Quality Evaluation of Arbitrary Style Transfer: Subjective Study and Objective Metric

Hangwei Chen, Feng Shao, *Member, IEEE*, Xiongli Chai, Yuese Gu, Qiuping Jiang, *Member, IEEE*, Xiangchao Meng, *Member, IEEE*, Yo-Sung Ho, *Fellow, IEEE*

***Abstract*—Arbitrary neural style transfer is a vital topic with great research value and wide industrial application, which strives to render the structure of one image using the style of another. Recent researches have devoted great efforts on the task of arbitrary style transfer (AST) for improving the stylization quality. However, there are very few explorations about the quality evaluation of AST images, even it can potentially guide the design of different algorithms. In this paper, we first construct a new AST images quality assessment database (AST-IQAD), which consists 150 content-style image pairs and the corresponding 1200 stylized images produced by eight typical AST algorithms. Then, a subjective study is conducted on our AST-IQAD database, which obtains the subjective rating scores of all stylized images on the three subjective evaluations, i.e., content preservation (CP), style resemblance (SR), and overall vision (OV). To quantitatively measure the quality of AST image, we propose a new sparse representation-based method, which computes the quality according to the sparse feature similarity. Experimental results on our AST-IQAD have demonstrated the superiority of the proposed method. The dataset and source code will be released at https://github.com/Hangwei-Chen/AST-IQAD-SRQE**

***Index Terms*—Arbitrary style transfer (AST), Image quality assessment (IQA), Content preservation (CP), Style resemblance (SR), Overall vision (OV), Sparse coding, Sparse feature similarity.**

I. Introduction

A. Background

STYLE transfer is a process that strives to render natural images with particular style characteristics from one image (e.g., the style image) while synchronously maintaining the detailed structure information of the content image. Such unique technique not only builds a bridge between the computer vision and appealing artworks, but also gets rid of the dilemma that it would take a long time for a well-trained artist to draw an image in a special style [1]. As shown by an example in Fig. 1, style transfer model can automatically generate a new stylized image based on the content and style of an image. Additionally, style transfer also plays an important role in many computer vision tasks, such as person re-identification [2], semantic segmentation [3], and image reconstruction [4].

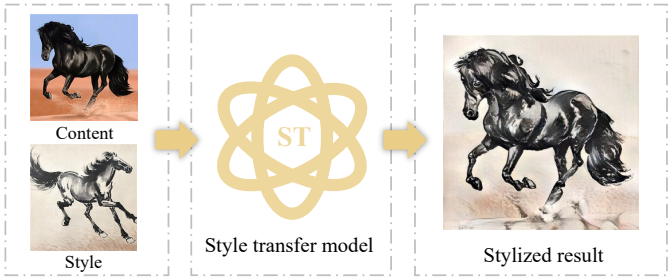


Fig. 1. An example of style transfer model that generates a stylized image.

As customary, style transfer is commonly cast as the study of the texture generation [5]. Early works [6] cope with the texture generation using local statistics or similarity measures on the pixel values. Recently, the field of Neural Style Transfer (NST) was ignited by the groundbreaking work of Gatys *et al.* [7], which is the process of using Convolutional Neural Network (CNN) to perform image translation and stylization. Then, lots of follow-up studies were conducted on the NST algorithm based on the deep neural network in order to promote the transfer efficiency and generation effects. Meanwhile, NST also has rapidly evolved from single-style to infinite-style models, also known as Arbitrary Style Transfer (AST) [8]–[20]. With the ability of utilizing one model to transfer arbitrary artistic style, AST has become a hot topic on computer vision, which could promote the creation of artworks and social communication. However, lacking of a good quantitative evaluation makes it difficult to measure the merits and drawbacks of the AST algorithms. It is therefore necessary to study and eventually evaluate the AST images.

B. Style Transfer Techniques

Mathematically, the style transfer model can be described as a translation process [21]:

$$I_{AB} \in B : \mathcal{M}_{A \longmapsto B}(I_A). \tag{1}$$

where I_A is the input content image from a source domain A to a target domain B, $\mathcal{M}_{A \longmapsto B}$ is a mapping that generates image $I_{AB} \in B$ indistinguishable from style image $I_B \in B$ on the target domain given the input source image $I_A \in A$.

Following the growth of neural networks, numerous NST methods have been proposed to study the problem of style

This work was supported by the Natural Science Foundation of China (grant 62071261, 62076013, 62021003, 62271277), and Zhejiang Province Natural Science Foundation of China (grant R18F010008, LR22F020002). (*Corresponding author: Feng Shao.*)

Hangwei Chen, Feng Shao, Xiongli Chai, Yuese Gu, Qiuping Jiang, and Xiangchao Meng are with the Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China (e-mail: 1010075746@qq.com; shaofeng@nbu.edu.cn; 747866472@qq.com; 805682724@qq.com; jiangqiuping@nbu.edu.cn, mengxiangchao@nbu.edu.cn).

Yo-Sung Ho is with the School of Information and Communications, Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea (e-mail: hoyo@gist.ac.kr).

transfer. According to the optimization ways adopted in the NST task, the existing NST methods can be divided into two categories: Image-Optimization-Based Online Neural Methods [8], [9] and Model-Optimization-Based Offline Neural Methods [12]–[20]. The former category could produce appealing stylized results through the iterative image optimization process, while the latter uses generated models with feed-forward networks to produce special style patterns. Particularly, considering the efficiency issue, it is more meaningful to design the Model-Optimization-Based Offline Neural Methods in practice.

The AST method is one of the Model-Optimization-Based Offline Neural Methods, which can accept an arbitrary artistic style as input and produce stylized results in a single feed-forward network once upon the model is trained. Thus, the AST has received substantial attention due to the increasing scientific and artistic values. In below, we review the details of the AST studies. For more comprehensive introduction, readers can refer to the survey [1].

1) Non-Parametric Methods: The common idea of the non-parametric methods [10], [11] is to seek the similarities between the patches in content and style images and swap them. Chen *et al*. [10] realized AST for the first time that developed a Style-Swap operation to swap the feature patches of content images with the best matching style feature patches. Another work by Gu *et al*. [11] also proposed a patch-based method considering the matching of both global statistics and local patches. However, if the structures of content image and style image are largely different, these methods cannot efficiently protect the shape with unsatisfactory style patterns.

2) Parametric Methods: The characteristic of these parametric methods [12]–[19] is to optimize a target function that reflects the similarity between the input and stylized images. These parametric methods [12]–[14], utilize the Gram-based VGG perceptual loss to produce stylization with a few modifications. Li *et al*. [12] proposed a linear transform function (LST) from content and style features for stylization. Li *et al*. [13] performed a pair of feature transforms, whitening and coloring (WCT), for feature embedding within a pre-trained encoder-decoder module. Huang *et al*. [14] proposed a novel adaptive instance normalization (AdaIN) layer that adjusts the mean and variance of the content input to match the style input. Another promising trend in parametric methods is to integrate attention mechanism into the deep neural network. Yao *et al*. [15] first considered multi-strokes with self-attention mechanism. Park*et al*. [16] introduced Style-Attentional Network (SANet) to match content and style features for achieving good results with evident style patterns. Deng*et al*. [17] proposed a multi-adaptation module that takes the global content structure and local style patterns into account. In addition, Zhang *et al*. [18] introduced a multi-modal style transfer (MST) via efficient graph cuts algorithms, which explicitly considers the matching of semantic patterns in content and style images. Inspired by MST, Chen *et al*. [19] developed a structure-emphasized multimodal style transfer (SEMST) model, which can flexibly match the content and the style clusters based on the cluster center norm.

C. Image Quality Assessment of AST

1) Motivations: Notwithstanding the current state-of-the-art methods have shown successful stream in style transfer, arbitrary style transfer image quality assessment (AST-IQA) has been a long-standing problem and relatively unexplored in the community. Nevertheless, with the exception of a few quantitative protocols [22], [23], almost all researches evaluate the stylization quality in a qualitative way (e.g., by side-by-side subjective vision comparisons or different user studies), which suffers from the following limitations. First, the stylization examples displayed for qualitative comparison are limited in number and often carefully selected to favor the cases where the algorithm works well. In other words, the results of these presentations are not comprehensive enough. Second, the selected observers often lack sufficient experience and expertise in art, which makes qualitative evaluation less convincing. Thus, it is practical and necessary to propose a reliable metric to quantitatively assess the stylization quality.

2) Challenges: Different from the traditional IQA tasks [24]–[29] that usually focus on general distortions generated by various stimuli, the AST-IQA is closely related to aesthetics and poses serious challenges in both subjective and objective assessment.

Subjective assessment: The first challenge is how to design and conduct a human subjective study that can obtain reliable ground truth labeling on a set of stylized images [30]. To our best knowledge, there is no publicly available database for AST-IQA. The most related benchmarks are the non-photorealistic rendering (NPR) benchmarks [31], [32], which are used for testing stylization algorithms without human opinion scores. Currently, there is no quality assessment standard for measuring the performance of style transfer, since the AST-IQA is a highly subjective task, e.g., different subjects tend to have various ideas towards the same stylized result, especially for the style evaluation.

Objective assessment: Once the subjective dataset is obtained, the next challenge is how to design a metric that can automatically evaluate the perceptual quality of the AST images closely consistent to human vision. In several image style transfer works [21], [33], [34], some Full-reference (FR) metrics (e.g., SSIM [35]) have been used to evaluate the similarity between the structures of the content and stylized images. However, strictly speaking, evaluating the AST quality is not a classic FR-IQA task. Straightforwardly applying the traditional FR-IQA strategy to the field of style transfer is problematic, since the stylized image has different content detailed information with the source and style images, and the ground truth image is unavailable for the stylized image (i.e., completely different with fidelity evaluation in the tradition IQA tasks). Although the general No-reference (NR) metrics (e.g., NIQE [28], TCLT [36] and BRISQUE [37]) have made great progress in the tradition IQA tasks without ground truth, they are also not applicable to AST-IQA because stylized images are closely related to aesthetics rather than naturalness. In addition, it is necessary for AST-IQA to fully take the original information of content and style images into account. Recent works target to address these challenges using some objective

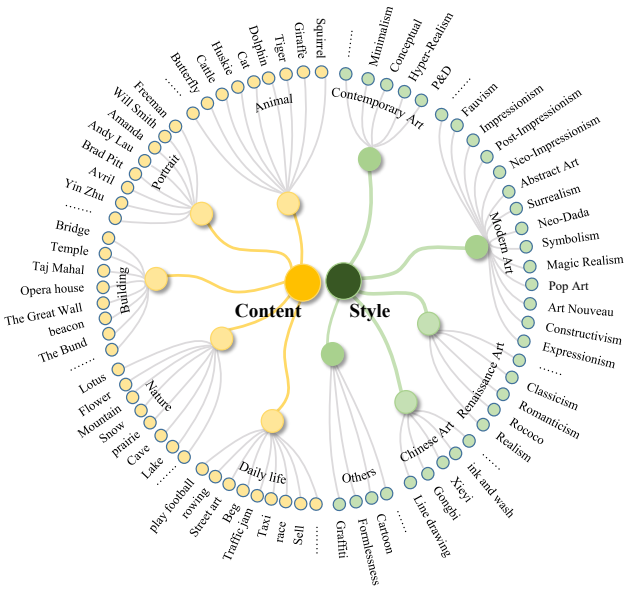


Fig. 2. Taxonomic structure of our source images.

