MULTITASK ATTENTIVE NETWORK FOR TEXT EFFECTS QUALITY ASSESSMENT

*Keqiang Yan, Shuai Yang, Wenjing Wang and Jiaying Liu**
Wangxuan Institute of Computer Technology, Peking University, Beijing, China

ABSTRACT

Along with the fast development of image style transfer, large amounts of style transfer algorithms were proposed. However, not enough attention has been paid to assess the quality of stylized images, which is of great value in allowing users to efficiently search for high quality images as well as guiding the designing of style transfer algorithms. In this paper, we focus on artistic text stylization and build a novel deep neural network equipped with multitask learning and attention mechanism for text effects quality assessment. We first select stylized images from TE141K [1] dataset and then collect the corresponding visual scores from users. Then through multitask learning, the network learns to extract features related to both style and content information. Furthermore, we employ an attention module to simulate the process of human high-level visual judgement. Experimental results demonstrate the superiority of our network in achieving a high judgement accuracy over the state-of-the-art methods. Our project website is available at `https://ykq98.github.io/projects/TEA/`.

***Index Terms*—** Image quality assessment, multitask, attention, text effects, style transfer

1. INTRODUCTION

With the fast development of image style transfer [2], a series of algorithms [2–11] were proposed, followed by an increasing number of stylized pictures on the Internet. While offering more diverse options, users are also faced with the problem of spending much time filtering high quality images. To tackle this problem, researches have been devoted to Image Quality Assessment (IQA) [12–21], which aims to automatically predict the perceptual quality of an image. With the help of a network that can characterize the perceptions of human beings, images of high visual quality can be quickly and easily found, eliminating the trouble of manual selecting. However, existing IQA networks are mostly designed for natural images with degradations such as noises, compression artifacts and blurring [17, 22, 23], which fail to characterize the stylish images, and consequently has poor performance on stylized image assessment. Besides, human eyes are especially sensitive to semantic structure of glyphs. Therefore, compared with ordinary image stylization, the task of assessing text effect transfer images is more challenging.

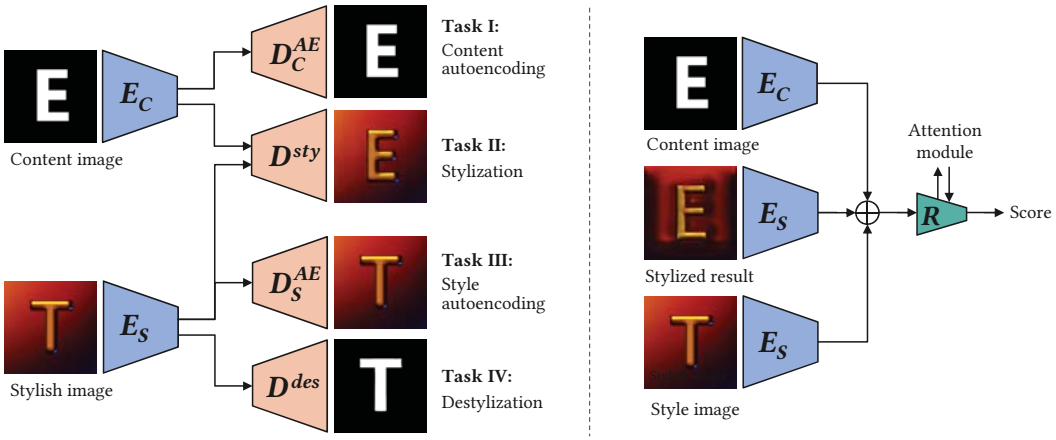
In this work, we focus on the style of text effects, and present a novel network for text effects quality assessment. The challenges of predicting the perceptual quality of text effects transfer results lie in three aspects: 1) Text effects transfer has only recently begun to be studied and therefore, there is a lack of large-scale corresponding datasets with user mean opinion scores (MOS). 2) Different from the conventional evaluation standard of IQA systems, the matching in style and content should also be taken into account. Moreover, style, as a high-level concept of human subjective cognition, is difficult to quantitatively evaluate. 3) The focus of human visual assessment may vary a lot for different text effects images. For example, when the background is relatively simple, users will pay more attention to the foreground text effects. Therefore, visual attention should be considered.

