Single Image Super-Resolution Quality Assessment: A Real-World Dataset, Subjective Studies, and An Objective Metric

Qiuping Jiang, *Member, IEEE,* Zhentao Liu, Ke Gu, *Member, IEEE,* Feng Shao, *Member, IEEE,* Xinfeng Zhang, *Senior Member, IEEE,* Hantao Liu, *Senior Member, IEEE,* and Weisi Lin, *Fellow, IEEE*

***Abstract*—Numerous single image super-resolution (SISR) algorithms have been proposed during the past years to reconstruct a high-resolution (HR) image from its low-resolution (LR) observation. However, how to fairly compare the performance of different SISR algorithms/results remains a challenging problem. So far, the lack of comprehensive human subjective study on large-scale real-world SISR datasets and accurate objective SISR quality assessment metrics makes it unreliable to truly understand the performance of different SISR algorithms. We in this paper make efforts to tackle these two issues. Firstly, we construct a real-world SISR quality dataset (i.e., *RealSRQ*) and conduct human subjective studies to compare the performance of the representative SISR algorithms. Secondly, we propose a new objective metric, i.e., *KLTSRQA*, based on the Karhunen-Loéve Transform (KLT) to evaluate the quality of SISR images in a no-reference (NR) manner. Experiments on our constructed *RealSRQ* and the latest synthetic SISR quality dataset (i.e., *QADS*) have demonstrated the superiority of our proposed *KLTSRQA* metric, achieving higher consistency with human subjective scores than relevant existing NR image quality assessment (NR-IQA) metrics. The dataset and the code will be made available at https://github.com/Zhentao-Liu/RealSRQ-KLTSRQA.**

***Index Terms*—Single image super-resolution, real-world, image quality assessment, no-reference metric, Karhunen-Loéve Transform.**

I. Introduction

Single image super-resolution (SISR) aims to reconstruct a latent high-resolution (HR) image from its corresponding low-resolution (LR) observation. Although SISR has been a long-standing problem in the community with a great progress made due to the progress of deep learning techniques over the past several years [1], how to fairly compare the performance of different SISR algorithms remains challenging.

Although the existing works generally conduct both qualitative evaluation and quantitative evaluation to compare different SISR algorithms, they suffer from several limitations. First, in terms of qualitative evaluation, only a limited number of images are presented for visual comparison and more importantly the selected visual images often vary in different works, hereby making the qualitative evaluation not convincing. Second, in terms of quantitative evaluation, several existing image quality assessment (IQA) metrics are adopted, without considering the suitability of these metrics for evaluating SISR results. Therefore, researchers may choose diverse IQA metrics to support their own methods, making it non-trivial and unfair to compare different SISR results objectively. Third, many SISR methods are only validated using synthetic LR images. However, due to the intrinsic discrepancies between synthetic and authentic degradations, evaluation on synthetic data does not necessarily reflect the true performance on real-world SISR with authentic degradations. Overall, the lack of comprehensive human subjective study on large-scale real-world SISR datasets and well-performed objective SISR quality metrics makes it impossible to fully understand the performance of different SISR algorithms.

In this paper, we make efforts to address the above problems. We firstly construct a real-world SISR quality dataset (*RealSRQ*) with comparative human subjective studies to compare the performance of several representative SISR algorithms. The ranking scores, obtained from our comparative subjective studies, are used as the ground truth scores indicating the perceived quality of SISR images. Then, we propose a new objective metric, i.e., *KLTSRQA*, based on the Karhunen-Loéve Transform (KLT) to evaluate the quality of SISR images in a no-reference (NR) manner. Finally, we conduct performance evaluation on our constructed *RealSRQ* and the latest synthetic SISR quality evaluation dataset (i.e., *QADS*). The experimental results demonstrate the superiority of our proposed *KLTSRQA* metric, achieving higher consistency with human subjective scores in comparison with all the competing NR-IQA metrics. To highlight, the main contributions of this work are twofold:

1) Benchmark dataset. We construct the first real-world

This work was supported in part by the Zhejiang Natural Science Foundation under Grant LR22F020002, in part by the Natural Science Foundation of China under Grants 61901236, 62071261, 62076013, 62071449, and U20A20184. in part by the Beijing Natural Science Foundation under Grant JQ21014, in part by the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant SJLZ2020003, and in part by the Fundamental Research Funds for the Central Universities. *(Corresponding author: Zhentao Liu)*

Qiuping Jiang, Zhentao Liu, and Feng Shao are with the School of Information Science and Engineering, Ningbo University, Ningbo 315211, China (e-mail: jiangqiuping@nbu.edu.cn, zhentaoliu@163.com, shaofeng@nbu.edu.cn).

Ke Gu is with Faculty of Information Technology, Beijing University of Technology, Engineering Research Center of Intelligent Perception and Autonomous Control, Ministry of Education, Beijing Laboratory of Smart Environmental Protection, Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing Artificial Intelligence Institute, Beijing 100124, China (e-mail: guke.doctor@gmail.com).

Xinfeng Zhang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China (e-mail: xfzhang@ucas.ac.cn).

Hantao Liu is with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K (e-mail: liuh35@cardiff.ac.uk).

Weisi Lin is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: wslin@ntu.edu.sg).

SISR quality dataset called *RealSRQ* with diverse scene types, scaling factors, and SISR algorithms. The subjective scores in *RealSRQ* are obtained via comparative human subjective studies. Consequently, *RealSRQ* provides a reliable platform to fairly compare the performance of different IQA metrics for evaluating SISR images.

2) Objective metric. We propose a new objective NR quality metric called *KLTSRQA* that can evaluate the quality of SISR images with high accuracy. Extensive performance comparisons with 15 relevant existing NR-IQA metrics (including two dedicated NR metrics for SISR images) on two datasets (i.e., *RealSRQ* and *QADS*) demonstrate the superiority of *KLTSRQA*.

The rest of this paper is organized as follows. Section II introduces related works. Section III illustrates the details of *RealSRQ*. Section IV presents the proposed *KLTSRQA* metric and performance comparisons. Finally, conclusions are drawn in Section IV.

II. Related Works

A. SISR Datasets

1) Synthetic SISR datasets: In the literature, there are several commonly used synthetic datasets for training and testing SISR algorithms. They are Set5 [2], Set14 [3], BSD500 [4], Train91 [5], Urban100 [6] and DIV2K [7, 8]. These datasets only provide the pristine images. In order to get the corresponding LR images, one need to down-sample the pristine images to different scales. A detailed comparison of the existing synthetic datasets are provided in Table I and corresponding brief descriptions of them are as follows:

Set5 [2]: It includes 5 pristine images with low resolution.

Set14 [3]: It includes 14 pristine images with low resolution. Compared with Set5 [2], the variation of image content is larger.

BSD500 [4]: The Berkeley Segmentation Dataset (BSD) is used for image segmentation and contour detection. It totally contains 500 pristine images with diverse scenes in the real-world.

Train91 [5]: It includes 91 pristine images. The images are mainly about flowers and other natural scenes.

Urban100 [6]: It includes 100 pristine images which are mainly about city buildings, including abundant structures in the real-world.

DIV2K [7, 8]: It includes 1,000 HR pristine images at 2K resolution collected from the Internet. This dataset owns rich image contents with high resolution and has been used in the NTIRE 2017 SR Challenge.

2) Real-world SISR datasets: In recent years, real-world SISR has drawn increasing attention. Some real-world SISR datasets have been constructed to train and test SISR algorithms. They are RealSR [9], SR-RAW[10], DRealSR [11], and ImagePairs [12]. Different from the synthetic datasets, real-world SISR datasets provide both HR images and the corresponding LR images captured in the real world instead of synthesized from the pristine HR image using simple degradation model. A comparison of the existing real-world SISR datasets are provided in Table II and corresponding brief descriptions of them are as follows:

TABLE I
Comparison of Existing Synthetic SISR Datasets.