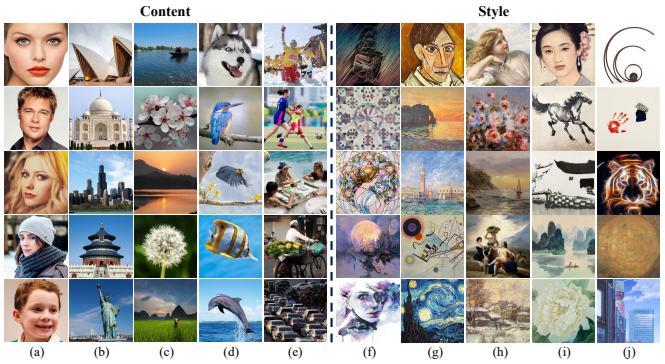


Fig. 3. Examples of content and style images in the AST-IQAD database. (a) Portraits. (b) Building. (c) Nature. (d) Animal. (e) Daily life. (f) Contemporary art. (g) Modern art. (h) Renaissance art. (i) Chinese art. (j) Others.

metrics from the perspective of different quality factors. Yeh *et al*. [22] proposed a metric with two factors (i.e., effectiveness and coherence) where the former factor is a measure of the extent to which the style was transferred, and the latter is a measure of the extent to which the transferred image is decomposed into the content objects. Wang *et al*. [23] first decomposed the quality of style transfer into three quantifiable factors, i.e., the content fidelity (CF), global effects (GE) and local patterns (LP), which cover the main aspects considered by different types of existing NST methods. However, these metrics either focus on the limited factors of style transfer quality (e.g., lacking of fine-grained quality factors), or are simple in quality pooling, which cannot effectively match the aesthetic perception of human observers in practice.

D. Overview of Our Work

In this paper, to resolve the above challenges, we carry out an in-depth investigation for AST-IQA from both subjective and objective perspectives. With the popularization of art education, a majority of people with similar background can make similar perceptual judgments about some basic elements in artistic painting (e.g., color tone, brush stroke, distribution of objects, and contents). Benefitted from the above conditions, to address the challenge in subjective assessment, we decompose the quality of AST into three quality factors that are easier to understand, namely content preservation (CP), style resemblance (SR), and overall vision (OV). These quality factors are assigned own preference labels by participants according to the knowledge in style transfer, intuition in vision and feed-back from the surveys. To address the challenge in objective assessment, as suggested by the recommendation system [38], we regard the problem as a data-driven modeling of user preference [39], and conduct quantitative evaluation of AST-IQA using sparse representation to dig intrinsic representation for content and style images. To sum up, the major contributions of our work are summarized as follows:

1) To carry out in-depth studies on perceptual quality assessment of AST stylized images from both subjective and objective aspects, we build a new AST images database named AST-IQAD, which consists 150 content-style image pairs and the corresponding 1200 stylized images produced by eight typical AST algorithms. Each stylized image contains the subject-rated CP, SR and OV scores. To our knowledge, it is the first large-scale AST image database with human opinion scores. Therefore, it can serve as a benchmark to objectively evaluate the existing AST methods and potentially guide the design of different AST methods.

2) We proposed a new sparse representation-based image quality evaluation metric (SRQE) for AST-IQA, which can quantitatively evaluate the quality factors of CP, SR, and OV. To be more specific, in the training phase, we learn multi-scale style and content dictionaries to represent the style characteristic and structure of the stylized images. In the quality estimation phase, the sparse feature similarities are further exploited to compute the qualities of CP and SR respectively, and the OV quality is obtained by combining the SR and CP qualities. Extensive experiments are conducted on our AST-IQAD dataset and the experimental results demonstrate the proposed method can well evaluate the AST quality.

The rest of this paper is organized as follows. Section II illustrates the details of AST-IQAD. Section III introduces the proposed method in detail. The experimental results are shown and discussed in Section IV. Finally, conclusions are drawn in Section V.

II. AST-IQAD DATABASE

To investigate quality assessment of AST images, we construct a new arbitrary style transfer database (AST-IQAD) for quality assessment, which includes 1200 stylized images generated by eight typical AST methods, and conduct a subjective quality evaluation study on the AST-IQAD database to capture the human option scores. To our knowledge, it is the first large-scale database for AST-IQA, and it can provide a better resource to evaluate and advance state-of-the-art style transfer algorithms. We will introduce the details of the AST-IQAD database in the following parts.

Fig. 4. Examples of AST images produced by different algorithms.

A. Source Images

Since the essence of style transfer is to migrate the color tone and stroke pattern from the source to target image while retaining the content structure information of the target image. To provide deeper and intuitive information, the selected source images should have clear and reasonable structures. Thus, we set up a hierarchical taxonomic system (shown in Fig. 2) for source image (the content images) and target images (the style images), respectively. Both content and style images are labeled with five categories.

1) Content Images: We collect 75 high quality images with a resolution of 512×512 pixels from the NPRgeneral benchmark [31] and other famous photography websites. According to the criteria of coverage [31], the content images are comprised of five categories (i.e., animal, portrait, building, nature, and daily life) with a wide range of characteristics (e.g., contrast, texture, edges and meaningful structures). Examples of the selected content images in the database are shown in Fig. 3.

2) Style Images: We select 126 style images from some artwork websites and WikiArt. WikiArt is the largest art encyclopedia in the visual arts from all over the world. Since these original style images are not provided with high resolution, all style images are set to the same resolution of 512×512 in line with the content images. These style images cover five categories including contemporary art, modern art, renaissance art, Chinese art, and others. Examples of style images in the AST-IQAD dataset are also shown in Fig. 3.

3) Content-Style Image Pairs: Pairing content images with appropriate and diverse style images can make the AST results more aesthetically pleasing. In our work, we provide two 'Paired' and 'Unpaired' mechanisms [40] for each content image, in which 'Paired' means that the content and the style images are semantically consistent (e.g., the same source of birds), while 'Unpaired' means that the content and the style images are the representations of different sources (e.g., the style images may be regular patterns or texture decorations). In total, we use 75 content images and 126 style images (including reused style images) to generate 150 content-style image pairs. Then, eight AST algorithms are conducted on the content-style image pairs to generate the AST images. More information of the image pairs can be found in our database.

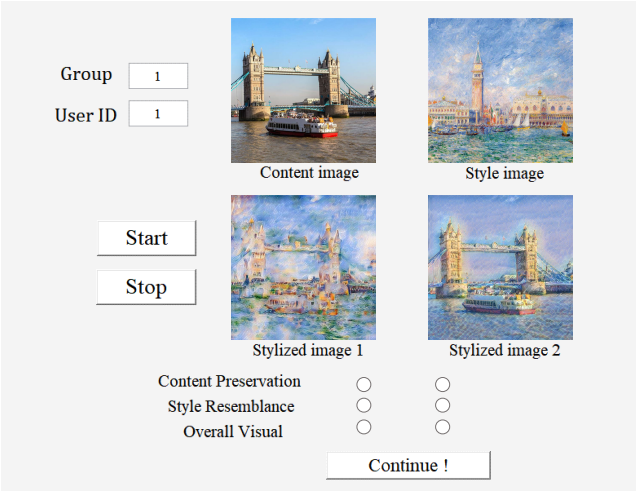


Fig. 5. Subjective interface in the experiment.

B. Arbitrary Style Transfer Algorithms

Different with traditional IQA databases that stimulated with different distortion stimuli, the testing images in our databases are generated from different AST methods. The eight representative AST methods used in the database are listed in

TABLE I
DESCRIPTION OF AST ALGORITHMS USED IN THE DATABASE.

Types	Methods	Descriptions
Gram-based	WCT [13]	Whitening and coloring transforms
	AdaIN [14]	Adaptive instance normalization
	LST [12]	Linear style transfer
Attention-based	AAMS [15]	Attention-aware Multi-stroke style transfer
	MANet [17]	Multi-adaptation networks
	SANet [16]	Style-attentional networks
Graph-based	MST [18]	Multimodal style transfer via graph cuts
Cluster-based	SEMST [19]	Structure-emphasized multimodal style transfer

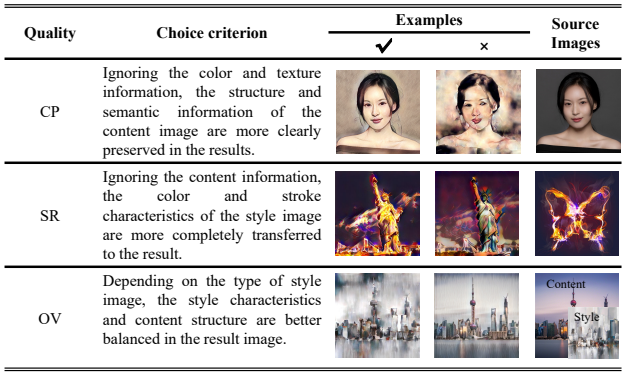


Fig. 6. Preference criteria for the AST-IQA task.

Table I, including AAMS [15], AdaIN [14], MANet [17], LST [12], WCT [13], MST [18], SEMST [19], SANet [16]. These algorithms cover a wide variety of techniques, including Gram-based, Attention-based, Graph-based and Cluster-based methods. As a result, we can obtain 1200 style-transferred images from 150 content-style image pairs. As shown by the examples of style-transferred images in Fig. 4, we have the following observations: 1) In the Gram-based methods, LST [12] and AdaIN [14] can well preserve the content information but may suffer from wash-out artifacts. On the contrary, WCT [13] is impressive in color and texture making the "painting taste" more intense while fails to preserve the main content structures. 2) The attention-based methods have distinct content structures and rich style patterns but may produce unpleasing visual artifacts. For example, SANet [16] and MANet [17] methods produce unpleasing eye-like artifacts, and AAMS [15] introduces imperceptible dot-wise artifacts. 3) The stylized results of MST [18] and SEMST [19] are similar and produce both visible content and proper stylization.

C. Human Subjective Study

Due to the different rating standards across different observers and the influence of visual content [41], the subjective quality scores evaluated by absolute category rating are imprecise, biased, and inconsistent, while the preference label, representing the relative quality of two images, is generally precise and consistent for the task. For this consideration, we adopt the pairwise comparison (PC) approach which aims to provide a binary preference label between a pair of stylized images.

The experiment was carried out in an indoor laboratory with low ambient illumination calibrating in accordance with the ITU-R BT.500-13 recommendations [42]. The subjective software interface (as shown in Fig. 5) is displayed at a 23-inch true color (32bits) LCD monitor with the screen resolution of 1920×1080 pixels. The ratio of background illumination behind the display to the image peak luminance is 0.15. The viewing distance is approximately six times the image height. In the interface specific to the AST task, participants are simultanesously shown two stylized images along with a style-content image pair, and are required to vote on three preferences for content preservation (CP), style resemblance (SR), and overall vision (OV). The detailed descriptions of preference criteria are summarized in Fig. 6.

A total number of 45 subjects aged from 18 to 30, including experts and students from the Faculty of Art, and under-graduate students with experience in image processing, were participated in the subjective study. For each content-style image pair producing eight AST results, we have 28 pairwise comparisons in the subjective ranking study. In total, 4200 pairwise comparisons from 150 content-style pairs are involved in the subjective study. To reduce the possible fatigue effect, we divide the experiment into three sub-sessions (each sub-session contains 15 subjects), in which each participant would take part in one sub-session completing 1400 PC voting. The content images presented in each sub-session included most of the scene types, and were displayed in a random order to reduce the possible memory interference. After completing every 200 pair comparisons, each subject was encouraged to look away to relax their eyes and asked to rest for about 15 minutes. Each participant took about six hours (including rest periods) to complete the whole subjective experiment. As a result, we can get 63,000 votes on each subjective evaluation.