IQA has been studied a lot, which can be categorized into classic methods [18, 22] and deep-based methods [12–14, 19, 24]. The main idea of classic methods is to manually design some features as evaluating indicators based on the domain knowledge. The values of these features are first calculated and organized as feature vectors, which are then used to fit the ground truth MOS in subjective evaluation. However, the artificial features are hard to design, especially for stylish images. Meanwhile, deep-based methods directly train neural networks to fit the ground truth MOS and can be further divided into non-reference and reference. The former uses the target image as input while the latter additionally requires a reference ground truth image. However, neither of these is suitable for our problem, since they fail to consider the matching of style and content images.

In this paper, we propose a multitask attentive network for text effects quality assessment. To train a network for accurate prediction of perceptual quality of text effects, we first select images from TE141K dataset [1], where $151,224$ text effects images are produced by 16 different style transfer models, and then we collect $28,001$ aesthetic opinion scores from invited users. Based on TE141K dataset, we train our network to accomplish multiple tasks of style autoencoding, content autoencoding, stylization and destylization, through which our network learns to characterize the robust style features and content features. These features empower our pre-

* Corresponding author
This work is partially supported by National Natural Science Foundation of China under contract No.61772043, in part by Beijing Natural Science Foundation under contract No.L182002.

978-1-7281-1331-9/20/$31.00



(a) Multitask Feature Extraction (b) Attentive Quality Prediction

Fig. 1: Framework of the proposed multitask attentive network for text effects quality assessment.

diction network to find the relationship between the input stylized image and its corresponding reference style/content images, which helps to prevent over-fitting problems. Furthermore, we incorporate visual attention modules into our prediction network to adaptively localize the most important regions in the target image. By giving high weights to the feature maps in these regions, the network can learn to focus on key positions and details, thus better simulating the human visual decision process.

In summary, the contributions of this work are threefold:

- We explore a new issue of text effect assessment for estimating the quality of images generated by text effect transfer models. To address this problem, we select images from TE141K dataset and collect 28k aesthetic opinion scores.
- We propose a novel multitask network for no reference text effect assessment. The tasks of text effect reconstruction, destylization and stylization make the network better in extracting image features.
- We propose an attentive network to simulate the process of human high-level visual judgement, paying more attention to interested areas, which shows excellent performance on the task of Text Effects Quality Assessment.

2. MULTITASK ATTENTIVE NETWORK FOR TEXT EFFECT ASSESSMENT

In this section, we describe our multitask attentive network to predict the perceptual quality y of text effects transfer result x based on its content image c and style image s. As shown in Figure 1, we first pretrain encoder-decoder-based networks on the tasks of style autoencoding, content autoencoding, stylization and destylization (Section 2.1), in the aim of making our encoders learn to extract robust style and content features. We then train our attentive prediction network where the encoders are further finetuned to predict the mean opinion scores (MOS) (Section 2.2). Detailed network architecture can be found in the supplementary material.

2.1. Multitask Feature Extraction

For multitask feature extraction, we build multitask network composed of the style encoder $E_{\mathcal{S}}$, the content encoder $E_{\mathcal{C}}$, the style decoder $D_{\mathcal{S}}^{AE}$, the content decoder $D_{\mathcal{C}}^{AE}$, the stylization decoder D^{sty} and the destylization decoder D^{des}. $E_{\mathcal{S}}$ and $E_{\mathcal{C}}$ extract the style features and content features from the stylish images (*i.e.*, style images and stylization results) and content images, respectively. To ensure the features extracted well characterize the key information for the perceptual assessment, we train the multitask network to accomplish the following tasks.