Dataset	Year	HR images	Characteristics
Set5 [2]	BMVC2012	5	Low resolution Limited content
Set14 [3]	LNCS2010	14	Low resolution Limited content
BSD500 [4]	TPAMI2011	500	Image segmentation Contour detection
Train91 [5]	CVPR2008	91	Natural image Flower
Urban100 [6]	CVPR2015	100	Urban image Building structure
DIV2K [7, 8]	CVPRW2017	1000	High resolution Rich content

TABLE II
Comparison of Existing Real-World SISR Datasets.

Dataset	Year	Scaling Factors	Characteristics
RealSR [9]	ICCV2019	$\times 2, \times 3, \times 4$	Focal length adjusting
SR-RAW [10]	CVPR2019	$\times 4, \times 8$	Focal length adjusting Raw sensor data
DRealSR [11]	ECCV2020	$\times 2, \times 3, \times 4$	Focal length adjusting
ImagePairs [12]	CVPRW2020	$\times 2$	Beam splitter

RealSR [9]: It includes 595 HR images and corresponding real-world LR images at three different scaling factors, i.e., $\times 2$, $\times 3$, and $\times 4$. The authors apply two DSLR camera (i.e., Nikon D810 and Canon 5D3) to capture various scenes in real world. Both the HR and LR images are captured by adjusting focal length at the same scene. Then, a new image registration approach is designed to align the LR-HR image pairs.

SR-RAW [10]: Similar to RealSR [9], the LR-HR image pairs in SR-RAW are also captured by adjusting focal length. However, different with RealSR [9], SR-RAW provides raw sensor data and HR RGB images because it is used for SR from raw data. In total, 500 seven-image sequences are taken in both indoor and outdoor scenes under seven different optical zoom settings using a 24-240mm zoom lens (i.e., Sony FE 24-240mm).

DRealSR [11]: DRealSR is also similar with RealSR [9], with a larger scale. Five DLSR cameras (i.e., Sony, Canon, Olympus, Nikon and Panasonic) are used to capture the LR-HR image pairs at four resolutions in both indoor and outdoor scenes. The SIFT algorithm is used to align the image contents with different resolutions. DRealSR totally contains 884, 783, and 840 LR-HR image pairs for the scaling factors $\times 2$, $\times 3$, and $\times 4$, respectively.

ImagePairs [12]: In this dataset, the authors apply a beam-splitter to make two cameras (i.e., a low-resolution camera and a high-resolution camera) capture images of the same scene simultaneously. The pixel-wise aligned LR-HR image pairs are obtained by applying a four-step process: ISP, image undistortion, pair alignment, and margin cropping. This dataset totally contains a collection of 11,421 LR-HR image pairs with a single scaling factor $\times 2$.

B. SISR Quality Datasets

Different from the SISR datasets used for training and testing SISR algorithms, an SISR quality dataset aims to

TABLE III
COMPARISON OF EXISTING SISR QUALITY DATASETS. INCLUDE OR EXCLUDE INDICATE WHETHER DEEP-BASED SISR METHODS ARE INVOLVED OR NOT.

Dataset	Yang et al. [13]	Ma et al. [14]	SISRSet [15]	QADS [16]	RealSRQ
Year	ECCV2014	CVIU2016	NC2019	TIP2019	-
Synthetic/Real	Synthetic	Synthetic	Synthetic	Synthetic	Real
HR Images	10	30	15	20	60
DS parameters	9	6	3	3	N.A.
Scaling factors	2,3,4	2,3,4,5,6,8	2,3,4	2,3,4	2,3,4
LR Images	90	180	45	60	180
SR methods	6 (Exclude)	9 (Exclude)	8 (Include)	21 (Include)	10 (Include)
SR Images	540	1,620	360	980	1,620
User study	ACR	ACR	PC	PC	PC

provide a platform for comparing different objective IQA metrics. The construction of SISR quality dataset generally involves human subjective studies to provide subjective quality scores of SISR images, such that the performance of different objective IQA metrics can be fairly compared by measuring how well they can predict the subjective quality scores. To our best knowledge, the earliest SISR quality datasets are built by Yang et al. [13] and Ma et al [14]. After that, SISRset [15] and QADS [16] were also constructed. A brief comparison of existing SISR quality datasets are listed in Table III and corresponding detailed descriptions of them are as follows:

Yang et al. [13]: As the first effort on this problem, this dataset contains 10 HR images based on which 90 LR images are generated at three scaling scales, i.e., $\times 2$, $\times 3$, and $\times 4$. Six classical (non-deep) SISR algorithms are applied to the LR images, obtaining 540 SISR results in total. The subjective user study is conducted in a absolute category rating (ACR) manner, generating a subjective quality score for each SISR image. Note that the evaluated SISR algorithms used in this study do not include deep learning-based SISR algorithms and the obtained SISR results are gray-scale.

Ma et al. [14]: This dataset is an extension of Yang et al.'s [13], with more HR images, scaling factors, and SISR algorithms. Specifically, it includes 30 HR images and the corresponding 180 LR images at six scaling factors, i.e., $\times 2$, $\times 3$, $\times 4$, $\times 5$, $\times 6$, and $\times 8$. Nine classical (non-deep) SISR algorithms are applied to the 180 LR images, thus obtaining 1,620 SISR results in total. Subjective user study is also conducted via ACR, finally generating a subjective score for each SISR image.

SISRset [15]: To investigate if the widely used metrics can well assess the DNN-based SISR results, Shi et al. [15] construct the SISRset. It contains 15 HR images and 45 LR images at three scaling factors, i.e., $\times 2$, $\times 3$, and $\times 4$. Four non-deep SISR algorithms and Four DNN-based SISR algorithms are applied to reconstruct the 45 LR images, and thus obtaining 360 SISR results in total. It applies pairwise comparison for human subjective study. Through intra-scaling comparisons and cross-scaling comparisons, each SISR result gets a mean opinion score (MOS).

QADS [16]: This dataset includes 20 HR images and 60 LR images at three scaling scales, i.e., $\times 2$, $\times 3$, and $\times 4$. A total number of 15 non-deep SISR algorithms and 6 DNN-based SISR algorithms are applied to the 60 LR images at three scaling scales, i.e., $\times 2$, $\times 3$, and $\times 4$. And it gets 980 SISR results in total. Note that not each algorithm is applied to all three scaling factors. It also applies pairwise comparison in its subjective study. It conducts comparisons across different SISR algorithms and different scaling factors in the same scene. Finally, each SISR result gets a MOS.

However, the common problem of these datasets is that the involved LR images are all synthetic ones, i.e., generated from the pristine HR images with a simple degradation model, rather than captured in the real world. Due to the intrinsic discrepancies between synthetic and authentic degradations, it is required to revisit this problem with real-world images. As far as we have known, the construction of *real-world SISR quality dataset* remains untouched.

C. IQA Metrics

Objective IQA aims to automatically evaluate the perceptual quality of distorted image in consistent with human subjective perception. Objective IQA metrics are roughly classified into three categories: full-reference (FR), reduce-reference (RR), and no-reference (NR). We mainly introduce the FR and NR metrics due to their wide applications in SISR studies.