D. Subjective Data Analysis

1) Global Ranking of AST Algorithms: To derive global ranking of the AST algorithms from the corresponding PC results, we adopt the Bradley-Terry [43] model to estimate the subjective score for each algorithm. The probability that the i-th method is favored over the j-th method is defined as:

$$P(i \succ j) = \frac{e^{u_i}}{e^{u_i} + e^{u_j}} \tag{2}$$

where u_i and u_j are subjective scores prefered for the i-th and the j-th methods, respectively. Then, the negative log-likelihood for the B-T scores $\mathbf{u} \in \mathbb{R}^{n_{ast}}$, where n_{ast} is the number of AST algorithms, can be jointly expressed as:

$$\mathcal{L}(\mathbf{u}) = -\log\left(\prod_{i=1}^{n_{ast}} \prod_{j=1, j\neq i}^{n_{ast}} P(i \succ j)^{\mathcal{W}_{ij}}\right) \tag{3}$$

where $\mathcal{W}_{ij}$ is the (i, j)-th element in the winning matrix $\mathcal{W} \in \mathbb{Z}^{n_{ast} \times n_{ast}}$, representing the number of times that the i-th method is preferred over the j-th method. By setting the

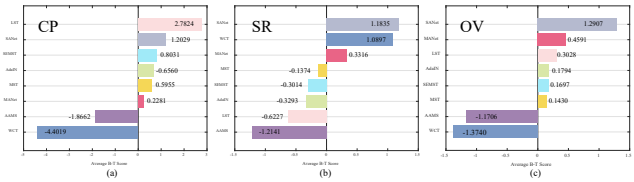


Fig. 7. Average B-T scores of different AST algorithms at each subjective evaluation.

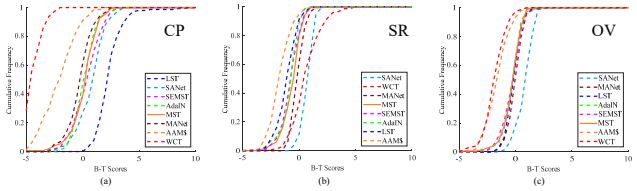


Fig. 8. Cumulative probability distribution curves of B-T scores at each subjective evaluation.

derivative of $\mathcal{L}(\mathbf{u})$ in Eq. (3) to zero to solve the optimization problem [44], the final B-T scores $\mathbf{u}$ are obtained via zero mean normalization, served as the ground truth subjective rating scores.

We can get a B-T score for each AST result in each subsession by applying the B-T model. Fig. 7 shows the average B-T scores of different AST methods for three subjective evaluations (i.e., CP, SR, and OV). A higher B-T score indicates a better performance. From the figure, some interesting observations could be drawn: 1) The B-T scores of different AST algorithms show various trends on the subjective evaluations of CP, SR and OV, which indicate that different methods have specific advantages. 2) For the CP evaluation, LST [12], which receives the best ranking, has shown significant advantage in maintaining structure information over other methods by a large margin. WCT [13] performs worst in the CP test due to treating diverse image regions in an indiscriminate way. 3) For the SR evaluation, SANet [16] performs best on average, attributed to the attention mechanism that generates more local style details. However, AAMS [15], which is also based on self-attention mechanism, does not perform well because the multi-stroke pattern generates imperceptible dot-wise artifacts. In addition, WCT [13] shows the competitive advantage, which indicates the effectiveness of whitening and coloring transformations. 4) For the OV evaluation, attention-based methods (e.g., SANet [16] and MANet [17]) rank top-2 on performance. It is not surprising because attention-based algorithms pay more attention to those feature-similar areas in the style image for stylizing a content image region. Furthermore, we plot the cumulative probability distribution curves of the B-T scores obtained from all AST results in Fig. 8. The AST method corresponding to the rightmost curve performs better because it accumulates higher B-T scores.

2) Correlation of B-T Scores: This part analyzes the correlation of B-T scores between the subjective evaluation of CP and SR, SR and OV, and CP and OV, as shown in Fig. 9. Taking SR and CP as an example, the method above (below) the diagonal indicates better performance in CP (SR), and vice versa. From

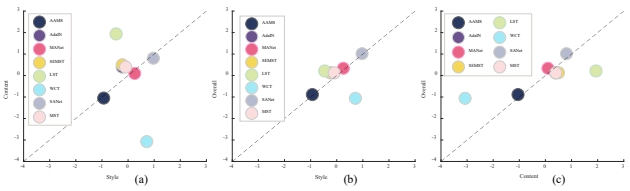


Fig. 9. Correlation of B-T scores between the subjective evaluations of CP and SR, SR and OV, and CP and OV.

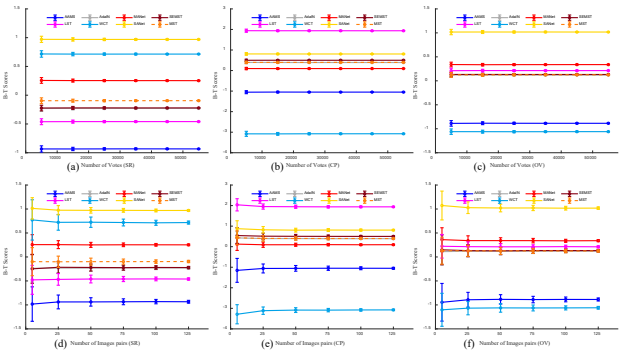


Fig. 10. Convergence analysis on the number of votes and image pairs at each subjective evaluation.

the Fig. 9 (a), it is observed that the SANet [16] method has achieved a better consistency in SR and CP evaluation, and the WCT [13] method shows excellent performance in SR but fails to sufficiently maintain the content structure. In addition, the correlation phenomena observed in Fig. 9 (b) and (c) also reveal the complex relationship between OV and CP (or SR).

3) Convergence Analysis: To demonstrate that the scale of the subjective study is large enough to support performance evaluation, we further analyze the convergence from the perspectives of the number of subject votes and images pairs.

Number of votes: We randomly sample λ (λ = 5000, 15000, 25000..., 55000) votes from a total of 63,000 voting results, and calculate the B-T scores for each AST algorithm. To avoid the possible bias, we repeat this process 1000 times with different samples of votes. Fig. 10 (a)-(c) show the mean and standard deviation of B-T scores for each sample on three subjective evaluations. It is observed that as the number of votes grows, the B-T scores tend to be stable, which demonstrates that the number of votes is sufficiently large for performance evaluation.

Number of images pairs: Similar to the above convergence analysis, we randomly sample μ (μ = 5, 25, 50, 75, 100, 125) content-style image pairs from our dataset and then plot the means and the standard deviations in Fig. 10 (d)-(f). Obviously, as the number of images pairs grows, standard deviation of B-T Scores decreases, which indicates that the B-T scores obtained from the subjective study are stable.

Overall, the above two kinds of convergence analysis demonstrate the reliability of the subjective rating scores.

III. OBJECTIVE QUALITY EVALUATION

In this paper, we propose a new sparse representation-based image quality evaluation metric (SRQE) for AST-IQA, as

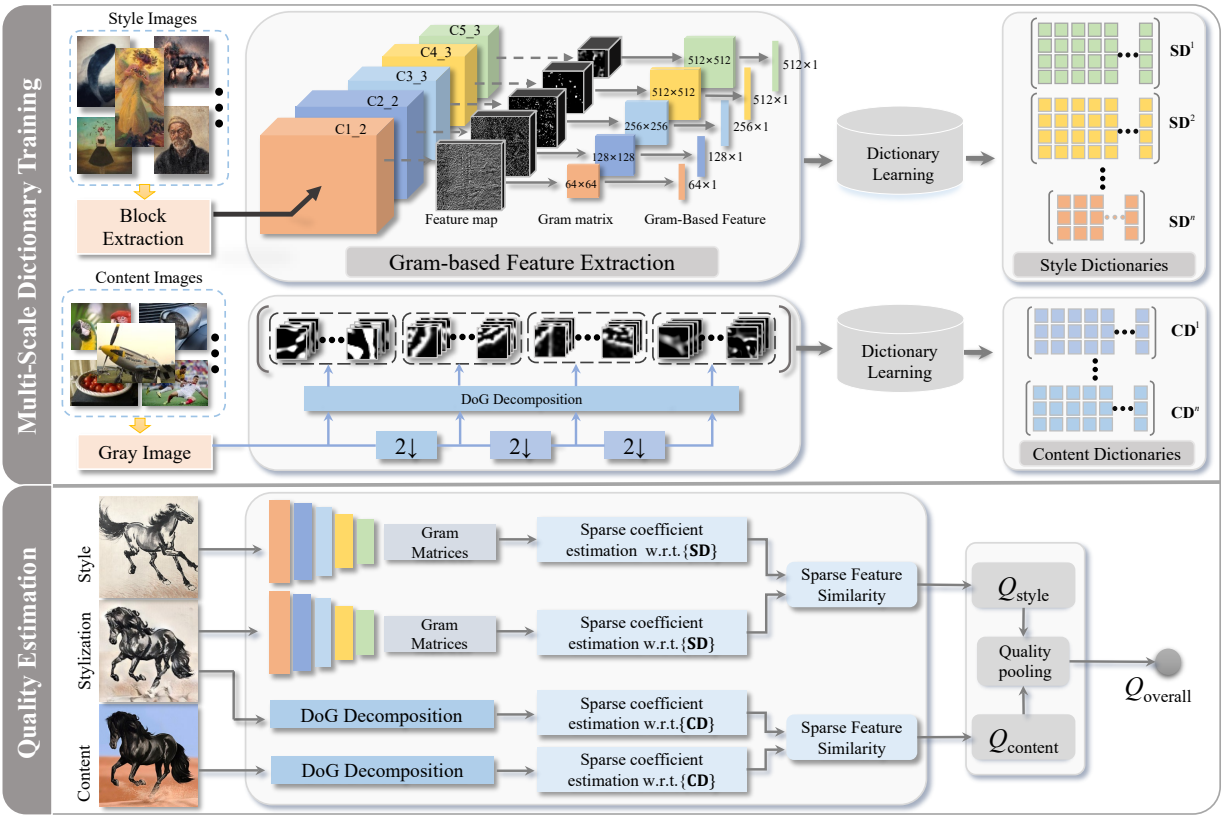


Fig. 11. Framework for the proposed SRQE method.

shown in Fig. 11. The process is composed of two phases: multi-scale dictionary training and quality estimation. In the training phase, the multi-scale style and content dictionaries, learnt from the training databases via sparse representation, are utilized to build style representation for style images and capture inherent structures for the content images, respectively. In the quality estimation phase, the quality of SR (or CP) is obtained by estimating the sparse feature similarity between the stylized image and the style (or content) image. Finally, the OV quality is acquired by combining the SR quality and CP quality together. In what follows, we elaborate on each step of the proposed method.

A. Style Feature Extraction

1) Selection of Training Database: As show in Fig. 12 (a), we re-collected 100 new style images, covering a wide variety of categories, as training images for dictionary learning. Note that there is no overlap between these images and the above collected style images (in the Section II-A) to ensure the complete independence of the training and test data. In addition, to avoid overfitting, we augment the training database by evenly partitioning the whole images. The impacts of the number of image blocks and training database will be discussed in the following Section IV-B.

2) Gram-Based Feature: The work in [7] demonstrated that the correlations between convolution responses at the same layer (i.e., Gram matrices) yielded effective texture synthesis and can effectively grasp the image style. Additionally, sparse representation technique shows great prospects in image visual style analysis [45]. Inspired by this, we construct a style perception model based on sparse representation, while incorporating high-level perceptual information (Gram matrices) extracted from deep neural network.

In this paper, we resort to compute gram matrices from style images using a pre-trained VGG network of the state-of-the-art full-reference IQA model (DISTS) [46] which has superior performance in evaluating texture similarity. Assumed that the feature map of a sample style image I_s at layer l of DISTS [46] is denoted as $\mathbf{F}^l(I_s) \in \mathbb{R}^{C\times H\times W}$, where C is the number of channels, and H and W represent the height and width of the feature map, the Gram-based representation is computed from the feature map $\mathbf{F}^l(I_s)' \in \mathbb{R}^{C\times(HW)}$ aggregated from the $\mathbf{F}^l(I_s) \in \mathbb{R}^{C\times H\times W}$:

$$\mathbf{G}\left(\mathbf{F}^l(I_s)'\right) = \left[\mathbf{F}^l(I_s)'\right]\left[\mathbf{F}^l(I_s)'\right]^T \tag{4}$$

In the implementation, we use the first to fifth VGG network layers of DISTS [46] to produce a set of Gram-based representations at different layers, namely $\{\mathbf{G}^l \in \mathbb{R}^{C\times C},\ l = 1,2,\ldots,L\}$, where L=5 denotes the highest layer and $C \in \{64,\ 128,\ 256,\ 512,\ 512\}$ correspond to the numbers of feature maps at each layer.