First of all, given a set of content text images $\{c_i|i = 1, 2, ..., M\}$, the encoded content feature of c_i is required to preserve the core information of the glyph in c_i. Therefore, we impose a reconstruction constraint that forces the content feature to fully reconstruct the input content image, leading to the standard L_2 loss on the autoencoder $D_{\mathcal{C}}^{AE} \circ E_{\mathcal{C}}$:

$$\mathcal{L}_{\mathcal{C}}^{AE} = \mathbb{E}_i \left[\| D_{\mathcal{C}}^{AE}(E_{\mathcal{C}}(c_i)) - c_i \|^2 \right]. \quad (1)$$

Similar to the content autoencoder, we sample stylish images containing results of the style transfer methods and ground truth style transfer reference images from the dataset and impose L_2 loss on the style autoencoder $D_{\mathcal{S}}^{AE} \circ E_{\mathcal{S}}$. Given N different kinds of text effects styles and K style transfer models, let s_{ij} ($i = 1, 2, ..., M$ and $j = 1, 2, ..., N$) be the ground truth style image with the style j and the glyph of c_i. Let x_{ij}^k denote the style transfer result of model k conditioned on the content image c_i and the style j. Then, the loss for the

Fig. 2: Representative examples of different MOS in TE141K dataset. From top to bottom: reference style image and content image, text effects transfer results with their average quality scores in the bottom left.

style autoencoder can be written as,

$$\begin{aligned}\mathcal{L}_{\mathcal{S}}^{AE} &= \mathbb{E}_{i,j}\left[\|D_{\mathcal{S}}^{AE}(E_{\mathcal{S}}(s_{ij})) - s_{ij}\|^2\right] \\ &+\lambda_1 \mathbb{E}_{i,j,k}\left[\|D_{\mathcal{S}}^{AE}(E_{\mathcal{S}}(x_{ij}^k)) - x_{ij}^k\|^2\right],\end{aligned} \tag{2}$$

where λ_1 is a weight that compromises between the reconstruction error of s and that of x.

Considering the matching of the glyph is an important factor in determining the perceptual quality of text effects transfer results, we force the style encoder to pay attention to the glyph in the stylish image. To this end, $E_{\mathcal{S}}$ is accomplished with D^{des} to form a destylization network to obtain the glyph information from the stylization result by removing its style elements,

$$\mathcal{L}^{des} = \mathbb{E}_{i,j,k}\left[\|D_{\mathcal{S}}^{des}(E_{\mathcal{S}}(x_{ij}^k)) - c_i\|^2\right]. \tag{3}$$

For the same reason, the matching of the glyph is taken into account throught a stylization network $D^{STY} \circ (E_{\mathcal{C}}, E_{\mathcal{S}})$, where the content and style features are extracted and combined to yield a new style transfer result,

$$\mathcal{L}^{sty} = \mathbb{E}_{i\neq l,j}\left[\|D^{sty}(E_{\mathcal{C}}(c_i), E_{\mathcal{S}}(s_{lj})) - s_{ij}\|^2\right]. \tag{4}$$

Finally, the objective function of our multitask feature extraction can be defined as

$$\mathcal{L}^{multi} = \lambda_{\mathcal{C}}^{AE}\mathcal{L}_{\mathcal{C}}^{AE} + \lambda_{\mathcal{S}}^{AE}\mathcal{L}_{\mathcal{S}}^{AE} + \lambda^{des}\mathcal{L}^{des} + \lambda^{sty}\mathcal{L}^{sty}, \tag{5}$$

where the losses of four tasks are weighted by $\lambda_{\mathcal{C}}^{AE}$, $\lambda_{\mathcal{S}}^{AE}$, λ^{des} and λ^{sty}, respectively.