FR-IQA metrics treat the reference image as the ideal one with perfect quality and compute the distance/similarity between the reference and distorted images as quality score. The most popular FR metric is PSNR which is highly efficient. However, it suffers from low correlation with human perception. SSIM [17] brings FR-IQA from pixel-wise error visibility to structural similarity. MS-SSIM [18] and IW-SSIM [19] improved SSIM from the perspectives of multi-scale mechanism and information content weighting, respectively. IFC [20] proposes a novel information fidelity criterion. VIF [21] is an extension of IFC [20] which quantifies the mutual information between reference and distorted images. FSIM [22] combines phase congruency and gradient magnitude for feature similarity calculation. GMSD [23] uses global variation of gradient based local quality map and applies standard deviation for pooling. Inspired by the internal generative mechanism theory, IGM [24] adopts an autoregressive prediction algorithm to decompose an image into order and disorder portions for separate quality calculation. VSI [25] uses visual saliency as feature to compute the local quality map and a weighting function to get the final quality score. Inspired by the fact that human visual system (HVS) is highly sensitive to edges, ESIM [26] extracts three salient edge features, i.e., edge contrast, edge width, and edge direction to assess screen content images (SCIs) quality. To conduct performance evaluation, a new SCI database is also established in this work. Later, a FR-IQA metric for SCIs called GFM [27] is proposed by using the Gabor filter response features.

NR-IQA is more challenging due to the lack of reference image. Most of the current NR metrics share a common two-step procedure. First, extracting image quality-related features from the distorted image. Second, using regression tools to map the extracted features to subjective scores. The main differences of these NR metrics lie in the extracted features. Some representative two step-based NR-IQA metrics include GM-LOG [28], BLIINDS-II [29], CurveletQA [30], BRISQUE

[31], OG-IQA [32], SSEQ [33], DIIVINE [34], RISE [35], BMPRI [36], FRIQUEE [37]. In sharp contrast with the regression-based NR-IQA metrics, NIQE [38], IL-NIQE [39], and HVS-MaxPol [40] are training-free. There are also some NR quality metrics specifically designed for image restoration and image super-resolution. PCRL [41] presents a pairwise-comparison-based rank learning framework for benchmarking image restoration algorithms. SR-metric [14] extracts features from three aspects: local frequency features, global frequency features, and spatial features, for blind quality assessment of image super-resolution.

III. RealSRQ: A Real-World SISR Quality Dataset

A. Real-World LR-HR Image Pairs

We aim to compare the performance of different SISR algorithms for real-world LR images. In this work, the real-world LR images are collected either from the RealSR dataset [9] or captured by ourselves following the same method in [9], i.e., adjust the focal length of a fixed digital single-lens reflex camera (*i.e.*, Canon 5D3). For each scene, images are taken using four focal lengths: 105mm, 50mm, 35mm, and 28mm. Images taken by the largest focal length are used to generate the HR images, and images taken by the other three focal lengths are used to generate the three LR versions. Due to the influence of lens distortion and optical center shift caused by focal length adjustment, the same image registration approach in [9] is adopted to align the LR-HR image pairs. Overall, we collect 60 HR images and corresponding 180 LR images at three different scales, including 60 for scale 2 (↓2), 60 for scale 3 (↓3), and 60 for scale 4 (↓4), respectively.

B. SISR Algorithms

The existing SISR algorithms can be classified into two branches: non-deep models and deep models. In this study, we evaluate 10 representative SISR algorithms, including BCI, ASDS [42], SPM [43], Aplus [44], AIS [45], SRCNN [46], CSCN [47], VDSR [48], SRGAN [49], and USRnet [50]. The set of SISR methods considered in our study equally samples from the two branches, i.e., the former five methods are non-deep methods while the latter five methods are deep learning-based SISR methods, and covers recent major publications in the field (either to be widely used or the latest ones). Especially, the USRnet is reported to be suitable for real-world SISR. As a result, we finally obtain 1,620 SR images in total. Table IV lists the evaluated SISR methods, the implemented scaling factors, and the number of generated SR images by each method. Note that we only implement SPM, AIS, and SRGAN methods under two scaling factors because the released codes or priors only support these two scaling factors.

C. Human Subjective Study

Human subjective study aims to compare different SISR results according to human visual perception. Early datasets use the ACR method where the test images are presented one at a time and are rated independently on a discrete category scale (e.g., ITU 5-point quality scale). Nevertheless, the ACR method will result in a huge bias and uncertainty when the observers do not have sufficient experience. For this consideration, the pairwise comparison (PC), which aims to provide a binary preference label between a pair of stimuli instead of rating an absolute quality level to a single stimulus, is adopted.

TABLE IV
List of the used SISR methods and the Corresponding Number of SR Images.

SISR methods		Scaling factor	No. of SR images
Non-deep models	BCI	2,3,4	180
	ASDS [42]	2,3,4	180
	SPM [43]	2,3	120
	Aplus [44]	2,3,4	180
	AIS [45]	2,3	120
Deep models	SRCNN [46]	2,3,4	180
	CSCN [47]	2,3,4	180
	VDSR [48]	2,3,4	180
	SRGAN [49]	2,4	120
	USRnet [50]	2,3,4	180
Total			1,620

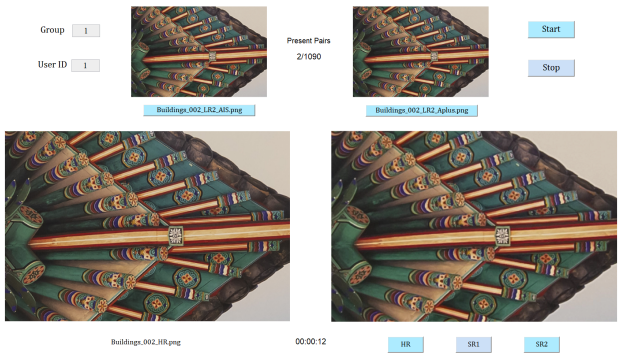

Fig. 1. Our designed GUI for pairwise comparison (PC).

Our PC-based human subjective study is detailed as follows. First, keeping in mind that our goal is to fairly compare different SISR algorithms, it is only meaningful to generate comparison pairs from the same scene and the same scaling factor. For scaling factor $\times 2$, we have 10 SISR results and thus $\binom{10}{2} = 45$ pairwise comparisons per scene. For scaling factor $\times 3$, we have 9 SISR results and thus $\binom{9}{2} = 36$ pairwise comparisons per scene. For scaling factor $\times 4$, we have 8 SISR results and thus $\binom{8}{2} = 28$ pairwise comparisons per scene. That is, $45 + 36 + 28 = 109$ pairwise comparisons are involved for each scene and a total number of 109 per scene$\times$60 scenes$=$ 6540 pairwise comparisons are involved for all scenes. Then, PC is performed via a customized GUI. A screenshot of our designed GUI is illustrated in Fig. 1. At the beginning of the experiment, participants would input their group number and user ID at the top-left corner. If they would like to start, they could press the "Start" button at the top-right corner. They can also press the "Stop" button for a rest. There are four image windows show on the screen simultaneously. The top row presents two SISR results to be compared. The HR image is shown in the bottom-left window. As suggested in [16], subjects can make their decisions more quickly and precisely by flipping the three images at the same position. Thus, we use the bottom-right window to show these three

TABLE V
ADDITIONAL QUESTIONNAIRE.

Factors	ID	Reason Descriptions
Noise	1	Another picture introduces more noises and checkerboards.
Detail	2	Another picture loses more details and its details are blurrier.
Contour	3	The contours of objects are not clear, with more severe shaking, ghosting, blurring or jaggies.
	4	The contours of objects are more distorted.
Texture	5	Textures are not clear, with more severe shaking, ghosting, blurring or jaggies.
	6	Textures are more distorted.
Color	7	The colors are dimmer, lighter, lower contrast, less saturated and look blurrier.
	8	Another picture shows a more severe discoloration.
Other	9	No specific reason. Another picture is just worse.

images for finer comparison. Participants can switch different images by pressing the three buttons "HR", "SR1", "SR2" below the bottom-right window. After careful comparison, subjects can make their choices by pressing the corresponding button below the SR image windows to submit a binary preference label, i.e., "1" or "-1".