After the above processing, each Gram-based representation then generates a Gram-based style feature vector $\mathbf{g}^l \in \mathbb{R}^{C\times 1}$ through averaging each row, described as:

$$\mathbf{g}^l = \mathbf{G}^l \cdot \mathbf{x}^l. \tag{5}$$

$$\mathbf{x}^l = [x_1^l, x_2^l, \ldots, x_C^l]^T. \tag{6}$$

Fig. 12. Some of the images of the training databases used in the paper. (a) Style training database. (b) Content training database.

where $x_1^l = \ldots = x_C^l = 1/C$

Finally, numerous style feature vectors at the same layer extracted from different style images are used to form a style matrix $\mathbf{SM}$. As a result, we can obtain five different style matrices (corresponding to five layers) $\mathbf{SM}^l = [\{\mathbf{g}^l\}_1, \{\mathbf{g}^l\}_2, \ldots, \{\mathbf{g}^l\}_{N^l}] \in \mathbb{R}^{C \times N^l}$. All style matrices will be used for the subsequent style dictionary learning.

3) Multi-Scale Style Dictionary Learning: Using the above style matrices $\mathbf{SM}^l = [\{\mathbf{g}^l\}_1, \{\mathbf{g}^l\}_2, \ldots, \{\mathbf{g}^l\}_{N^l}] \in \mathbb{R}^{C \times N^l}$ as input, we learn multi-scale style dictionary $\mathbf{SD}^l$ by seeking a sparse representation for each style feature vector $\mathbf{g}^l$ under specific sparsity constraint τ. Each style sub-dictionary $\mathbf{SD} = [\mathbf{sd}_1, \mathbf{sd}_2, \ldots, \mathbf{sd}_U] \in \mathbb{R}^{C \times U}$ contains U basic elements. Formally, the process of multi-scale style dictionary learning can be formulated as:

$$\left\langle \mathbf{SD}^l,\ \hat{\boldsymbol{\alpha}}_i \right\rangle = \arg\min \sum_{i=1}^{N} \left\| \{\mathbf{g}^l\}_i - \mathbf{SD}^l \boldsymbol{\alpha}_i \right\|_2^2 \quad \text{s.t. } \forall i,\ \|\boldsymbol{\alpha}_i\|_0 \le \tau \tag{7}$$

where $\|\cdot\|_2$ is the l_2-norm operator, $\|\cdot\|_0$ denotes the l_0-norm that counts the number of non-zero elements in a vector, and $\boldsymbol{\alpha}_i$ is the sparse coefficient vector of $\{\mathbf{g}^l\}_i$. Typically, both $\mathbf{SD}^l$ and $\boldsymbol{\alpha}_i$ are unknown in this stage. We resort to the online dictionary learning (ODL) algorithm implemented in the SPArse Modeling Software [47] to solve this NP-hard problem. Details of dictionary learning can refer to [47].

B. Content Feature Extraction

1) Selection of Training Database: Since the essence of the proposed content evaluation model is to restore the structure information of the source content images and stylized images based on dictionary learning, we only select natural images to construct the content dictionary. Refer to [48], we randomly select ten natural images from the TID 2013 [49] database and NPRgeneral [31], which have different scenes in the images, as shown in Fig. 12 (b).

2) DoG Response Feature: As known, human visual perception is highly sensitive to the edge information, the major objects in the painting emphasized by the artists often contain distinct edges in most cases [39]. Intuitively speaking, the outline of the main objects largely reflects the content information of artworks. Inspired by this, the edge information, as the significant component in painting content, needs to be deeply investigated for evaluating the CP quality. Furthermore, an outstanding painting will be appreciated by humans in local details and global perception. Thus, it is necessary to utilize multi-scale strategy to better describe the content of the painting from coarse to fine level of detail [50]. As shown in Fig. 13, the multi-scale Difference of Gaussian (DoG) is applied to represent the content feature [51], which can properly simulate the receptive field of retinal cells. First, the DoG signals, $DoG(x, y)$, at different scales can be computed by:

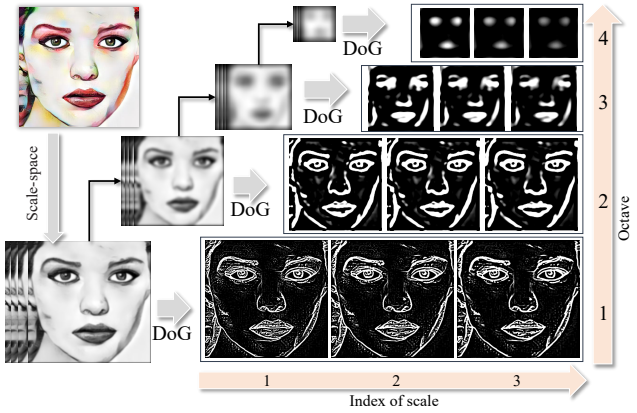


Fig. 13. An example of DoG multi-scale space.

$$DoG(x, y) = \left| R_{\sigma, k\sigma}(x, y) \otimes I(x, y) \right|. \tag{8}$$

where $I(x, y)$ denotes the pixel location (x, y) of the input image, the symbol $\otimes$ denotes the convolution operation, and $R_{\sigma, k\sigma}(x, y)$ is defined as the difference between two Gaussian kernel with nearby scales σ and $k\sigma$:

$$R_{\sigma, k\sigma}(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) - \frac{1}{2\pi k^2\sigma^2} \exp\left(-\frac{x^2 + y^2}{2k^2\sigma^2}\right) \tag{9}$$

where σ and k are used to control the scales of DoG. Refer to [48], we set k = 1.6, and $\sigma \in \{0,\ 1,\ 1.6,\ 2.56,\ 4.096\}$ in the experiment. Here, σ = 0 denotes the original scale.

Once the DoG signals at the current octave are computed, the last scale-space image was selected as the new input and was down-sampled by a factor of two to repeat the above process, thereby producing a set of DoG signals with a variety of octaves and scales, namely $\{DoG^{z,o}(x, y)\}$, where $z \in \{1, \ldots, Z\}$ denotes the z-th scale, and $o \in \{1, \ldots, O\}$ denotes o-th octave.

After the above processing, each DoG signal is partitioned into numerous patches with size of 8×8, subtracted by the mean value. In the implementation, 1000 overlapped patches having rich details and structures are selected as training samples. Then, these patches are vectorized into column vectors to form a content matrix $\mathbf{CM}$, $\mathbf{CM} = [\mathbf{y}_1, \ldots, \mathbf{y}_k] \in \mathbb{R}^{T \times K}$, based on which the subsequent overcomplete content dictionary is learned. Each patch $\mathbf{y}_k \in \mathbb{R}^{T \times 1}$ contains T pixels and $k = 1, \ldots, K$. Here, K = 1000.

3) Multi-Scale Content Dictionary Learning: Similar to the above multi-scale style dictionary learning, the multi-scale

content dictionary $\mathbf{CD}^{z,o}$ can be learned from multi-scale content matrices $\mathbf{CM}^{z,o}$. Each content sub-dictionary $\mathbf{CD} = [\mathbf{cd}_1, \mathbf{cd}_2, \ldots, \mathbf{cd}_U] \in \mathbb{R}^{C\times U}$ contains V basic elements. In the experiment, we set V = 256. Similarly, the process of multi-scale content dictionary learning can be formulated as:

$$\left\langle \mathbf{CD}^{z,o},\ \hat{\boldsymbol{\beta}}_i \right\rangle = \arg\min \sum_{i=1}^{K} \|\mathbf{y}_i - \mathbf{CD}^{z,o}\boldsymbol{\beta}_i\|_2^2 \quad \text{s.t. } \forall i,\ \|\boldsymbol{\beta}_i\|_0 \le \tau \tag{10}$$

where $\boldsymbol{\beta}_i$ is the sparse coefficient vector of $\mathbf{y}_i$. Note that we also apply the ODL algorithm to solve Eq. (10)

C. Feature Similarity Measurement

Through the above efforts, we obtain two types of over-complete multi-scale dictionaries, containing U and V basic atoms as the column vectors in $\mathbf{SD}^l$ and $\mathbf{CD}^{z,o}$, respectively. Thus, each style feature vector $\mathbf{g}^l$ (or content patch $\mathbf{y}_k$) can be sparsely represented as a linear combination of basic atoms contained in $\mathbf{SD}^l$ (or $\mathbf{CD}^{z,o}$).

1) Style Sparse Coefficients Estimation: Given the testing stylized image I_t and style image I_s, we can obtain two Gram-based representations using the DISTS network, denoted as $\mathbf{G}_t^l \in \mathbb{R}^{C\times C}$ and $\mathbf{G}_s^l \in \mathbb{R}^{C\times C}$. Then, the style sparse coefficient vectors can be estimated by a weighted linear combination of previously learnt dictionary elements, i.e.,

$$\mathbf{s}^l = \mathbf{G}_s^l \times (\mathbf{SD}^l)^+ \tag{11}$$

$$\mathbf{ts}^l = \mathbf{G}_t^l \times (\mathbf{SD}^l)^+ \tag{12}$$

where $(\mathbf{SD}^l)^+$ denotes the generalized inverse matrices of $(\mathbf{SD}^l)$.

2) Content Sparse Coefficients Estimation: For the testing stylized image I_t and the source content image I_c, after the same processing steps as in the training phase, we can obtain patch vectors $\mathbf{y}_c^{z,o}$ from I_c and its corresponding patch vectors $\mathrm{y}_t^{z,o}$ from I_t. Similarly, the content sparse coefficient $\mathbf{c}$ and $\mathbf{tc}$ can be computed by:

$$\mathbf{c}^{z,o} = \mathbf{y}_c^{z,o} \times (\mathbf{CD}^{z,o})^+ \tag{13}$$

$$\mathbf{tc}^{z,o} = \mathbf{y}_t^{z,o} \times (\mathbf{CD}^{z,o})^+ \tag{14}$$

where $(\mathbf{CD}^l)^+$ denote the generalized inverse matrices of $(\mathbf{CD}^l)$.

3) Sparse Feature Similarity Measure: From the above estimation phase, we generate the sparse coefficients $\mathbf{s}^l, \mathbf{c}^{z,o}, \mathbf{ts}^l, \mathbf{tc}^{z,o}$ on style, content and stylized images. Considering that these sparse coefficients are represented as a linear combination of basis vectors, meaning that the similarity between the style feature vectors or content patches can be directly measured using their sparse coefficient vectors. Thus, the style and content similarities are respectively defined as:

$$SS^l\left[\mathbf{G}_s^l, \mathbf{G}_t^l\right] = \frac{2\left\langle \mathbf{s}^l, \mathbf{ts}^l \right\rangle + \eta}{\|\mathbf{s}^l\|_2 \cdot \left\|\mathbf{ts}^l\right\|_2 + \eta} \tag{15}$$

$$CS^{z,o}\left[\mathbf{y}_c^{z,o}, \mathbf{y}_t^{z,o}\right] = \frac{2\left\langle \mathbf{c}^{z,o}, \mathbf{tc}^{z,o} \right\rangle + \eta}{\|\mathbf{c}^{z,o}\|_2 \cdot \|\mathbf{tc}^{z,o}\|_2 + \eta} \tag{16}$$

where $\langle\cdot\rangle$ calculates the inner product, η is a constant with a small value added to prevent the denominator to be zero. The SS measures the style similarity between the style and the stylized image, and CS measures the content similarity between the content and the stylized image.