2.2. Attentive Quality Prediction

Visual attention models have been applied to localizing interested regions in an image to capture its features. The idea is naturally consistent with human's behavior of assessing image quality. People tend to give evaluations based on some key regions such as a region with evident checkerboard artifacts or color deviation rather than the whole images. It is especially evident for image stylization since the style tends to be completely differently rendered in different regions of an image, making the regions that have a significant impact on ratings unevenly distributed. Therefore we consider visual attention to be important for text effects quality assessment, which allows the prediction network to know where the critical regions should be focused on.

Specifically, we employ the Convolutional Block Attention Module (CBAM) [25] into our prediction network. Our prediction network is composed by the proposed content and style encoders $E_{\mathcal{C}}$ and $E_{\mathcal{S}}$, and a three-layer convolutional score regression network R, with CBAM between its first and second convolutional layers. For a stylization result x_{ij}^k, its corresponding reference c_i and s_j, and its ground truth MOS y_{ij}^k in our dataset, $E_{\mathcal{C}}$ first extracts the content feature from c_i. Meanwhile $E_{\mathcal{S}}$ extracts the style features of x_{ij}^k and s_j. These three features are concatenated and fed into R. Within R, the intermediate feature $\mathbf{f}$ are weighted by the attention map calculated by the CBAM as

$$\begin{aligned}\mathbf{f}' &= M_c(\mathbf{f}) \otimes \mathbf{f}, \\ \mathbf{f}'' &= M_s(\mathbf{f}') \otimes \mathbf{f}',\end{aligned} \tag{6}$$

where M_c and M_s are 1D channel and 2D channel attention maps inferred by CBAM, respectively, and $\otimes$ denotes element-wise multiplication. $\mathbf{f}''$ is the final refined feature, which is used to predict the final scores:

$$\mathcal{L}^R = \mathbb{E}_{i,j,k}\left[\left(R(E_{\mathcal{C}}(c_i), E_{\mathcal{S}}(s_{ij}), E_{\mathcal{S}}(x_{ij}^k)) - y_{ij}^k\right)^2\right]. \tag{7}$$

3. DATA COLLECTION

To train our multitask attentive network, we use TE141K dataset which contains stylized images of 16 stylization methods and collect the corresponding MOS from the invited users. In this paper, we label these 16 methods as #1 - #16. For the details of these 16 methods and their MOS scores, please refer to the supplementary material. In this section, we will introduce the details of our data collection.

For user evaluation, we select 20 styles, which is the subset named TE141K-S of the TE141K dataset. For each model, we sample four stylization results of different content references from each style, adding up to 1,280 images. Then we conduct user studies where 135 invited observers were shown these 1,280 images with their corresponding content and style reference images, and were asked to score 1 to 5 on the image visual quality by taking both content consistency and style conformity into consideration. We first gave the users a tutorial about the score labeling, and then show them the query images with corresponding content and style references. As shown in Fig. 2, user score of 1 is of lowest quality and 5

Table 1: Performance evaluation on the TEA dataset in terms of SRCC.

Method	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	**All**
NIMA [21]	-0.06	0.49	-0.12	-0.16	0.79	0.39	0.02	-0.15	0.02	0.12	0.38	0.02	0.13	0.24	0.27	0.49	0.38
CNNIQA [12]	0.64	0.07	0.26	-0.02	0.41	0.29	0.14	0.52	0.12	-0.14	-0.03	0.67	-0.37	0.42	0.52	0.54	0.61
DeepBIQ [26]	0.22	0.59	0.22	0.53	0.53	0.33	0.14	0.19	0.43	0.30	-0.05	0.53	0.12	0.62	0.53	0.61	0.64
Ours	0.55	0.57	0.28	0.41	0.67	0.59	0.17	0.14	0.26	0.34	0.61	0.31	0.42	0.34	0.67	0.47	**0.68**

Table 2: Performance evaluation on the TEA dataset in terms of PLCC.