A total number of 60 subjects, including 32 males and 28 females, participated in our human subjective study. Before the experiment, we train them adequately to ensure that they are familiar with the background knowledge and the GUI. We divide the 60 participants into 6 groups, i.e., 10 participants per group. Each group is responsible for one scene type attribute group. Thus, each participant should finish 1,090 PC voting (109 pairs per scene $\times 10$ scenes per group$= 1,090$) and every comparison pair is voted by 10 times. As a result, we can get 65,400 votes totally ($6,540$ pairs $\times 10$ votes per pair$= 65,400$). Every time they finish 100 pair comparisons, they are asked to have a rest for 5 minutes to avoid the influence of potential visual fatigue. Each participant will take about 172 minutes to complete the subjective experiment.

To further ensure the reliability of human subjective studies, we also set check points to avoid random selections. Each participant would go through 30 check points in which the correct choice is easy to select. Each check point is a comparison pair consisting of two SR results from the same scene and the same SISR algorithm, but at two different scaling factors. As long as one fails the check point more than twice, his/her votes will be discarded. Fortunately, all of our participants successfully pass the check points.

To get more insights on the pros and cons of different SISR algorithms, we are also interested in asking the participants "Why don't you like another picture?" We provide nine reasons to form the questionnaire, as shown in Table V. These reasons are put forward based on the diverse factors that we think would affect the quality of SR results. To reduce the experiment period, the questionnaire only appears randomly on the GUI with a probability of $1/6$.

D. Subjective Study Result Analysis

This section performs comprehensive statistical analyses on the results obtained from our PC-based subjective studies, including global ranking of SISR algorithms, convergence analysis, and human preference analysis.

1) Global Ranking of SISR Algorithms: We adopt the Bradley-Terry model [51] to derive the global ranking of SISR algorithms from the corresponding PC results.

First, we define C_{ij} as

$$C_{ij} = \begin{cases} \text{number that } i \text{ beats } j, & i \neq j \\ 0, & i = j \end{cases} \tag{1}$$

where C_{ij} denotes the number that method i beats method j. Suppose there are M SISR methods and their subjective rating scores are denoted by $\mathbf{s} = [s_1, s_2, \cdots, s_M]$. Based on the results of PC, we can construct a winning matrix $\mathbf{C} \in \mathbb{R}^{M \times M}$, where each element is defined by Eq. (1). Suppose the probability of users prefer method i over method j is

$$P_{ij} = \frac{e^{s_i}}{e^{s_i} + e^{s_j}} \tag{2}$$

Then, the probability of $\mathbf{s}$ is

$$P(\mathbf{s}) = \prod_{i=1}^{M} \prod_{\substack{j=1 \\ j \neq i}}^{M} (P_{ij})^{C_{ij}} \tag{3}$$

By minimizing the negative log likelihood of $P(\mathbf{s})$, we can obtain an estimation of $\mathbf{s}$. The negative log likelihood of $P(\mathbf{s})$ is expressed as

$$L(\mathbf{s}) = -\sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} C_{ij} \log P_{ij} \tag{4}$$

The partial derivative of $L(\mathbf{s})$ with respect to s_k is

$$\frac{\partial L(\mathbf{s})}{\partial s_k} = \sum_{\substack{i=1 \\ i \neq k}}^{M} \frac{(C_{ki} + C_{ik}) \cdot e^{s_k}}{e^{s_k} + e^{s_i}} - \sum_{\substack{j=1 \\ j \neq k}}^{M} C_{kj} \ , k = 1, 2, \cdots, M \tag{5}$$

Let $\frac{\partial L(\mathbf{s})}{\partial s_k} = 0$, and the $t+1$-th iteration of s_k, $k = 1, 2, \cdots, M$ is obtained as

$$s_k^{t+1} = \log \left(\frac{\sum_{\substack{j=1 \\ j \neq k}}^{M} C_{kj}}{\sum_{\substack{i=1 \\ i \neq k}}^{M} \frac{C_{ki} + C_{ik}}{e^{s_k^t} + e^{s_i^t}}} \right), k = 1, 2, \cdots, M \tag{6}$$

Since $C_{kk} = 0$, the above equation can be rewritten as

$$s_k^{t+1} = \log \left(\frac{\sum_{j=1}^{M} C_{kj}}{\sum_{i=1}^{M} \frac{C_{ki} + C_{ik}}{e^{s_k^t} + e^{s_i^t}}} \right), k = 1, 2, \cdots, M \tag{7}$$

After we get an estimation $\hat{\mathbf{s}}$ of $\mathbf{s}$, zero mean normalization is further performed on $\hat{\mathbf{s}}$ to obtain the final B-T scores as the ground truth subjective rating scores.

Note that the SISR results of the same scene and the same scaling factor constitute a group for PC in our study. By applying the B-T model on each group, we can get a B-T score for each SISR result in this group. Then, we rank the evaluated SISR results/algorithms based on the average B-T scores, as shown in Fig. 2. A higher B-T score indicates a better performance. We find that, ASDS [42] gets the highest B-T Score at all scaling factors. For $\times 2$ and $\times 3$, USRnet [50], which gets the second ranking, has shown significant advantage over other methods by a large margin. BCI owns the lowest B-T score on both $\times 3$ and $\times 4$ scaling factors, but performs moderately on the scaling factor $\times 2$.

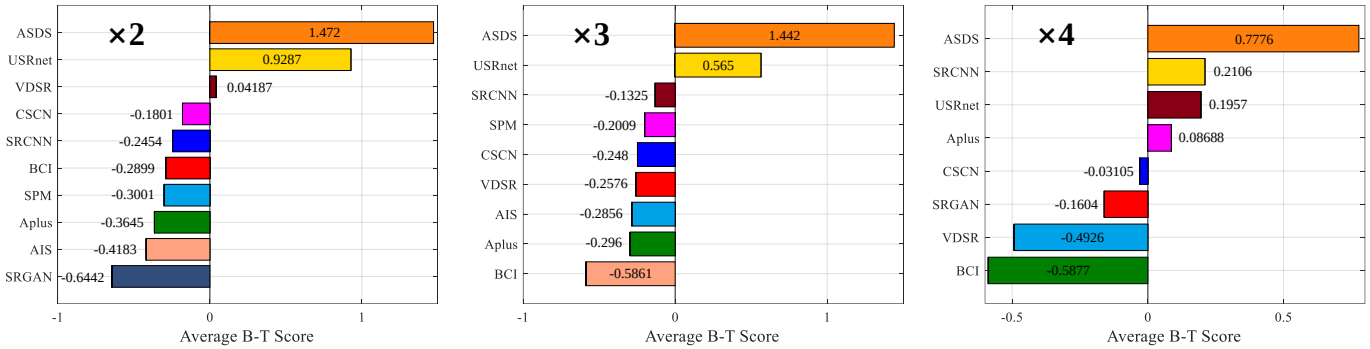

Fig. 2. Average B-T scores of different SISR algorithms at each scaling factor.