D. Final Quality Pooling

To measure the final quality between a stylized image and its corresponding content and style images, we need to pool the above spare feature similarities into a single score. In our pooling strategy, we first pool the style and content sparse feature similarities into SR and CP scores across all scales or octaves, and then combine the scores to measure the OV quality score. First, the SR quality score Q_{style} is defined as:

$$Q_{\text{style}} = \prod_{l=1}^{L} (SS)^l \tag{17}$$

Then, the CP quality score Q_{content} is defined as:

$$Q_{\text{content}} = \frac{1}{Z^2} \prod_{o=1}^{O} \left(\sum_{z=1}^{Z} (CS)^{z,o} \right) \tag{18}$$

where Z and O denote the number of scales and octaves.

Finally, the OV quality Q_{overall} is calculated by combining Q_{style} and Q_{content} into a score:

$$Q_{\text{overall}} = (Q_{\text{content}})^{\omega_1} \cdot (Q_{\text{style}})^{\omega_2} \tag{19}$$

where the parameters ω_1 and ω_2 are used to adjust the relative importance of the two portions. In this paper, we set ω_1 = 0.4 and ω_2 = 0.6 based on the performance analyses in Section IV-C. Of course, there is a large room to manipulate the importance weights for better quality prediction. A more meaningful practice may be to explore the proper combination of Q_{style} and Q_{content} that best fits human subjective study.

IV. EXPERIMENTAL RESULTS

A. Evaluation Criteria

Similar to [52]–[54], four criteria are adapted to measure the performance of different methods: the Spearman Rank order Correlation Coefficient (SRCC), Kendall Rank-order Correlation Coefficient (KRCC), Pearson Linear Correlation Coefficient (PLCC), and Hit Rate (HITR). Specifically, the SRCC and KRCC measure the prediction monotonicity. PLCC is utilized to evaluate the prediction linearity after fitting a five-parameter logistic function:

$$f(x) = \kappa_1 \left(\frac{1}{2} - \frac{1}{1 + \exp(\kappa_2(x\text{-}\kappa_3))} \right) + \kappa_4 \cdot x + \kappa_5 \tag{20}$$

where x and $f(x)$ represent the objective and mapped scores respectively, and $\{\kappa_i | i = 1, 2, \ldots, 5\}$ are the five parameters to be fitted. Additionally, HITR can measure the classification accuracy [52], defined as:

$$HITR = R_i / R_n \tag{21}$$

where R_i denotes the number of correct judgments in PC, and R_n is the total number of PC. Considering that the subjective experiments are based on PC within the same group of images,

TABLE II
PERFORMANCE COMPARISON OF CP AND SR FOR THE PROPOSED METHOD WITH DIFFERENT TRAINING DATABASES.

Evaluation	Dictionary	Database	Training images	Training samples for each sub-dictionary	SRCC	KRCC	HITR	PLCC
CP	Dict. I	TID 2013	10	1000 (patches)	0.7906	0.6773	0.8376	0.8637
	Dict. II	NPRgeneral	10	1000 (patches)	0.7903	0.6783	0.8383	0.8637
	Dict. III	TID 2013+NPRgeneral	40	4000 (patches)	0.7900	0.6774	0.8379	0.8638
	Dict. IV	Proposed	10	1000 (patches)	0.7921	0.6807	0.8393	0.8635
SR	Dict. I	TAD66K-art	100	400, 400, 900, 1600, 1600 (vectors)	0.6034	0.4827	0.7386	0.6236
	Dict. II	TAD66K-art	400	1600, 1600, 3600, 6400, 6400 (vectors)	0.6028	0.4816	0.7378	0.6227
	Dict. III	Proposed	100	400, 400, 900, 1600, 1600 (vectors)	0.6062	0.4886	0.7412	0.6278

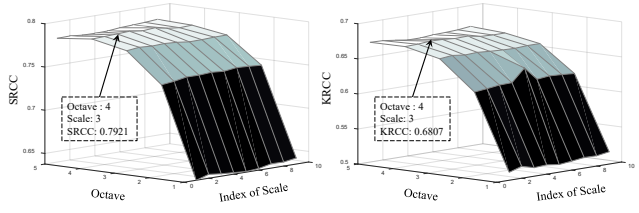


Fig. 14. Impacts of various octave and scale combinations on the performance of the Q_{content} on CP evaluation.

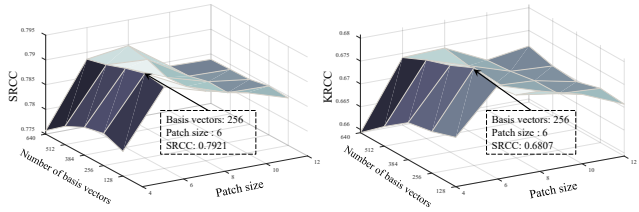


Fig. 15. Impacts of various patch sizes and numbers of basis vectors combinations on the performance of the Q_{content} on CP evaluation.

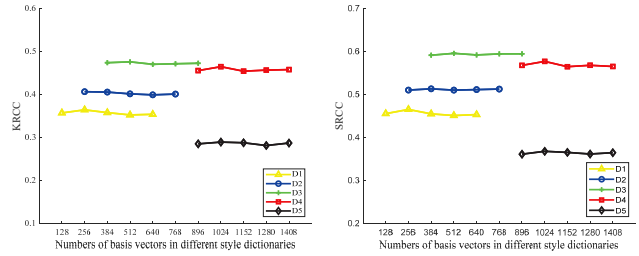


Fig. 16. Impacts of numbers of basis vectors on the performance of the Q_{style} on SR evaluation.

so that the ground truth (B-T scores) are only meaningful within the same group. Therefore, these criteria are computed respectively for each group from the same source image. Then, the average value of all 150 groups is reported as the final performance score. A superior metric should have higher criteria values (with a maximum of 1).

B. Parameter and Training Database Setting

Since the fundamental of proposed quality metrics (i.e., Q_{style}, Q_{content} and Q_{overall}) is the dictionary operating in a multi-scale framework with several parameters, it makes sense to explore the influences of parameters and training databases on the performance evaluation. Therefore, this subsection first tunes the multi-scale dictionaries parameters, and then changes the type and number of images in the training databases.

1) Parameters in Q_{content}: For multi-scale content dictionary, we first visualize the influence of different combinations of scale and octave with at most 10 scales and 5 octaves on the performance of Q_{content} in Fig. 14. It shows that the performance is greatly affected by the number of octaves while slightly affected by the number of scales. The reason is that the CP evaluation is more concerned with semantic structural changes rather than detail fidelity. As shown in the Fig. 13, octave adjustment causes large structural changes, while scale adjustment mainly affects small detail information. Here, we set the number of octaves O = 4 and the number of scales Z = 3, which can achieve the best performance. In addition, we show the performance of tuning the patch size and the number of basis vectors in Fig. 15. From the results, we can observe that the optimal performance is obtained by selecting patch size 6 and the number of basis vectors 256. Therefore, we set T = 36 (patch size = 6) and V = 256 in this work.

2) Parameters in Q_{style}: For multi-scale style dictionary, since the Gram-based style feature vector of each layer is fixed, we only test the effect of the number of basis vectors. As shown in Fig. 16, the evaluation accuracy varies relatively slight over the numbers of basis vectors, which indicates that the Q_{style} does not highly depend on the training configurations. In this paper, we set $U \in \{256, 256, 512, 1024, 1024\}$.

3) Training database: In addition to the parameters in the dictionary, the selection of the training database is also worth analyzing, which can validate whether the performance is dependent on a particular training database. To this end, we train several overcomplete content and style dictionaries based on different training images respectively. The training images are selected from the TID 2013 [49], NPRgeneral [31], and TAD66K [55]. The implementation details and performance results are listed in Table II. Note that the number of training samples is maintained consistent for each content sub-dictionary, while the training samples (i.e., Gram-based style feature vectors) of style sub-dictionaries are set to a different number to match the size of the Gram matrix. The results represent that the performance with different dictionaries are quite similar. This demonstrates that the proposed method is insensitive to the selection of the training databases.

TABLE III
THE IMPACTS OF DIFFERENT POOLING STRATEGY ON THE PERFORMANCE OF THE PROPOSED METHOD.

Pooling strategy	Multiplication (c, d)=(1,0) (w1, w2)					Summation (c, d)=(0,1) (w3, w4)					Combination (w1, w2, w3, w4)=(0.4, 0.6, 0.4, 0.6) (c, d)				
	(0.5,0.5)	(0.6,0.4)	(0.8,0.2)	(0.4,0.6)	(0.2,0.8)	(0.5,0.5)	(0.6,0.4)	(0.8,0.2)	(0.4,0.6)	(0.2,0.8)	(0.5,0.5)	(0.6,0.4)	(0.8,0.2)	(0.4,0.6)	(0.2,0.8)
SRCC	0.5980	0.5666	0.5053	**0.6077**	0.4495	0.5637	0.5397	0.4892	0.5990	0.4894	0.5985	0.5997	0.5984	0.5968	0.6046
KRCC	0.4746	0.4493	0.3920	**0.4855**	0.3563	0.4502	0.4254	0.3792	0.4760	0.3989	0.4779	0.4793	0.4784	0.4754	0.4836
HITR	0.7362	0.7243	0.6964	**0.7410**	0.6776	0.7248	0.7124	0.6900	0.7367	0.6938	0.7374	0.7381	0.7376	0.7360	0.7400
PLCC	0.6668	0.6500	0.5792	**0.6510**	0.4704	0.6445	0.6168	0.5593	0.6561	0.5277	0.6581	0.6584	0.6578	0.6575	0.6551

C. Analysis of Pooling Methods

As mentioned above, there is a complex relationship between OV and other quality factors (e.g., Q_{style} and Q_{content}). Hence, it is meaningful to analyze the impacts of different pooling methods on the performance results. In this connection, we refer to the experimental settings [56] making a modification to Equation (19):

$$\begin{aligned} Q_{\text{overall}} = c \times ({Q_{\text{content}}}^{\omega_1}) \times ({Q_{\text{style}}}^{\omega_2}) + \\ d \times (\omega_3 Q_{\text{content}} + \omega_4 Q_{\text{style}}) \end{aligned} \tag{22}$$

where c and d are utilized to balance the significance of the summation and multiplication pooling items, ω_1, ω_2, ω_3 and ω_4 are set to balance the importance of different quality factors. The performance of the three pooling strategy (i.e., multiplication(c=1, d=0), summation (c=0, d=1) and combination) are listed in the Table III. From the results, we can find that both multiplication and summation strategies achieve the optimal performance when (ω_1, ω_2) and (ω_3, ω_4) are set as (0.4, 0.6), but the performance of multiplication is better than that of summation. Based on the optimal weight setting for the quality factors, the combination pooling strategy obtains similar results among the five (c, d) combinations. In summary, different pooling strategies generate significant impacts on the performance, in contrast, the multiplication pooling strategy with parameter (ω_1=0.4, ω_2=0.6) obtains the optimal results. As a consequence, we employ this multiplication pooling strategy as the final pooling method in this work.

D. Performance Test on Different Quality Factors

For the performance test on the CP, SR and OV quality evaluations, we compare the proposed SRQE with existing IQA metrics, including two categories: (1) Fourteen state-of-the-art general-purpose FR-IQA metrics: SSIM [35], FSIM [26], MS-SSIM [57], IW-SSIM [25], Peak Signal-to-Noise Ratio (PSNR), MAD [58], VIF [24], VSI [27], GMSD [59], UQI [60], IFC [61], RFSIM [62], DISTS [46] and LPIPS [63]. (2) Eight state-of-the-art general-purpose NR-IQA metrics: NIQE [28], TCLT [36], BMPRI [64], BLIINDS-II [65], BRISQUE [37], UNIQUE [66], WaDIQaM [67], TReS [68]. For the traditional learning-based models, we randomly divide the each individual dataset into two non-overlapping subsets (80% for training and 20% for testing). Then, we resort to the support vector regression (SVR) to train the models and report the average results after training-testing process 1000 times. For the deep learning-based methods, we divide the data set into five fixed subsets with non-overlapping contents, and use four subsets for training and one subset for testing at each time to ensure that each image has been tested. For the methods that are training-free or require specific variance of human opinions, we directly report the performance results using the pre-trained model.