Method	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	**All**
NIMA [21]	-0.13	0.42	-0.09	0.02	0.74	0.28	0.08	-0.11	0.05	0.14	0.35	0.12	0.12	0.25	0.57	0.58	0.39
CNNIQA [12]	0.66	-0.04	0.22	0.15	0.38	0.35	0.06	0.44	0.16	-0.14	0.12	0.61	-0.24	0.26	0.47	0.41	0.63
DeepBIQ [26]	0.26	0.57	0.25	0.60	0.47	0.40	0.10	-0.10	0.45	0.32	-0.08	0.63	0.16	0.61	0.49	0.60	0.65
Ours	0.54	0.55	0.38	0.50	0.69	0.54	0.14	0.19	0.23	0.30	0.55	0.49	0.38	0.50	0.38	0.48	**0.67**

Table 3: Ablation study in terms of SRCC, PLCC and accuracy (%).

Method	SRCC	PLCC	Accuracy
Baseline	0.67	**0.67**	72.02
Baseline + AT	0.66	0.65	73.39
Baseline + MT	**0.68**	**0.67**	72.94
Full model	**0.68**	**0.67**	**75.23**

is of highest quality. Image with lower score suffers from mismatching of reference content image and reference style image. And even if the image is visualy pleasing, the users tend to give a low score of 2 because of the mismathing of the content as shown in Fig. 2. We end up with 28,001 opinion scores, where each image has around 22 scores in average. The final ground truth MOS of each image is calculated as the mean of the scores obtained.

4. EXPERIMENTAL RESULTS

4.1. Implementation Details

We adapt our multitask network from the basic encoder-decoder architecture, which contains three convolutional layers. After each convolutional layer we utilize batch normalization layer [27] except the first layer. The architecture of our attentive quality prediction model contains two blocks. Each block has one convolutional layer which downsamples the feature maps to half of their original size, one batch normalization layer and a leaky relu layer. We first resize the image to the size of 300×300 and then randomly crop the images to 256×256 for training. Data augmentation of random rotation ($\pm$ 20 degree) and flip is then performed. The Adam optimizer is adopted with the fixed learning rate of 0.00005 and batch size of 1. We select 668 images for training and 167 for validation and 436 images for testing. For all experiments, we set $\lambda_1 = \lambda_{\mathcal{C}}^{AE} = 0.25$, $\lambda^{sty} = \lambda_{\mathcal{S}}^{AE} = 1$, and $\lambda^{des} = 1.25$.

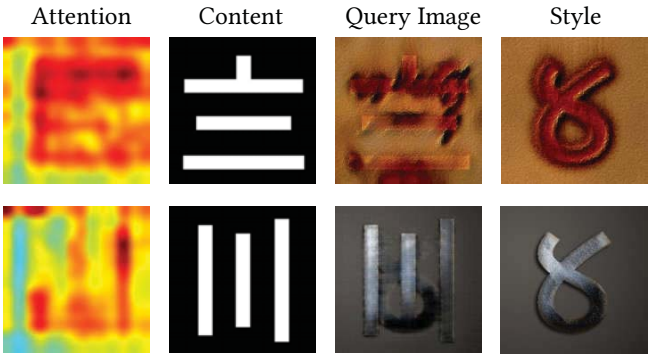


Fig. 3: Representative visualiztaion examples of the pixel-wise attention map. The attention maps have been upsampled to 256×256 from the original size of 16×16. The first line shows our attentive network pays attention to the matching of the content and notices the dismatching of the style and the second line shows it can find out the style inconsistency in the background.

4.2. Comparison of Previous IQA Methods

Quantitative evaluations. Correlation and accuracy values of our evaluations on the IQA models on the collected data are presented in Table 1 and Table 2, respectively. We report the evaluation performance for 16 style transfer models in Supplementary material and the overall performance on all these models. As shown in Table 2, NIMA [21], CNNIQA [12] and DeepBIQ [26] have 0.38 SRCC with 0.39 PLCC, 0.61 SRCC with 0.63 PLCC and 0.64 SRCC with 0.65 PLCC, respectively. By comparison, our method obtains 0.68 SRCC with 0.67 PLCC, outperforming the state-of-the-art methods.

