perceptual field s is set to 192, and the p is adjusted according to the image size in training phase so that we sample

about 48 patches. Hence, we improve the efficiency about 8 times in the case of the image of size 384×512.

In this way, we can augment datasets implicitly without inconsistency by sampling different patch sets from one image as long as the patch size and the number of patches are large enough. Note that p is set to 1.0 in test phase so that all patches are taken into account for the prediction without randomness. Compared with dropout which is typically interpreted as bagging a large number of models sharing parameters and used to avoid over-fitting e.g. WaDIQaM-FR, the mask in MGCN aims to resolve the inconsistency between the patch quality score and the image quality score, and to avoid the redundant computation.

3.3. Predictor

The high level features of multiple patches are not only shared in computation, but also in distortion classification and quality regression. The predictor module receives the masked high level feature map, then identifies the distortion and predicts the quality score simultaneously as illustrated in Fig. 1. To classify the distortion, the predictor first evaluates the (unnormalized) distribution of all distortion types for each sampled patch, then averages these distributions over the patches, finally we apply softmax to obtain the normalized distribution of all distortion types for the image. As for the regression, the predictor evaluates the quality score for each sampled patch, and combine them to predict the image quality score. More specifically, the image quality score equals to the average or the weighted average of patch quality scores. Learning the predictor module can be viewed as a way of learning to represent the image with a set of patches within it.

4. EXPERIMENTS

We evaluate the performance of MGCN on three well-known FR IQA datasets, LIVE2 [1], TID2008 [2], TID2013 [3].

4.1. Training strategy and experiment protocol

Some prior works e.g. [18, 5] preprocess images by local contrast normalization. This preprocessing hinders the assessment for distortion types such as intensity shift and contrast variation. According to [6], the training of the GAE requires that both inputs are contrast-normalized. So we normalize the input in the same way as [6]. During the pre-training, we employ Adam [19] to train GAE for 200 epoches. The learning rate starts with 0.002. The pre-trained GAE is used to initialize the encoder module in MGCN. In the training phase, we fine-tune the MGCN end-to-end with learning rate 10^{-4} for 300 epoches. we simply augment data by flipping images horizontally without any other augmentation. Batch normalization technique [20] is applied to accelerate convergence.

To quantize the performance, we choose three metrics including Spearman Rank-Order Correlation Coefficient (SROCC), Pearson Linear Correlation Coefficient (PLCC), and classification accuracy.

MGCN are compared with seven FR IQA methods including four non-learning based models (PSNR, MS-SSIM [14], FSIMc [15], VSI [16]), three learning based models (DOG-SSIMc [17], WaDIQaM-FR [10], DeepQA [9]), and two NR IQA methods including IQA-CNN++ [12], MEON [11]. Here, MGCN series contains two subclasses: MGCN-ave averages patch quality scores to predict the image quality score and MGCN-weight calculates the weighted average. λ is set as 2.0 for TID2008/TID2013 and 5.0 for LIVE2 in the multi-task learning framework according to the validation set.

In experiments, we use Monte-Carlo cross validation, 60% data for training, 20% data for validation, and 20% data for test. The reference images together with the respective distorted versions are randomly split into three disjoint groups. Specially there are 17 reference images for training, 6 reference images for validation, and 6 reference images for test in LIVE2. And there are 15 reference images for training, 5 reference images for validation, 5 reference images for test in TID2008/TID2013. We repeat experiments 10 times and report the median for each metric. Tab. 1 summarizes the performance of different models in three datasets. It's worth mentioning that the data settings and the experiment protocols are different in prior works. For fair comparison, we run all FR models in the same data setting. For fair comparison between learning based models and non-learning based models, we evaluate the performance of the non-learning based models including PSNR, MS-SSIM, FSIMc, VSI only on the test set. We train DeepQA using the codes released by the author and train WaDIQaM-FR[1] using the codes implemented by ourselves in the following experiments.

4.2. Multi-task learning

From Tab. 1, we can see that MGCN ($\lambda = 0$) which focuses on evaluating the image quality achieves the state of the art performance in quality assessment, but guesses the type of the distortion randomly. In the multi-task learning framework, MGCN-ave/MGCN-weight improve the classification accuracy dramatically and reach approximate performance in quality prediction compared with the corresponding single task MGCN. This validates the feasibility of MGCN in the multi-task learning framework. Compared with NR methods for multi-task learning, MGCN surpasses IQA-CNN++ and MEON across-the-aboard in TID2008 as expected but wins by a narrow margin in LIVE2. This may be due to the data preprocessing in LIVE2 where we resize all images into the same size 512x640. Remember that MGCN operates on the images but other models take patches as input. So MGCN suffers from the distortion introduced by resizing in LIVE2. This may explain why MGCN is worse in terms of distortion identification on the simple database LIVE2 than on TID2008/TID2013 where all images have the same size.