1) Performance Test on CP: Although the evaluation of CP between the stylized images and source content images is not a classic FR-IQA problem, since the stylized image targets to maintain the structure information of source content image, the structure measurement module commonly included in existing FR-IQA methods is relatively suitable for comparison. As a consequence, we compare the proposed Q_{content} with the state-of-the-art general-purpose FR-IQA metrics. In addition, we also utilize the NR-IQA method to establish the functional mapping from the stylized images to the CP quality scores. Each content subset conducts the same training-testing strategies described above. The performance compassion results are listed in Table IV, where the top three metrics are highlighted in bold. It can be seen that the traditional general-purpose FR-IQA metrics perform better than NR-IQA metric. It is reasonable since the purpose of CP evaluation is to measure the content structure similarity between the stylized and the original content images. Thus, ignoring the content image and directly extracting features from the stylized image for regression not only lacks enough useful information, but also has no practical significance. Our Q_{content} achieves the best performance for all dataset on the three most important ranking-related performance criteria: SRCC, KRCC and HITR, but is inferior to MS-SSIM [57] on PLCC. The reason is that MS-SSIM [57] also applies the multi-scale feature extraction strategy to simulate the visual characteristics of humans appreciating art works from different scales.

2) Performance Test on SR: To our best knowledge, there is no related methods for evaluating the quality of stylized images on SR. Although SR evaluation cannot be taken as a classical FR/NR-IQA issue, we are curious about how the performance of these methods in AST-IQA task, especially given the lack of comparison methods. Thus, we report the performance comparison results in Table V. It is clear that the above FR-IQA methods are not suitable for quantitative evaluation of SR, because of the difference in contents between the stylized images and the style images. DISTS [46] obtains the better performance among these FR-IQA metrics, since it is designed to evaluate structural and texture (related to style) similarities, allowing for slight pixel misalignment. In addition, although the NR-IQA method can map the stylized

TABLE IV
PERFORMANCE COMPARISON ON CP EVALUATION. '*' INDICATES THAT THE METHOD IS RE-TRAINED ON THE AST-IQAD.

Method		Portrait				Building				Nature				Animal				Daily Life				All			
		SRCC	KRCC	HITR	PLCC	SRCC	KRCC	HITR	PLCC	SRCC	KRCC	HITR	PLCC	SRCC	KRCC	HITR	PLCC	SRCC	KRCC	HITR	PLCC	SRCC	KRCC	HITR	PLCC
NR-IQA	NIQE	0.0490	0.0290	0.5145	0.0258	0.0565	0.0623	0.5310	0.0376	0.0534	0.0699	0.5357	0.0899	0.0506	0.0446	0.5226	0.0529	0.0907	0.0687	0.5357	0.1158	0.0601	0.0549	0.5281	0.0644
	TCLT	0.2546	0.1957	0.5964	0.4184	0.2941	0.2344	0.6131	0.3564	0.3009	0.2252	0.3340	0.6119	0.2398	0.1800	0.5881	0.3812	0.1131	0.0643	0.5321	0.2986	0.2405	0.1799	0.5721	0.3577
	BMPRI*	0.4887	0.3942	0.6966	0.5327	0.4549	0.3485	0.6752	0.5678	0.3672	0.2741	0.6373	0.4255	0.3927	0.2916	0.6452	0.4874	0.4608	0.3414	0.6725	0.5631	0.4396	0.3347	0.5274	0.6678
	BLIINDS-II*	0.3912	0.3225	0.6623	0.4455	0.5610	0.4397	0.7201	0.6399	0.5066	0.3961	0.6970	0.5973	0.4626	0.3609	0.6814	0.5418	0.5660	0.4467	0.7206	0.6983	0.5491	0.4361	0.7180	0.6493
	BRISQUE*	0.4276	0.3387	0.6701	0.4734	0.4633	0.3728	0.6868	0.4872	0.5033	0.4033	0.7025	0.5615	0.3467	0.2667	0.6320	0.4179	0.4987	0.3921	0.6950	0.5330	0.5557	0.4415	0.7202	0.6247
FR-IQA	SSIM	0.6282	0.5178	0.7607	0.7314	0.6664	0.5537	0.7750	0.7639	0.5884	0.4781	0.7381	0.7315	0.6460	0.5306	0.7643	0.7444	0.7561	0.6164	0.8060	0.8450	0.6570	0.5393	0.7688	0.7632
	FSIM	0.7141	0.5895	0.7952	0.7688	0.7154	0.5947	0.7976	0.7738	0.6642	0.5496	0.7750	0.7861	0.6924	0.5688	0.7833	0.7800	0.7903	**0.6668**	0.8321	0.8776	0.7153	0.5939	0.7967	0.7973
	MS-SSIM	**0.7923**	**0.6706**	**0.8357**	**0.8613**	**0.7809**	**0.6637**	**0.8321**	**0.8716**	**0.7349**	**0.6070**	**0.8024**	**0.8557**	**0.7504**	**0.6404**	**0.8190**	**0.8597**	**0.8208**	**0.7000**	**0.8488**	**0.9163**	**0.7759**	**0.6563**	**0.8276**	**0.8729**
	IW-SSIM	**0.7736**	**0.6564**	**0.8286**	**0.8658**	07272	0.6063	0.8048	0.8522	0.6953	0.5714	0.7857	**0.8357**	0.6984	0.5760	0.7869	0.8296	0.7748	0.6477	0.8226	**0.8866**	**0.7339**	**0.6115**	**0.8057**	**0.8540**
	PSNR	0.4150	0.3296	0.6655	0.5268	0.5389	0.4270	0.7119	0.6447	0.4570	0.3613	0.6798	0.5806	0.5860	0.4780	0.7381	0.6917	0.6766	0.5450	0.7690	0.7660	0.5347	0.4282	0.7129	0.6420
	MAD	0.6867	0.5705	0.7857	0.7673	0.7227	0.6117	0.8048	0.8073	0.6793	0.5616	0.7798	0.8031	0.7369	0.6117	0.8048	0.8246	0.7834	0.6547	0.8250	0.8616	0.7218	0.6020	0.8000	0.8128
	VIF	0.7384	0.6088	0.8060	0.7843	0.6248	0.5061	0.7536	0.7455	0.5909	0.4640	0.7321	0.7505	0.6769	0.5424	0.7714	0.7567	0.7214	0.5828	0.7917	0.8137	0.6705	0.5408	0.7710	0.7701
	VSI	0.5960	0.4793	0.7417	0.7295	0.7114	0.5875	0.7929	0.7779	0.7290	0.5926	0.7952	0.8044	0.6706	0.5473	0.7726	0.8128	0.7480	0.6257	0.8119	0.8513	0.6910	0.5665	0.7829	0.7952
	GMSD	0.5941	0.4846	0.7440	0.6587	0.5801	0.4589	0.7298	0.6156	0.6047	0.4878	0.7440	0.6719	0.5527	0.4401	0.7190	0.6676	0.7422	0.6167	0.8071	0.8014	0.6148	0.4976	0.7488	0.6830
	UQI	0.3682	0.2771	0.6405	0.4520	0.5323	0.4384	0.7179	0.6610	0.3949	0.2895	0.6440	0.4883	0.5045	0.3969	0.6976	0.5998	0.5265	0.4137	0.7060	0.6109	0.4653	0.3631	0.6812	0.5624
	IFC	0.7411	0.6112	0.8071	0.7945	0.6601	0.5395	0.7702	0.7597	0.6008	0.4760	0.7381	0.7541	0.6730	0.5401	0.7702	0.7589	0.7214	0.5805	0.7905	0.8179	0.6793	0.5494	0.7752	0.7770
	RFSIM	0.5878	0.4797	0.7405	0.5645	0.6485	0.5249	0.7619	0.6080	0.5606	0.4495	0.7238	0.5468	0.6932	0.5520	0.7750	0.6538	0.7802	0.6546	0.8238	0.7234	0.6541	0.5321	0.7650	0.6193
	DISTS	0.6323	0.5084	0.7548	0.7409	**0.7773**	**0.6611**	**0.8298**	**0.8765**	0.6389	0.5020	0.7500	0.7939	0.6734	0.5380	0.7679	0.8265	0.7795	0.6549	0.8262	0.8695	0.7003	0.5729	0.7857	0.8215
	LPIPS	0.6625	0.5383	0.7714	0.7600	0.7166	0.5863	0.7940	0.7954	**0.7325**	**0.6305**	**0.8143**	0.7931	0.6728	0.5463	0.7714	0.7792	**0.7935**	0.6633	**0.8333**	0.8721	0.7156	0.5929	0.7969	0.7999
AST-IQA	Q_{content}	**0.7871**	**0.6707**	**0.8357**	**0.8280**	**0.7953**	**0.6973**	**0.8464**	**0.8788**	**0.7528**	**0.6285**	**0.8131**	**0.8507**	**0.7750**	**0.6617**	**0.8298**	**0.8517**	**0.8502**	**0.7454**	**0.8714**	**0.9084**	**0.7921**	**0.6807**	**0.8393**	**0.8635**

TABLE V
PERFORMANCE COMPARISON ON SR EVALUATION. '*' INDICATES THAT THE METHOD IS RE-TRAINED ON THE AST-IQAD.

Methods / Criteria	NR-IQA					FR-IQA									AST-IQA
	NIQE	TCLT	BMPRI*	BLIINDS-II*	BRISQUE*	RFSIM	IFC	MAD	VIF	SSIM	IW-SSIM	MS-SSIM	LPIPS	DISTS	Q_{style}
SRCC	0.3845	0.2019	**0.4479**	0.4162	**0.4885**	0.0016	0.1008	0.2769	0.1744	0.3093	0.1243	0.3252	0.2471	0.4233	**0.6062**
KRCC	0.3075	0.1396	**0.3483**	0.3301	**0.3952**	0.0001	0.0725	0.2146	0.1196	0.2300	0.0978	0.2425	0.1955	0.3327	**0.4886**
HITR	0.6538	0.5702	**0.6744**	0.6641	**0.6975**	0.5002	0.5360	0.6076	0.5590	0.6145	0.5469	0.6202	0.5971	0.6631	**0.7412**
PLCC	0.4295	0.2183	**0.5169**	0.4651	**0.5254**	0.0135	0.0890	0.3389	0.1634	0.3270	0.1314	0.3925	0.2608	0.4654	**0.6278**

images to SR quality scores with powerful learning machines, it lacks practical relevance. In addition, it is also doubtful whether the NR-IQA method can maintain high performance when testing more diverse images that are not included in training. Compared with other methods, our Q_{style} achieves the best performance, since it builds a strong association (e.g., style pattern and brush stoke) with AST.

3) Performance Test on OV: Since there is no ground truth for stylized image to directly compare on OV evaluation, instead of using the FR-IQA methods, we utilize the CP/SR quality score of the FR-IQA method for performance comparison, which is beneficial to analyze and explore the contribution of the CP/SR quality components on the overall quality OV. Additionally, we also perform NR-IQA on the OV quality evaluation task, which will be useful to understand how challenging this task is for the existing NR-IQA metrics. The performance comparison results of all methods are shown in Table VI. In the table, the subscript "SR/CP" represents the quality scores generating from the SR/CP evaluation. From the results, we observe that: 1) All CP and SR quality score generated by vanilla FR metrics have weak correlation with the overall quality OV. It indicates that OV quality evaluation is a complex process where multiple factors need to be considered. 2) The learning-based NR-IQA methods show good performance results, in particular, transformer-based TReS [68] achieves competitive results with our Q_{overall}. Obviously, utilizing deep neural network to perform AST-IQAD task is a promising way. 3) It can be seen that Q_{overall} has a higher performance than Q_{style} or Q_{content} alone, which also demonstrates the effectiveness of our pooling strategy. As an unsupervised learning method based on sparse representation, our Q_{overall} can stably evaluate OV quality scores and achieve the best performance on three criteria via properly combining

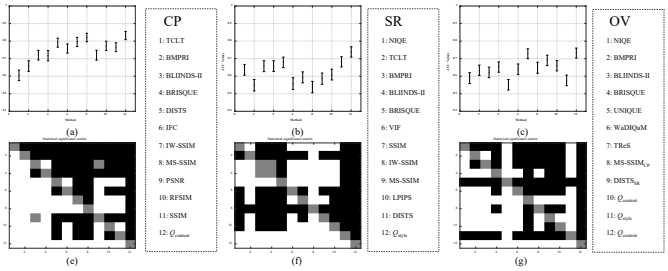


Fig. 17. Statistical significance analyses. (a)-(c): AUC values of different method on CP, SR, and OV evaluation, respectively. (e)-(f): two-sample t-test results (statistical significance matrix) on CP, SR, and OV evaluation, respectively.

the Q_{style} and Q_{content}. Actually, as discussed in Section IV-C, there still leaves a large space for improving the evaluation accuracy via proper importance weights and combination strategies.