Visual evaluations. In Fig. 4, we further show examples of representative MOS for visual evaluation. Fig. 4 suggests that our network could well characterize key factors such as

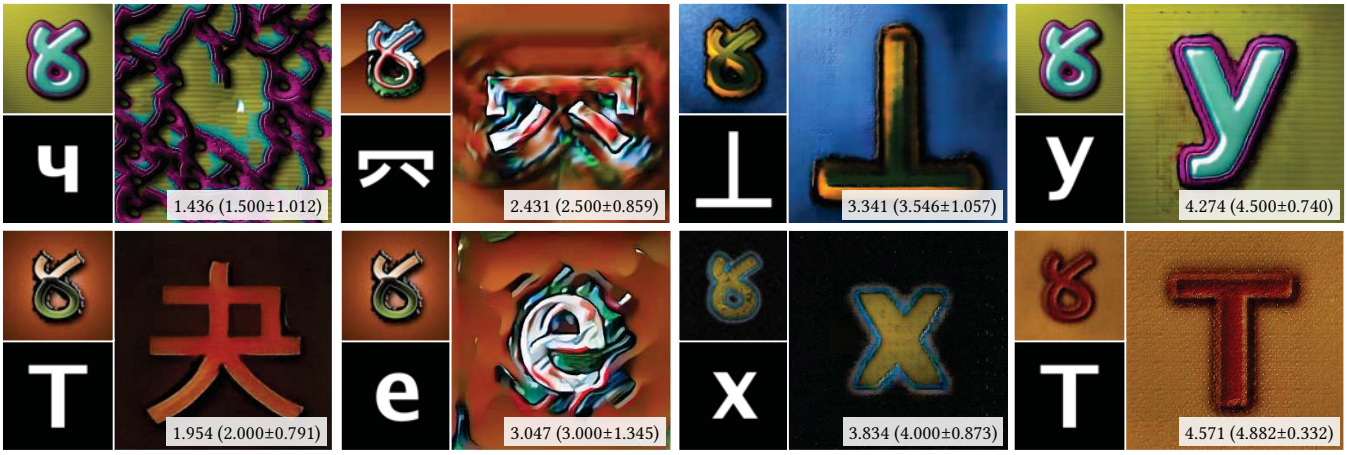


Fig. 4: Our predicted scores on examples of representative MOS of 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, and 4.9, respectively. The quality scores are shown at the bottom right of each image in form of y $(\mu \pm \sigma)$, where y is our predicted score, μ is the mean opinion score with its standard error σ.

Fig. 5: Ranking some examples from TE141K without labels using the proposed network. Predicted scores are shown at the bottom right of each image.

the matching of style, the matching of glyph, bleeding artifacts and other artifacts in determining aesthetic qualities. For more results of our proposed network, readers may refer to supplementary material.

4.3. Ablation Study

Multitask pretraining. In Table 3, we study the effect of the multitask pretraining (Section 2.1). For accuracy calculation, we follow the setting of NIMA [21] to report the accuracy evaluations of two-class quality categorization, where predicted scores are compared to 3 as cut-off score. Comparing to Baseline, Baseline+MT (multitask) brings a gain of 0.9% in accuracy. Comparing to our full model, without multitask pretraining, the accuracy drops by 1.84%. The high judgement accuracy with multitask pretraining verifies that our multitask attentive network effectively extracts robust style features and content features, which makes it easier to find the relationship between the input stylized image and its corresponding reference style/content images.

Attention module. In Table 3, we examine the effects of our attention module through a comparative experiment. Comparing to Baseline+AT (attention), the judgement accuracy of the baseline without attention module drops by 1.37%, which means our network do learn to detect and focalize critical regions of the image as human do and predict more accurate scores. To visualize the effect of attention module, in Fig 3, we show some examples of the input images and their corresponding pixel-wise attention maps. By utilizing the attention module, our network tends to pay more attention to the critical parts of the given image.