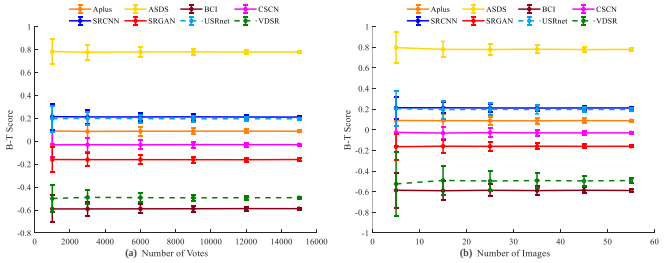

Fig. 3. Convergence analysis on the number of votes and images at scaling factor ×4.

Fig. 4. Vote percentages of different reasons.

2) Convergence Analysis: In order to demonstrate that the number of votes and images are sufficient to obtain stable subjective rating scores, we perform convergence analysis from two perspectives: the number of votes and the number of images.

Number of votes: We collect 65400 votes in total. There are 27000 votes at scaling factor ×2, 21600 votes at scaling factor ×3 and 16800 votes at scaling factor ×4. We randomly sample α ($\alpha = 1000, 3000, 6000, 9000, 12000, 15000$) votes from all votes at each scaling factor, and compute B-T score for each SISR algorithm. We repeat this process 1000 times for each α. Fig. 3(a) show the mean and standard deviation of B-T scores for each α at scaling factor ×4. Obviously, with the increasing of the number of votes, standard deviation of B-T scores decreases, indicating the subjective rating scores tend to be stable.

Number of images: We collect 60 HR images in total. We randomly sample β ($\beta = 5, 15, 25, 35, 45, 55$) HR images from our dataset and compute B-T Score for each SISR algorithm at each scaling factor. We repeat this process 1000 times for each β. Fig. 3(b) show the mean and standard deviation of B-T Scores for each β at scaling factor ×4. Obviously, with the increasing of the number of images, standard deviation of B-T Scores decreases, indicating the subjective rating scores tend to be stable.

3) Human Preference Analysis: This part analyzes the results collected in the additional questionnaire during the human subjective study. Our question is "Why don't you like another picture?" and we provide nine options for users to choose, as shown in Table V.

We show the vote percentages of all the reasons in Fig. 4. The i-th element in each row represents the vote percentage of the Reason #i for corresponding SISR algorithm in this row. In general, Reason #2 gets the highest vote percentage and Reasons #3, #5, #7 get relatively higher vote percentage than the rest reasons. As for ASDS [42], i.e., the one with the best overall performance at all scaling factors, Reason #1 gets the highest vote percentage followed by Reason #8. After observing our dataset, we found that the SISR results generated by ASDS [42] appear to have relatively clear object contours and texture details. However, these images on the other hand suffer from relatively severe noise (Reason #1) and color artifacts (Reason #8) around edges. Based on the vote percentage map, we can summarize that several influential factors on SISR image quality. First, the loss and ambiguity of details; Second, shaking, ghosting, blurring or jaggies of the edge and texture; Third, color shift, low contrast and saturation. We hope the future development of advanced SISR algorithms can better take these factors into account.

IV. KLTSRQA: An Objective NR Quality Metric for SISR

A. Motivation

As analyzed previously, we are aware that the distortions on macro-structures (e.g., edges and contours) and micro-structures (e.g., texture and details) are the main factors affecting the visual quality of SISR results. Thus, it is of critical importance to characterize the underlying distortions from different image components (i.e., macro-structures and

(a) Original Image (b) k=1 (c) k=2 (d) k=4

(e) k=8 (f) k=16 (g) k=32 (h) k=64

Fig. 5. Image reconstruction results with different numbers of spectral components. (a) is the original image, (b)-(h) are the reconstructed results by using the first $k, k = \{1, 2, 4, 8, 16, 32, 64\}$ spectral components. Zoom-in for best viewing.

micro-structures) for SISR image quality evaluation. As we will show in the next, the KLT coefficients in different spectral components actually account for different image components, i.e., the front part of spectral components in the KLT coefficient matrix accounts for macro-structures while the latter part accounts for micro-structures. Therefore, in this work we are inspired to extract quality-aware features from different spectral components to more accurately evaluate SISR image quality in a NR manner.

1) Theory of KLT: KLT is a signal dependent linear transform, the kernels of which are derived by computing the principal components along eigen-directions of the autocorrelation matrix of the input data.

Given an image $\mathbf{X}$ with size $M \times N$, a set of non-overlapping patches with size $\sqrt{K} \times \sqrt{K}$ are extracted. These image patches are vectorized and combined together to form a new matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_S] \in \mathbb{R}^{K \times S}$, where $\mathbf{x}_s \in \mathbb{R}^K, s = 1, 2, \cdots, S$ represents the s-th vectorized patch and S is the total number of image patches in $\mathbf{X}$. The covariance matrix of $\mathbf{X}$ is defined as follows

$$\mathbf{C} = \mathbb{E}[(\mathbf{x}_s - \bar{\mathbf{x}})(\mathbf{x}_s - \bar{\mathbf{x}})^{\mathrm{T}}] \tag{8}$$

$$= \frac{1}{S-1} \sum_{s=1}^{S} (\mathbf{x}_s - \bar{\mathbf{x}})(\mathbf{x}_s - \bar{\mathbf{x}})^{\mathrm{T}} \tag{9}$$

where $\bar{\mathbf{x}} = \frac{1}{S} \sum_{s=1}^{S} \mathbf{x}_s$ denotes the mean vector obtained by averaging each row of $\mathbf{X}$ and $\mathbf{C} \in \mathbb{R}^{K \times K}$. Then, the eigenvalues and eigenvectors of $\mathbf{C}$ are calculated via eigenvalue decomposition. The eigenvectors are arranged according to their corresponding eigenvalues in the descending order to form the KLT kernel $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_K] \in \mathbb{R}^{K \times K}$ where $\mathbf{p}_k \in \mathbb{R}^K, k = 1, 2, \cdots, K$ represents the k-th eigenvector. Using the KLT kernel $\mathbf{P}$, the KLT of $\mathbf{X}$ is expressed as follows:

$$\mathbf{Y} = \mathbf{P}^{\mathrm{T}} \mathbf{X} \tag{10}$$

where $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_K]^{\mathrm{T}} \in \mathbb{R}^{K \times S}$ is the KLT coefficient matrix and $\mathbf{y}_k \in \mathbb{R}^S, k = 1, 2, \cdots, K$ refers to the k-th spectral component obtained by $\mathbf{y}_k = (\mathbf{p}_k)^{\mathrm{T}} \mathbf{X}$.

2) Relationship Between KLT and SISR Image Quality Evaluation: Ideally, we can reconstruct the original image based on the KLT coefficient matrix $\mathbf{Y}$ and the KLT kernel $\mathbf{P}$. Note that $\mathbf{P}$ is an orthogonal matrix, thus

$$\mathbf{P}\mathbf{P}^{\mathrm{T}} = \mathbf{I} \tag{11}$$

where $\mathbf{I} \in \mathbb{R}^{K \times K}$ represents the identity matrix with size $K \times K$. So, the original image $\mathbf{X}$ can be reconstructed as follows

$$\mathbf{X} = \mathbf{P}\mathbf{Y} \tag{12}$$

In order to understand the role of different spectral components in image reconstruction, we take the first k spectral components in KLT coefficient matrix as the reconstruction KLT coefficient matrix $\mathbf{Y}^{(k)}$, which is defined as follows:

$$\mathbf{Y}^{(k)} = \begin{bmatrix} y_{1,1} & \cdots & y_{k,1} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y_{1,S} & \cdots & y_{k,S} & 0 & \cdots & 0 \end{bmatrix}^{\mathrm{T}} \tag{13}$$

where $\mathbf{Y}^{(k)} \in \mathbb{R}^{K \times S}, k = 1, 2, \cdots, K$. By setting different values of k, different numbers of spectral components are involved in the reconstruction process. The image is reconstructed as follows:

$$\mathbf{X}^{(k)} = \mathbf{P}\mathbf{Y}^{(k)} \tag{14}$$

and $\mathbf{X}^{(k)}$ denotes the reconstructed image by only considering the first k spectral components.