E. Statistical Analysis

In the above experiments, the proposed SROE demonstrates a better correlation between the prediction scores and ground truth. We further adopt the hypothesis testing approach based on t-statistics [69] to prove that our SRQE is statistically better than other metrics. Specifically, we first calculate the area under curve (AUC) values (i.e., the area covered by receiver operating characteristic (ROC)) with 95% confidence interval (CI) for all image pairs (from PC in the subjective study). A higher value of AUC indicates better performance of the method. Next, we carry out the two-sample t-test between the pair of AUC values with 95% CI. We show the results of the AUC values and statistical significance of difference in Fig. 17, where the white/gray/black square manifests that the

TABLE VI
PERFORMANCE COMPARISON ON OV EVALUATION. '*' INDICATES THAT THE METHOD IS RE-TRAINED ON THE AST-IQAD. THE SUBSCRIPT "SR/CP" OF FR-IQA METHODS REPRESENT THE QUALITY SCORE GENERATING FROM THE SR/CP EVALUATION.

Type	Method	SRCC	KRCC	HITR	PLCC
NR-IQA	NIQE	0.2615	0.2078	0.6036	0.3041
	TCLT	0.0189	0.0177	0.5081	0.0601
	BMPRI*	0.3705	0.2832	0.6413	0.4739
	BLIINDS-II*	0.3259	0.2467	0.6228	0.4201
	BRISQUE*	0.4124	0.3179	0.6577	0.4685
	UNIQUE	0.2038	0.1507	0.5776	0.3063
	WaDIQaM*	0.3779	0.2840	0.6421	0.4183
	TReS*	**0.5993**	**0.4816**	**0.7398**	**0.6779**
FR-IQA	UQI_{CP}	0.3252	0.2365	0.6176	0.3626
	IFC_{CP}	0.3976	0.2955	0.6483	0.5015
	VIF_{CP}	0.3970	0.2927	0.6469	0.5032
	VSI_{CP}	0.3714	0.2744	0.6369	0.4477
	PSNR_{CP}	0.3275	0.2386	0.6188	0.3791
	MAD_{CP}	0.3482	0.2501	0.6250	0.4122
	SSIM_{CP}	0.3495	0.2530	0.6267	0.4268
	SSIM_{SR}	0.1890	0.1512	0.5757	0.1930
	FSIM_{CP}	0.3749	0.2773	0.6388	0.4399
	GMSD_{CP}	0.2814	0.2062	0.6033	0.3250
	RFSIM_{CP}	0.3585	0.2616	0.6300	0.2968
	RFSIM_{SR}	0.0254	0.0209	0.5095	0.0204
	DISTS_{CP}	0.3626	0.2681	0.6340	0.4875
	DISTS_{SR}	**0.4695**	**0.3693**	**0.6829**	**0.5169**
	$\text{MS-SSIM}_{\text{CP}}$	0.4258	0.3201	0.6598	0.5174
	$\text{MS-SSIM}_{\text{SR}}$	0.2625	0.2063	0.6033	0.3334
	$\text{IW-SSIM}_{\text{CP}}$	0.4177	0.3089	0.6550	0.5258
	$\text{IW-SSIM}_{\text{SR}}$	0.1238	0.0988	0.5486	0.1285
AST-IQA	Q_{content}	0.4520	0.3439	0.6724	0.5163
	Q_{style}	0.2313	0.1779	0.5886	0.2310
	Q_{overall}	**0.6077**	**0.4855**	**0.7410**	**0.6510**

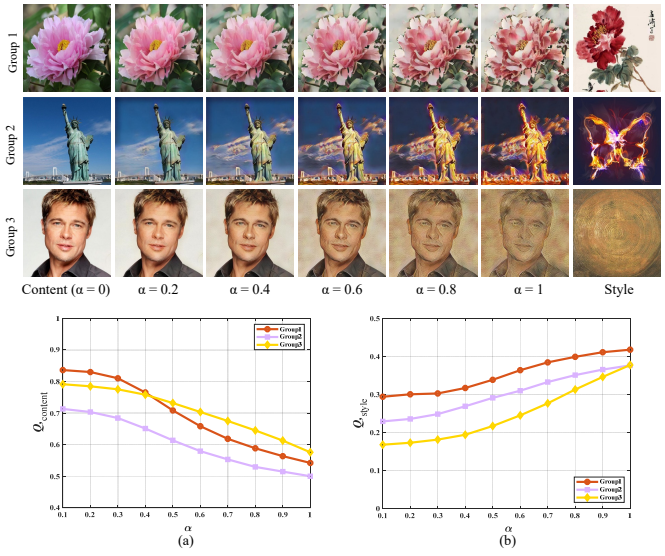


Fig. 18. Robustness analysis of the proposed Q_{style} and Q_{content} in content and style trade-off application.

method in row is significantly better/indistinguishable/worse than the method in column. It can remark from the results that our SRQE performs significantly better than all competitors, indicating the superiority of our SRQE method.

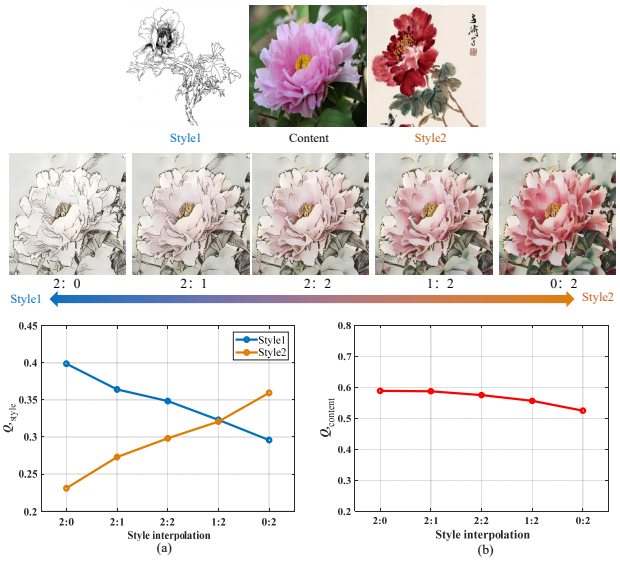


Fig. 19. Robustness analysis of the proposed Q_{style} and Q_{content} in style interpolation application.

F. Robustness Analysis

In this subsection, we present the robustness of our proposed Q_{style} and Q_{content} in two AST applications (i.e., content-style trade-off and style interpolation) which are included in many AST methods.

1) Content-style trade-off: This application can adjust the degree of stylization. Three style degree groups with smooth changes generated by MANet [17] are presented in Fig. 18. When α = 1, the fully stylized image is obtained. Fig. 18 (a)-(b) present Q_{content} and Q_{style} results with different degrees of stylization. It can be seen that as α increases from 0 to 1, the Q_{content} (or Q_{style}) value is consequently decreasing (or increasing) gradually which indicates that our Q_{style} and Q_{content} can effectively capture the changes in the style patterns and content structure of the image.

2) Style interpolation: This application is to merge multiple style images into a single generated result. Here, we also utilize MANet [17] to generate a group of stylized images with different interpolations, and then use Q_{style} (or Q_{content}) to evaluate SR (or CP) of the stylized images. As shown in the Fig. 19 (a), with the continuous decline of the weights for the specific styles, the Q_{style} value is consequently decreasing gradually which indicates that our Q_{style} can clarify style characteristics and accurately evaluate SR even under the interference of multiple styles. Fig. 19 (b) presents that our Q_{content} predicts a set of slowly decreasing quality values as the style changes from Style1 (Line drawing) to Style2 (Xieyi). It is reasonable because the Line drawing pays more attention to the maintenance of structure and shape than Xieyi. The results of the above predictions prove the robustness of our method.

G. Ranking Capability

1) Performance Test on Rank-n accuracy: To comprehensively compare the performance of the IQA metrics, we also

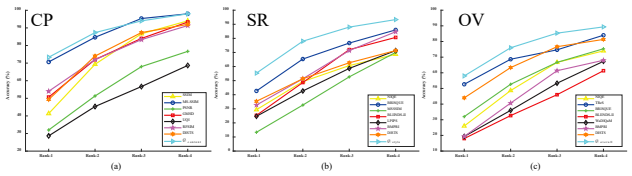


Fig. 20. Rank-n accuracy on the three quality factors.

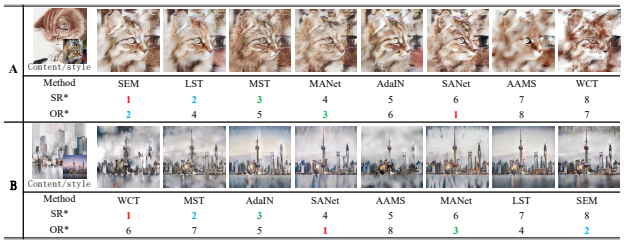


Fig. 21. Some typical failure ranking of our method on the OV evaluation. SR* denotes the subjective ranking, and OR* denotes the objective ranking.

focus on the performance of Rank-n accuracy [70]. Given the eight stylized images for each Content-Style image pair, the rank-n accuracy is the percentage of the objective scores where the subjective-rated best one is within their top n positions. The results are presented in Fig. 20. We can observe that the proposed method achieves a fairly good performance in the SR, CP and OV ranking accuracy tests. Since one of the most important applications of AST-IQA metric is to guide the generation of stylized images, the proposed method is a promising tool for automatic selection of the optimal transferring result from a set of candidates.

2) Failure Ranking Analysis: As aforementioned, we demonstrate the ranking capability of the proposed method on the three quality factors. However, in some special situations, the proposed method encounters challenges to achieve the expected ranking results. As shown in Fig. 21, we present two groups of representative failure rankings on the OV evaluation, which can be divided into two categories. In the first category, our method fails to capture extraneous artifacts, as shown in Fig. 21 (A). We select the stylization produced by SANet [16] as the optimal result, which produces unpleasing eye-like artifacts (zoom in for greater clarity) due to the fine-grained nature. The reason behind this lies in that our method mainly uses global sparse representation and ignores the local extraneous artifacts. In the second category, our method fails to effectively balance the importance of quality factors, as show in Fig. 21 (B). Actually, the CP and SR do not always complement each other. A non-realistic style leads to lower content retention in the final stylization result. Obviously, there is a large room to manipulate the importance weights for the quality factors. Overall, how to further dig the aesthetic information behind the stylization and propose a better strategy to balance different quality factors are the key issues to be explored in the future work.

V. Conclusion

In this paper, we first constructed a new database (AST-IQAD) to collect the subject-rated scores on the three quality factors of content preservation (CP), style resemblance (SR), and overall vision (OV). Then, a new sparse representation-based method (SRQE) is proposed to predict the human perception toward different stylized results. Experimental results show that our proposed method produces very promising AST-IQA results compared with existing general-purpose IQA methods. Overall, our new database creates a reliable platform to evaluate the performance of different AST algorithms and our method is helpful for guiding the design of different algorithms. In our future work, we will further mine the aesthetic information behind the stylization and propose a better strategy to balance different quality factors.