When our model is trained with only one label, the results are shown in Tab. 2 and Tab. 1. The negative sign in

[1]The author only release the trained models without training codes.

Table 1. Performance comparison of different models. The best results are highlighted in bold. "-" indicates that the results are not reported in the original paper or the method is not able to identify the distortion. "*" indicates that the metric is evaluated on the subset. We repeat all FR methods in the same data setting and cite the results of NR methods from the original papers. Note IQA-CNN++ reports the results of TID2008 on 13 distortions. MEON reports the results of TID2013 on 4 distortions.

Class	Models	LIVE2			TID2008			TID2013		
		SROCC	PLCC	Accuracy	SROCC	PLCC	Accuracy	SROCC	PLCC	Accuracy
NR	IQA-CNN++ [12]	0.950	0.950	0.951	0.870*	0.880*	0.929*	-	-	-
	MEON [11]	-	-	-	-	-	-	0.912*	0.912*	0.940*
FR	PSNR	0.905	0.883	-	0.563	0.585	-	0.660	0.696	-
	MS-SSIM [14]	0.953	0.759	-	0.855	0.857	-	0.784	0.834	-
	FSIMc [15]	0.965	0.859	-	0.879	0.879	-	0.850	0.878	-
	VSI [16]	0.956	0.757	-	0.902	0.882	-	0.892	0.900	-
	DOG-SSIMc [17]	0.957	0.958	-	0.888	0.896	-	0.857	0.878	-
FR-DL	WaDIQaM-FR [10]	0.960	0.963	-	0.916	0.925	-	0.918	0.924	-
	DeepQA [9]	**0.977**	**0.981**	-	0.887	0.894	-	0.883	0.893	-
MGCN	MGCN-ave ($\lambda = 0$)	0.966	0.965	0.194	0.923	0.920	0.056	0.924	0.916	0.038
	MGCN-ave	0.969	0.969	0.950	0.933	0.925	0.985	0.932	0.929	**0.979**
	MGCN-weight ($\lambda = 0$)	0.971	0.971	0.196	**0.945**	**0.941**	0.055	**0.940**	**0.946**	0.043
	MGCN-weight	0.966	0.967	**0.958**	0.940	0.937	**0.988**	0.934	0.942	0.972

Table 2. Performance of MGCN-weight trained only with the label of the distortion type.

Dataset	SROCC	PLCC	Accuracy
LIVE2	-0.288	-0.139	0.950
TID2018	-0.170	-0.223	0.985
TID2013	-0.106	-0.074	0.968

SROCC/PLCC indicates that the prediction are contrary to the ground-truth. When we train the network only with the single label (the distortion types/the image quality scores), the other branch in the predictor module is not trained at all, and its output is meaningless.

4.3. Quality prediction

In LIVE2, most models achieve satisfactory results and are comparable with each other, due to the limited numbers of the distortion types, levels and reference images.

In TID2008/TID2013, MGCN and WaDIQaM-FR surpass other methods and report the state-of-the-art performance. As stated in the experiment protocol, TID2008/TID2013 has more data with more distortion types and levels than LIVE2. Traditional or shallow learning based methods have limited capacity for modeling data distribution confined by the models themselves. They don't benefit from the increasement of dataset scale but deteriorate due to the complexity of the dataset. As for the deep learning models which explicitly augment data by assigning image quality scores to patch quality scores, the benefit of the increasement of dataset scale is limited because it introduces more inconsistency between image quality scores and patch quality scores simultaneously. Instead, MGCN and WaDIQaM-FR can capture data distribution more accurately and gain the full boost of performance steming from the increasement of training data without inconsistency.

Compared with WaDIQaM-FR, MGCN benefits from two points: 1) WaDIQaM-FR uses 10 convolutional layers to extract the high level feature for both the reference images and the distorted images independently, and MGCN uses 4 convolutional layers to extract the high level feature from the encoded low level feature. MGCN has less parameters and is more resistant to over-fitting. 2) WaDIQaM-FR fuses the high level feature which may lose details related to the image quality while MGCN encodes the transformation of two images in the pixel space.