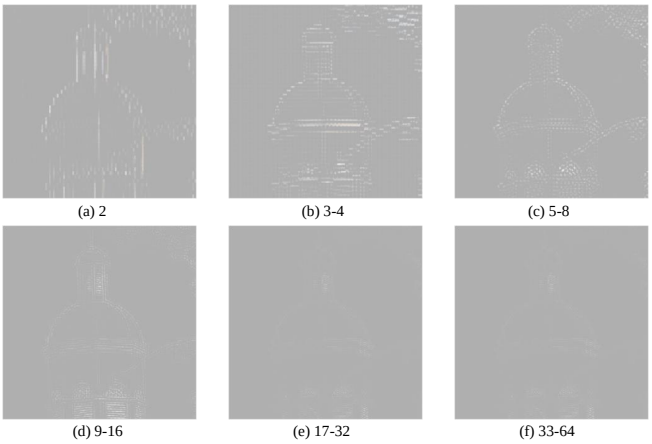

Fig. 6. Difference maps between the adjacent reconstruction results shown in Fig. 5. These difference maps are exactly the results reconstructed by the spectral components whose indexes are shown below each image.

A visual example is shown in Fig. 5 where we set $K = 64$ and $k = 1, 2, 4, 8, 16, 32, 64$. Fig. 5(a) is the original image, Fig. 5(b)-Fig. 5(h) are the reconstructed images corresponding to $k = 1, 2, 4, 8, 16, 32, 64$, respectively. As shown in Fig. 5(b), when only the 1^{st} spectral component is involved in the reconstruction process, almost all the macro-structures are recovered. As k increases, the small textures and details become richer and clearer. We then present more results to demonstrate the role of the last k spectral components in image reconstruction. Fig. 6 shows the difference maps between the adjacent reconstruction results shown in Fig. 5. Note that, the image shown in Fig. 6(a) refers to the difference map between Fig. 5(b) and Fig. 5(c) and is exactly the reconstruction result by only using the 2^{nd} spectral component. The image shown in Fig. 6(b) refers to the difference map between Fig. 5(c) and Fig. 5(d) and is exactly the reconstruction result by only using the 3^{rd} and 4^{th} spectral components, and so on.

From these results, we can have the following observations. First, the first spectral component (see Fig. 5(b)) can reconstruct most image structures. Second, Fig. 6(a)(b)(c) contain some basic contours and edges while Fig. 6(d)(e)(f) only contain some extremely small details. In other words, we can say that the front part of spectral components in the KLT coefficient matrix accounts for the reconstruction of image macro-structures such as the basic contour and main structures while the latter part of spectral components accounts for the reconstruction of image micro-structures such as the small textures and details. As we have reported that the distortions on macro-structures (e.g., edges and contours) and micro-structures (e.g., texture and details) are also the main factors affecting the visual quality of SISR results. Thus, by specifying different spectral components, KLT provides an effective way to characterize the underlying SISR image distortions from different image components (i.e., macro-structures and micro-structures).

B. KLTSRQA

1) Overview: The flow chart of *KLTSRQA* is depicted in Fig. 7. The input of *KLTSRQA* is a to-be-evaluated SISR image and the output is an estimated quality score. For an input SISR image in the RGB format, we first convert it into the opponent color space [52, 53] and then perform a local mean subtraction and divisive normalization on each color channel to obtain three mean subtracted contrast normalized (MSCN) coefficient maps. The KLT is performed on the corresponding MSCN maps using the KLT kernels that we have constructed offline. Therefore, we can obtain three KLT coefficient matrices corresponding to the three opponent color channels. Based on the obtained KLT coefficient matrix for each channel, quality-aware feature extraction is performed from two aspects. On the first aspect, we use the asymmetric generalized Gaussian distribution (AGGD) model [54] to fit the KLT coefficients in different spectral components and the estimated AGGD parameters are taken as the first part of features. On the second aspect, we compute the energy of the KLT coefficients in different spectral components as the second part of features. These two parts of features are combined together and aggregated over three channels to yield a final quality-aware feature vector for quality score prediction via the SVMrank model.

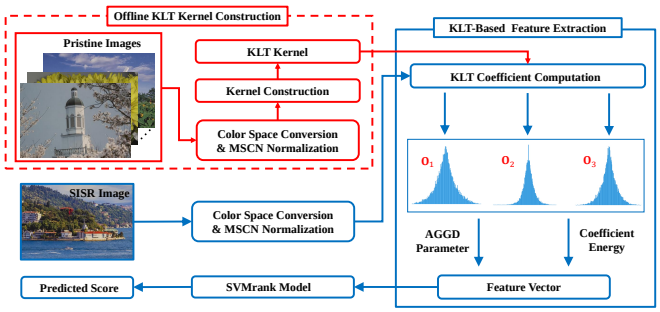

Fig. 7. The flow chart of our proposed KLTSRQA metric.

2) Color Space Conversion & MSCN Normalization: Instead of the original RGB color space, our method is implemented in a more perceptually relevant opponent color space which has been demonstrated to be better correlated with the color perception of HVS. The color space conversion is formulated as follows:

$$\begin{bmatrix} \mathbf{O}_1 \\ \mathbf{O}_2 \\ \mathbf{O}_3 \end{bmatrix} = \begin{bmatrix} 0.06 & 0.63 & 0.27 \\ 0.30 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{bmatrix} \begin{bmatrix} \mathbf{R} \\ \mathbf{G} \\ \mathbf{B} \end{bmatrix} \tag{15}$$

For each color channel $\mathbf{O}_l, l = 1, 2, 3$, we perform a local mean subtraction and divisive normalization by computing the corresponding MSCN coefficients [31]:

$$\widehat{\mathbf{O}}_l(i,j) = \frac{\mathbf{O}_l(i,j) - \mu(i,j)}{\sigma(i,j) + c} \tag{16}$$

where i and j are the spatial coordinates of a pixel, $c = 1$ is a constant that prevents instabilities from occurring when the denominator tends to zero and we set $c = 1$ here.

$$\mu(i,j) = \sum_{p=-P}^{P} \sum_{q=-Q}^{Q} w^{p,q} \mathbf{O}_l^{p,q}(i,j) \tag{17}$$

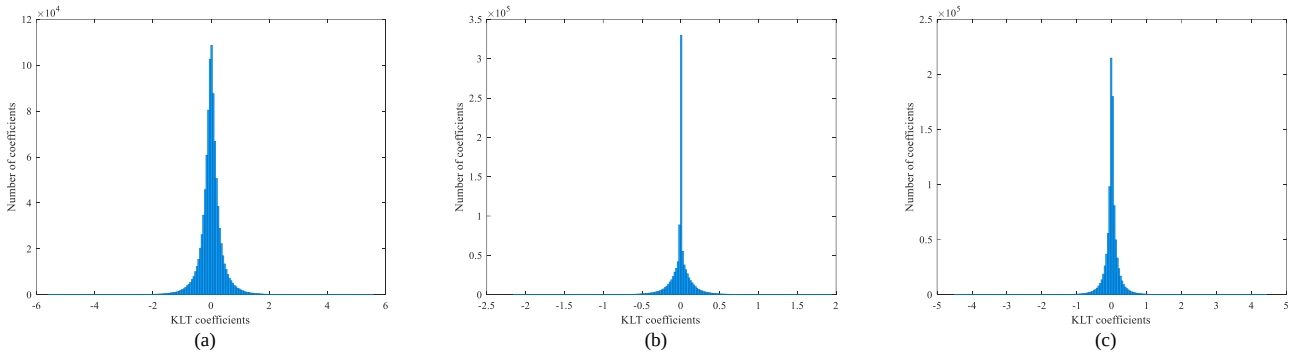

Fig. 8. Distributions of the KLT coefficients in three channels. (a) $\mathbf{O}_1$; (b) $\mathbf{O}_2$; (c) $\mathbf{O}_3$.