References

[1] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 11, pp. 3365–3385, 2019.

[2] Z. Pang, J. Guo, Z. Ma, W. Sun, and Y. Xiao, "Median stable clustering and global distance classification for cross-domain person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 3164–3177, 2021.

[3] Y. Zhao, Z. Zhong, Z. Luo, G. H. Lee, and N. Sebe, "Source-free open compound domain adaptation in semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 7019–7032, 2022.

[4] Z. Chen, Y. Wang, T. Guan, L. Xu, and W. Liu, "Transformer-based 3d face reconstruction with end-to-end shape-preserved domain transfer," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[5] Y. Gao, X. Feng, T. Zhang, E. Rigall, H. Zhou, L. Qi, and J. Dong, "Wallpaper texture generation and style transfer based on multi-label semantics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1552–1563, 2021.

[6] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, 2001, pp. 341–346.

[7] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

[8] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman, "Controlling perceptual factors in neural style transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3985–3993.

[9] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," *arXiv preprint arXiv:*1701.01036, 2017.

[10] T. Q. Chen and M. Schmidt, "Fast patch-based style transfer of arbitrary style," *arXiv preprint arXiv:*1612.04337, 2016.

[11] S. Gu, C. Chen, J. Liao, and L. Yuan, "Arbitrary style transfer with deep feature reshuffle," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8222–8231.

[12] X. Li, S. Liu, J. Kautz, and M.-H. Yang, "Learning linear transformations for fast image and video style transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3809–3817.

[13] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.

[15] Y. Yao, J. Ren, X. Xie, W. Liu, Y.-J. Liu, and J. Wang, "Attention-aware multi-stroke style transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1467–1475.

[16] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5880–5888.

[17] Y. Deng, F. Tang, W. Dong, W. Sun, F. Huang, and C. Xu, "Arbitrary style transfer via multi-adaptation network," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2719–2727.

[18] Y. Zhang, C. Fang, Y. Wang, Z. Wang, Z. Lin, Y. Fu, and J. Yang, "Multimodal style transfer via graph cuts," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5943–5951.

[19] C. Chen, "Structure-emphasized multimodal style transfer," *MasterTokyo Institute of Technology:* Tokyo, Japan, 2020.

[20] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Stylebank: An explicit representation for neural image style transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1897–1906.

[21] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Transactions on Multimedia*, vol. 24, pp. 3859–3881, 2022.

[22] M.-C. Yeh, S. Tang, A. Bhattad, and D. A. Forsyth, "Quantitative evaluation of style transfer," *arXiv preprint arXiv:*1804.00118, 2018.

[23] Z. Wang, L. Zhao, H. Chen, Z. Zuo, A. Li, W. Xing, and D. Lu, "Evaluate and improve the quality of neural style transfer," *Computer Vision and Image Understanding*, vol. 207, p. 103203, 2021.

[24] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[25] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2010.

[26] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[27] L. Zhang, Y. Shen, and H. Li, "Vsi: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[28] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.

[29] H. Chen, X. Chai, F. Shao, X. Wang, Q. Jiang, M. Chao, and Y.-S. Ho, "Perceptual quality assessment of cartoon images," *IEEE Transactions on Multimedia*, 2021.

[30] T. O. Aydın, A. Smolic, and M. Gross, "Automated aesthetic analysis of photographic images," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 1, pp. 31–42, 2014.

[31] D. Mould and P. L. Rosin, "Developing and applying a benchmark for evaluating image stylization," *Computers & Graphics*, vol. 67, pp. 58–76, 2017.

[32] P. L. Rosin, T. Wang, H. Winnemller, D. Mould, and M. Wand, "Benchmarking non-photorealistic rendering of portraits," in *the Symposium*, 2017, pp. 1–12.

[33] Y. Huang, M. Jing, J. Zhou, Y. Liu, and Y. Fan, "Lccstyle: Arbitrary style transfer with low computational complexity," *IEEE Transactions on Multimedia*, 2021.

[34] Z. Ma, J. Li, N. Wang, and X. Gao, "Image style transfer with collection representation space and semantic-guided reconstruction," *Neural Networks*, vol. 129, pp. 123–137, 2020.

[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[36] Q. Wu, H. Li, F. Meng, K. N. Ngan, B. Luo, C. Huang, and B. Zeng, "Blind image quality assessment based on multichannel feature fusion and label transfer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 425–440, 2015.

[37] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[38] P. Resnick and H. R. Varian, "Recommender systems," *Communications of the ACM*, vol. 40, no. 3, pp. 56–58, 1997.

[39] C. Li and T. Chen, "Aesthetic visual quality assessment of paintings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 236–252, 2009.

[40] N. Kolkin, J. Salavon, and G. Shakhnarovich, "Style transfer by relaxed optimal transport and self-similarity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 051–10 060.

[41] F. Gao, D. Tao, X. Gao, and X. Li, "Learning to rank for blind image quality assessment," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2275–2290, 2015.

[42] R. I.-R. BT, "Methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union*, 2002.

[43] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[44] W.-S. Lai, J.-B. Huang, Z. Hu, N. Ahuja, and M.-H. Yang, "A comparative study for single image blind deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1701–1709.

[45] J. M. Hughes, D. J. Graham, and D. N. Rockmore, "Quantification of artistic style through sparse coding analysis in the drawings of pieter bruegel the elder," *Proceedings of the National Academy of Sciences*, vol. 107, no. 4, pp. 1279–1283, 2010.

[46] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022.

[47] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 689–696.

[48] F. Shao, K. Li, W. Lin, G. Jiang, M. Yu, and Q. Dai, "Full-reference quality assessment of stereoscopic images by learning binocular receptive field properties," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 2971–2983, 2015.

[49] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti et al., "Color image database tid2013: Peculiarities and preliminary results," in *European Workshop on Visual Information Processing (EUVIP)*, 2013, pp. 106–111.

[50] T. Lindeberg, "Scale-space theory: A basic tool for analyzing structures at different scales," *Journal of Applied Statistics*, vol. 21, no. 1-2, pp. 225–270, 1994.

[51] Y. Fu, H. Zeng, L. Ma, Z. Ni, J. Zhu, and K.-K. Ma, "Screen content image quality assessment using multi-scale difference of gaussian," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2428–2432, 2018.

[52] B. Hu, L. Li, H. Liu, W. Lin, and J. Qian, "Pairwise-comparison-based rank learning for benchmarking image restoration algorithms," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 2042–2056, 2019.

[53] Z. Peng, Q. Jiang, F. Shao, W. Gao, and W. Lin, "LGGD+: Image retargeting quality assessment by measuring local and global geometric distortions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3422–3437, 2022.

[54] Q. Jiang, Y. Gu, C. Li, R. Cong, and F. Shao, "Underwater image enhancement quality evaluation: Benchmark dataset and objective metric," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5959–5974, 2022.

[55] S. He, Y. Zhang, R. Xie, D. Jiang, and A. Ming, "Rethinking image aesthetics assessment: Models, datasets and benchmarks," in *Proceeding of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022.

[56] Y. Zhang, H. Zhang, M. Yu, S. Kwong, and Y.-S. Ho, "Sparse representation-based video quality assessment for synthesized 3d videos," *IEEE Transactions on Image Processing*, vol. 29, pp. 509–524, 2019.

[57] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, vol. 2. Ieee, 2003, pp. 1398–1402.

[58] E. C. Larson and D. M. Chandler, "Most apparent distortion: Fullreference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, 2010.

[59] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2013.

[60] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[61] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117– 2128, 2005.

[62] L. Zhang, L. Zhang, and X. Mou, "RFSIM: A feature based image quality assessment metric using riesz transforms," in *2010 IEEE International Conference on Image Processing*, 2010, pp. 321–324.

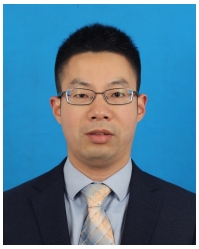
[63] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[64] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 508–517, 2018.

[65] 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.

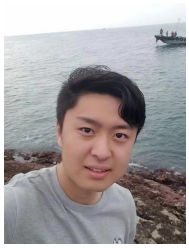
[66] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, 2021.
[67] S. Bosse, D. Maniry, K.-R. M¨uller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2017.
[68] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and selfconsistency," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1220–1230.
[69] Y. Fang, H. Zhu, K. Ma, Z. Wang, and S. Li, "Perceptual evaluation for multi-exposure image fusion of dynamic scenes," *IEEE Transactions on Image Processing*, vol. 29, pp. 1127–1138, 2019.
[70] Y. Zhang, W. Lin, Q. Li, W. Cheng, and X. Zhang, "Multiple-level feature-based measure for retargeted image quality," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 451–463, 2017.

Hangwei Chen received the B.S. degree from Ningbo University, China, in 2020. He is currently pursuing the Ph.D. degree in Signal and Information Processing at Ningbo University, Ningbo, China. His current research interests include image processing and quality assessment.



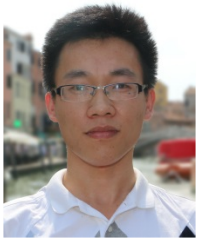
Feng Shao (M'16) received his B.S. and Ph.D degrees from Zhejiang University, Hangzhou, China, in 2002 and 2007, respectively, all in Electronic Science and Technology. He is currently a professor in Faculty of Information Science and Engineering, Ningbo University, China. He was a visiting Fellow with the School of Computer Engineering, Nanyang Technological University, Singapore, from February 2012 to August 2012. He received 'Excellent Young Scholar' Award by NSF of China (NSFC) in 2016. He has published over 100 technical articles in refereed journals and proceedings in the areas of 3D video coding, 3D quality assessment, and image perception, etc.

Xiongli Chai received the B.S. degree and M.S. degree from Ningbo University, China, in 2017 and 2020 respectively. He is currently pursuing the Ph.D. degree in Signal and Information Processing at Ningbo University, Ningbo, China. His current research interests include image/video processing and quality assessment.



Yuese Gu received the B.S. degree in communication engineering from Ningbo University, Ningbo, China, where he is currently pursuing the master's degree. His research interest lies in deep learning with applications in image processing and computer vision.

Qiuping Jiang (M'18) received the Ph.D. degree in signal and information processing from Ningbo University in 2018. From January 2017 to May 2018, he was a Visiting Student with Nanyang Technological University, Singapore. He is currently an Associate Professor with Ningbo University. His research interests include image processing, visual perception, and computer vision. He has received the 2017 Best Paper Honorable Mention Award of the Journal of Visual Communication and Image Representation. He serves as an Associate Editor for IET Image Processing, Journal of Electronic Imaging, and APSIPA Transactions on Signal and Information Processing.



Xiangchao Meng (M'18) received the B.S. degree in geographic information system from the Shandong University of Science and Technology, Qingdao, China, in 2012, and the Ph.D. degree in cartography and geography information system from Wuhan University, Wuhan, China, in 2017. He is currently a Lecturer with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China. His research interests include variational methods and remote sensing image fusion.

Yo-Sung Ho (SM'06–F'16) received the B.S. and M.S. degrees in electronic engineering from Seoul National University, Seoul, Korea, in 1981 and 1983, respectively, and the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, in 1990. He joined Electronics and Telecommunications Research Institute (ETRI), Daejon, Korea, in 1983. From 1990 to 1993, he was with Philips Laboratories, Briarcliff Manor, NY, where he was involved in development of the Advanced Digital High-Definition Television (AD-HDTV) system. In 1993, he rejoined the technical staff of ETRI and was involved in development of the Korean DBS digital television and high-definition television systems. Since 1995, he has been with Gwangju Institute of Science and Technology (GIST), Gwangju, Korea, where he is currently Professor of Information and Communications Department. His research interests include digital image and video coding, image analysis and image restoration, advanced video coding techniques, digital video and audio broadcasting, three-dimensional video processing, and content-based signal representation and processing.