4.4. Ablation experiments and efficiency analysis

To quantify the contributions of the encoder module and the mask separately, we perform the ablation experiments on TID2008/TID2013. The results are shown in Tab. 3. To test the contribution of the encoder module, we randomly initialize the gated encoder, and directly train MGCN end-to-end without pre-training the GAE by unsupervised learning. Other settings are kept the same. The results are shown in the first row "no pretrain". As for the mask, we set p to 1.0 in training phase, which indicates that all patches are kept to represent the image. The results are shown in the second row "no mask". The last row "pretrain + mask" indicates the performance of the complete MGCN-weight. According to the results, the encoder module with unsupervised learning contributes more to MGCN than the mask, which proves the representation power and effectiveness of extracting distortion features by the GAE module for IQA.

Table 3. Ablation experiments.

Dataset	Setting	SROCC	PLCC	Accuracy
TID2008	no pretrain	0.891	0.863	0.978
	no mask	0.915	0.913	0.971
	pretrain + mask	0.940	0.937	0.988
TID2013	no pretrain	0.883	0.887	0.946
	no mask	0.924	0.931	0.971
	pretrain + mask	0.934	0.942	0.972

Table 4. Time and memory cost on TID2013.

Model	Train	Epochs	Test	Memory
DeepQA [9]	1.89h	80	0.010s	448MB
WaDIQaM-FR [10]	45.83h	3000	0.019s	4390MB
MGCN-weight	29.87h	300	0.122s	5058MB

Tab. 4 shows the time and memory cost of different models on TID2013. The epoch numbers of DeepQA and WaDIQaM-FR are from the original paper. "Train" and "Test" refer to GPU time for training and testing one image respectively, and "Memory" refers to GPU memory.

5. CONCLUSION

In this paper, we propose a full reference framework for evaluating the image quality score and identifying distortions simultaneously. MGCN makes use of both reference and distorted images, and addresses the issue of over-fitting by narrowing the gap between image quality and patch quality. Comprehensive experiments are conducted to demonstrate the effectiveness of the proposed method.

Acknowledgment This work was partially supported by National Basic Research Program of China (973 Program) under contract 2015CB351803 and the NSFC under contracts 61572042, 61527804. We also acknowledge the high-performance computing platform of Peking University for providing computational resources.

6. REFERENCES

[1] Hamid Rahim Sheikh, Muhammad Farooq Sabir, and Alan Conrad Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *TIP*, vol. 15, no. 11, pp. 3440–3451, 2006.

[2] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti, "TID2008-a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[3] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C. C. J. Kuo, "Color image database TID2013: Peculiarities and preliminary results," in *Proceedings of the 4th European Workshop on Visual Information Processing*, June 2013, pp. 106–111.

[4] Jane Bromley, James W. Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah, "Signature verification using a "siamese" time delay neural network," *IJPRAI*, vol. 7, no. 4, pp. 669–688, 1993.

[5] Yudong Liang, Jinjun Wang, Xingyu Wan, Yihong Gong, and Nanning Zheng, "Image quality assessment using similar scene as reference," in *ECCV*. Springer, 2016, pp. 3–18.

[6] Roland Memisevic, "Learning to relate images," *PAMI*, vol. 35, no. 8, pp. 1829–1846, 2013.

[7] Afshin Dehghan, Enrique G Ortiz, Ruben Villegas, and Mubarak Shah, "Who do I look like? determining parent-offspring resemblance via gated autoencoders," in *CVPR*, 2014, pp. 1757–1764.

[8] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting.," *JMLR*, vol. 15, no. 1, pp. 1929–1958, 2014.

[9] Jongyoo Kim and Sanghoon Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *CVPR*, 2017.

[10] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek, "Neural network-based full-reference image quality assessment," in *PCS*. IEEE, 2016, pp. 1–5.

[11] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo, "End-to-end blind image quality assessment using deep neural networks," *TIP*, vol. 27, no. 3, pp. 1202–1213, 2018.

[12] Le Kang, Peng Ye, Yi Li, and David Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *ICIP*, 2015, pp. 2791–2795.

[13] Roland Memisevic, "Gradient-based learning of higher-order image features," in *ICCV*, 2011, pp. 1591–1598.

[14] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers*. IEEE, 2003, vol. 2, pp. 1398–1402.

[15] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang, "FSIM: A feature similarity index for image quality assessment," *TIP*, vol. 20, no. 8, pp. 2378–2386, 2011.

[16] Lin Zhang, Ying Shen, and Hongyu Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *TIP*, vol. 23, no. 10, pp. 4270–4281, 2014.

[17] Soo-Chang Pei and Li-Heng Chen, "Image quality assessment using human visual DOG model fused with random forest," *TIP*, vol. 24, no. 11, pp. 3282–3292, 2015.

[18] Le Kang, Peng Ye, Yi Li, and David Doermann, "Convolutional neural networks for no-reference image quality assessment," in *CVPR*, 2014, pp. 1733–1740.

[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.