4.4. Text Effects Ranking

Predicted scores can be used to rank text effects. Some stylization results from TE141K dataset without ground truth MOS are ranked in Fig. 5. Although not labelled, our network gives reasonable ranking of these images that matches human visual judgement, which could potentially benefit the development of text effects recommendation system.

5. CONCLUSION AND FUTURE WORK

In this paper, we present a multitask attentive network for text effects quality assessment. The network is trained on multiple tasks of content autoencoding, style autoencoding, stylization and destylization to learn to extract robust content and style features. An attentive module is exploited to empower the network to focus on the regions that are critical for the perceptual assessment. To train out network, we propose a text effects assessment dataset where more than 28,001 user scores are collected. We validate the effectiveness and robustness of our method by comparisons with state-of-the-art image quality assessment algorithms. In future work, we would

like to explore and formulate interpretable features that impact the visual text effects quality most, which could possibly enlighten the extension of our network to more general image style assessment.

References

[1] S. Yang, W. Wang, and J. Liu, "TE141K: Artistic text benchmark for text effects transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[2] L. A. Gatys, A. S Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2016.

[3] J. Johnson, A. Alahi, and F. F. Li, "Perceptual losses for real-time style transfer and super-resolution," *Proc. European Conf. Computer Vision*, 2016.

[4] C. Li and M. Wand, "Combining markov random fields and convolutional neural networks for image synthesis," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2016.

[5] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. Kang, "Visual attribute transfer through deep image analogy," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–15, 2017.

[6] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "StyleBank: An explicit representation for neural image style transfer," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2017.

[7] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance Normalization: The missing ingredient for fast stylization," 2017, arXiv:1704.00028.

[8] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," *Proc. Int'l Conf. Computer Vision*, 2017.

[9] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M. Yang, "Universal style transfer via feature transforms," *Proc. Advances in Neural Information Processing Systems*, 2017.

[10] S. Yang, J. Liu, W. Wang, and Z. Guo, "TET-GAN: Text effects transfer via stylization and destylization," *Proc. AAAI Conference on Artificial Intelligence*, 2019.

[11] W. Wang, J. Liu, S. Yang, and Z. Guo, "Typography with Decor: Intelligent text style transfer," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2019.

[12] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2014.

[13] X. Liu, J. van de Weijer, and A. D Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," *Proc. Int'l Conf. Computer Vision*, 2017.

[14] P. Zhang, W. Zhou, L. Wu, and H. Li, "SOM: Semantic obviousness metric for image quality assessment," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2015.

[15] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016.

[16] Y. Deng, C. L. Chen, and X. Tang, "Image Aesthetic Assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017.

[17] L. Liu, H. Dong, H. Hua, and A. C. Bovik, "No-reference image quality assessment in curvelet domain," *Signal Processing Image Communication*, vol. 29, no. 4, pp. 494–505, 2014.

[18] M. A. Saad, A. C. Bovik, and C. Charrier, "DCT statistics model-based blind image quality assessment," *Proc. IEEE Int'l Conf. Image Processing*, 2011.

[19] L. Kang, P. Ye, Y. Li, and D. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," *Proc. IEEE Int'l Conf. Image Processing*, 2015.

[20] K. Y. Lin and G. Wang, "Hallucinated-IQA: No-reference image quality assessment via adversarial learning," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2018.

[21] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[22] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513–516, 2010.

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

[24] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," *Proc. European Conf. Computer Vision*, 2016.

[25] S. Woo, J. Park, J. Lee, and In So Kweon, "CBAM: Convolutional block attention module," *Proc. European Conf. Computer Vision*, 2018.

[26] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image and Video Processing*, 2016.

[27] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating deep network training by reducing internal covariate shift," 2015, arXiv:1502.03167.