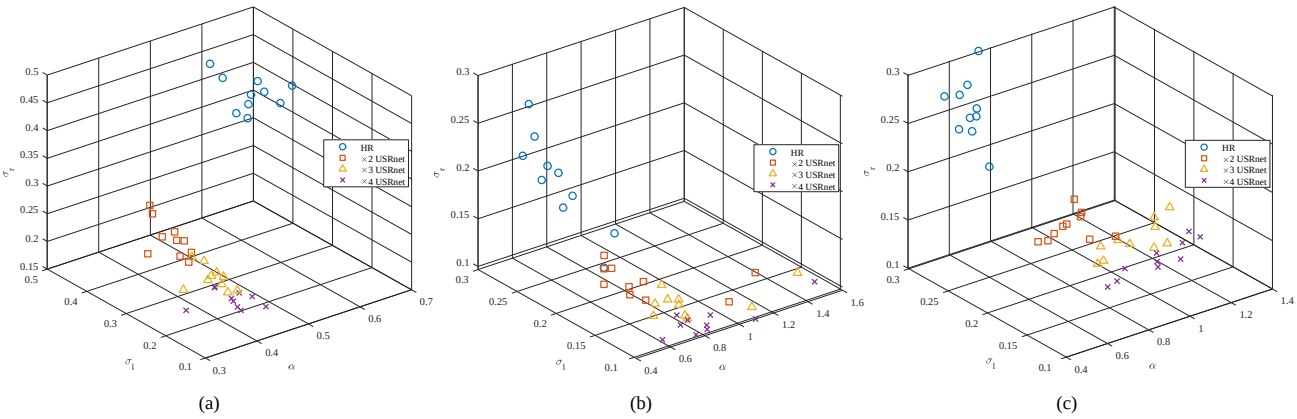

Fig. 9. Visualization of the estimated AGGD parameters of different images. (a) $\mathbf{O}_1$; (b) $\mathbf{O}_2$; (c) $\mathbf{O}_3$.

$$\sigma(i,j) = \sqrt{\sum_{p=-P}^{P}\sum_{q=-Q}^{Q} w^{p,q}(\mathbf{O}_l^{p,q}(i,j) - \mu(i,j))^2} \tag{18}$$

where $w^{p,q}$ is a 2D circularly-symmetric Gaussian weighting function sampled out to 3 standard deviations and re-scaled to unit volume. According to the implementations in [31], we also set $P = Q = 3$.

3) Offline KLT Kernel Construction: In this work, we resort to the 60 pristine HR images in *RealSRQ* to construct three KLT kernels with size 64×64 for the three color channels, respectively. The procedures of the offline KLT kernel construction are shown in the red lines in Fig. 7. Note that each pristine HR image is also preprocessed as described in *Color Space Conversion & MSCN Normalization*. For each color channel, the MSCN maps of all the HR images are used together to obtain the corresponding KLT kernel according to the procedures described in *Theory of KLT*. Finally, we can get three KLT kernels $\mathbf{P}_1$, $\mathbf{P}_2$, and $\mathbf{P}_3$ for the three channels $\mathbf{O}_1$, $\mathbf{O}_2$, and $\mathbf{O}_3$, respectively, which will be used for KLT of the input SISR images.

4) KLT-Based Feature Extraction: For an input to-be-evaluated SISR image, the KLT is performed on the corresponding MSCN maps $\widehat{\mathbf{O}}_l, l = 1,2,3$ using the KLT kernels $\mathbf{P}_l, l = 1,2,3$ that we have constructed offline:

$$\mathbf{Y}_l = \mathbf{P}_l{}^{\mathrm{T}}\widehat{\mathbf{O}}_l \tag{19}$$

where $\mathbf{Y}_l, l = 1,2,3$ represents the obtained KLT coefficient matrix for the three color channels. Based on $\mathbf{Y}_l, l = 1,2,3$, quality-aware feature extraction is performed from two aspects which will be detailed in the next.

AGGD Parameters: Since the quality-aware features are extracted based on the KLT coefficient matrices $\mathbf{Y}_l, l = 1,2,3$, we first conduct some statistical analyses of them. For a certain HR image in *RealSRQ*, the histograms of $\mathbf{Y}_l, l = 1,2,3$ are shown in Fig. 8. It is observed that they all present Gaussian-like distributions. Considering the distributions, instead of being completely symmetric, are somewhat asymmetric between the left and right sides, it is appropriate to use the asymmetric generalized Gaussian distribution (AGGD) [55] to fit the KLT coefficients. The probability density function of AGGD with zero mode is expressed as follows:

$$f(x;\alpha,\sigma_l,\sigma_r) = \begin{cases} \frac{\alpha}{(\beta_l+\beta_r)\gamma(\frac{1}{\alpha})}\exp\left(-\left(\frac{-x}{\beta_l}\right)^{\alpha}\right), & x < 0 \\ \frac{\alpha}{(\beta_l+\beta_r)\gamma(\frac{1}{\alpha})}\exp\left(-\left(\frac{-x}{\beta_r}\right)^{\alpha}\right), & x \geq 0 \end{cases} \tag{20}$$

where β_l and β_r are defined by the following equations.

$$\beta_l = \sigma_l\sqrt{\frac{\gamma(\frac{1}{\alpha})}{\gamma(\frac{3}{\alpha})}},\ \beta_r = \sigma_r\sqrt{\frac{\gamma(\frac{1}{\alpha})}{\gamma(\frac{3}{\alpha})}} \tag{21}$$

where $\gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}dt$ is the gamma function; $\alpha > 0$ is the shape parameter and $\sigma_l > 0, \sigma_r > 0$ are left-scale and right-scale parameter that control the spread on the left and

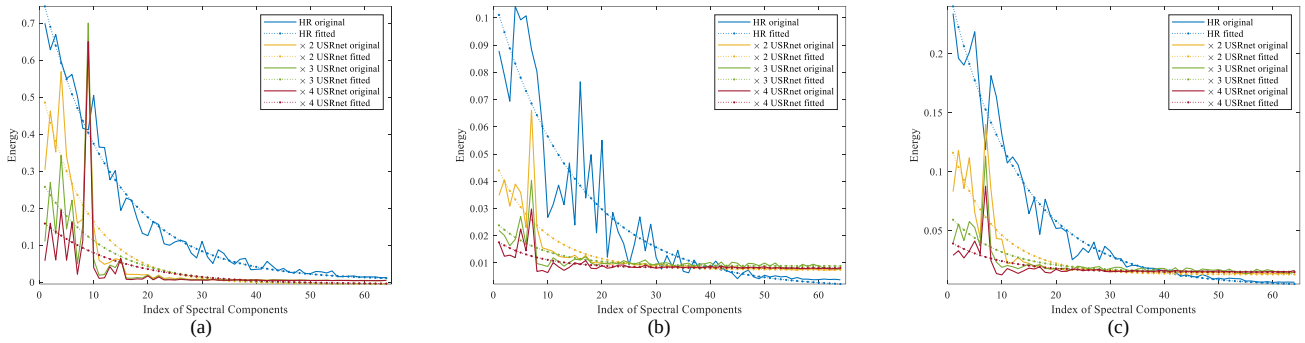

Fig. 10. Energy of KLT coefficients in different spectral components. (a) $\mathbf{O}_1$; (b) $\mathbf{O}_2$; (c) $\mathbf{O}_3$.

right side of the mode, respectively. When $\sigma_l = \sigma_r > 0$, AGGD reduces to GGD. The parameters of AGGD $(\alpha, \sigma_l, \sigma_r)$ can be obtained by the moment-matching-based approach proposed in [54].

In our method, both the KLT coefficient distribution in the whole matrix and the KLT coefficient distributions in each individual spectral component are fitted with AGGD and all the estimated AGGD parameters are combined together to form as the first part of our extracted quality-aware features.

Next, we will illustrate the effectiveness of the estimated AGGD parameters $(\alpha, \sigma_l, \sigma_r)$ in distinguishing the visual quality of SISR images. For this purpose, we visualize the distributions of the estimated AGGD parameters of different images in three-dimensional space. The distributions of the AGGD parameters estimated on the 10 HR images annotated as *Building* in *RealSRQ* ("*Building* HR") and their corresponding SISR results generated by the USRnet [50] algorithm at three scaling factors ("*Building* $\times 2$ USRnet", "*Building* $\times 3$ USRnet", and "*Building* $\times 4$ USRnet") are visualized in Fig. 9. Note that (a), (b), (c) correspond to three channels, respectively. Typically, for the same image content and the same SISR algorithm, the visual quality of SISR images will decrease as the scaling factor increases. It is clear that the distributions of AGGD parameters shown in Fig. 9 can well characterize such trend, i.e., the AGGD parameters of the intra-group images are concentrated together while the AGGD parameters of the inter-group images are well separated in the three-dimensional space. In addition, we also observe some slight differences among the three color channels. Specifically, in channel $\mathbf{O}_1$, $\alpha, \sigma_l, \sigma_r$ of "*Building* HR" are relatively large. As scaling factor increases (i.e., visual quality decreases), $\alpha, \sigma_l, \sigma_r$ show a decreasing trend. In channel $\mathbf{O}_2$, σ_l, σ_r of "*Building* HR" are relatively large. As scaling factor increases (i.e., visual quality decreases), σ_l, σ_r show a decreasing trend while α stays stable. In channel $\mathbf{O}_3$, σ_l, σ_r of "*Building* HR" are relatively large and α is relatively small. As scaling factor increases (i.e., visual quality decreases), σ_l, σ_r show a decreasing trend and α shows an increasing trend. Overall, the AGGD parameters estimated on the KLT coefficient matrices in three opponent color channels all have good capability in distinguishing the visual quality of SISR images.

Energy of KLT Coefficients: In addition to use AGGD to model the KLT coefficient distributions, we also calculate the energy of KLT coefficients. The KLT coefficient matrix on each color channel $\mathbf{O}_l$ of a SISR image is $\mathbf{Y}_l, l = 1, 2, 3$. Thus, the energy of KLT coefficients in each spectral component is defined as follows:

$$e_{l,k} = \frac{1}{S}\sum_{s=1}^{S} \mathbf{Y}_l(k,s)^2, \ (k = 1, 2, \cdots, K) \tag{22}$$

For an HR image *Building 001 HR* in *RealSRQ* and its corresponding SISR results generated by the USRnet [50] algorithm at three scaling factors (i.e., *Building 001* $\times 2$ *USRnet*, *Building 001* $\times 3$ *USRnet*, and *Building 001* $\times 4$ *USRnet*), their energy distributions in different spectral components are shown in Fig. 10. Again, (a), (b), (c) correspond to three opponent color channels, respectively. From these figures, we can find that for all the images thee energies generally decrease as the spectral component index increases, but the attenuation factors differ for different images. Specifically, the attenuation factor is proportional to visual quality, i.e., the attenuation factor is the largest for *Building 001 HR* and the lowest for *Building 001* $\times 4$ *USRnet*. Inspired by this, we use the exponential function to fit the energy distribution. The exponential function is defined as follows:

$$f(x; \lambda_1, \lambda_2, \lambda_3) = \lambda_1 e^{\lambda_2 x} + \lambda_3, \tag{23}$$

where λ_1, λ_2, and λ_3 are the parameters to be fitted. The fitted curves by exponential function are also shown in the dashed lines in Fig. 10. As we can see, the fitted curves can well characterize the general trend of energy changes along with the increase of spectral component index. However, instead of directly using the three parameters as the features, we resort to sample the continuous curves that we have fitted at an interval of 4 to better enhance the feature representation capacity. Finally, the sampled discrete values on the fitted curve are taken as the second part of our extracted features.

C. Quality Evaluation

The remaining issue is to map the extracted quality-aware features to predicted quality scores. This is a typical regression problem from a machine learning perspective and we can resort to any regression algorithm to implement it. Since the ground truth B-T scores are only meaningful within the same group, we resort to the classical SVMrank model [56] to learn the mapping function from extracted features to subjective quality scores, i.e., B-T scores in *RealSRQ*.

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT NR-IQA METRICS ON REALSRQ.

Metric	KROCC				SROCC				PLCC				RMSE			
	×2	×3	×4	All	×2	×3	×4	All	×2	×3	×4	All	×2	×3	×4	All
GM-LOG [28]	0.3944	0.5189	0.4579	0.4547	0.5055	0.6460	0.5777	0.5737	0.9328	0.8972	0.7853	0.8772	0.2982	0.2931	0.3149	0.3014
BLIINDS-II [29]	0.3492	0.5502	0.4829	0.4558	0.4436	0.6660	0.5961	0.5629	0.9312	0.9182	**0.8451**	0.9014	0.2989	0.2627	**0.2465**	0.2713
CurveletQA [30]	0.2957	0.3536	0.4405	0.3579	0.3855	0.4623	0.5589	0.4625	0.9410	0.8848	0.7988	0.8801	0.2898	0.3094	0.3018	0.2999
BRISQUE [31]	0.3520	0.4203	0.2995	0.3592	0.4476	0.5348	0.3748	0.4551	**0.9495**	0.9212	0.7289	0.8747	**0.2496**	0.2474	0.3861	0.2893
OG-IQA [32]	0.3327	0.4150	0.3201	0.3564	0.4420	0.5252	0.4087	0.4599	0.9258	0.8403	0.7774	0.8533	0.3492	0.3544	0.2875	0.3327
SSEQ [33]	0.3086	0.4852	0.3021	0.3655	0.4085	0.6002	0.3998	0.4698	0.9326	0.8672	0.6458	0.8258	0.3032	0.3175	0.3888	0.3333
DIIVINE [34]	0.3148	0.4750	0.4812	0.4175	0.4128	0.6111	0.5995	0.5342	0.9326	0.9184	0.8086	0.8911	0.2744	0.2573	0.2913	0.2737
RISE [35]	0.4142	0.3913	0.3518	0.3881	0.5201	0.5055	0.4559	0.4962	0.9313	0.8741	0.7262	0.8515	0.2784	0.3379	0.3076	0.3069
BMPRI [36]	0.3503	0.2795	0.1544	0.2687	0.4535	0.3956	0.1861	0.3550	0.9368	0.7651	0.6373	0.7908	0.2663	0.4065	0.3710	0.3441
FRIQUEE [37]	**0.5715**	**0.6210**	**0.5978**	**0.5958**	**0.6958**	**0.7378**	**0.7081**	**0.7134**	0.9093	**0.9278**	**0.8692**	**0.9036**	0.3002	**0.2362**	**0.2493**	**0.2638**
NIQE [38]	0.0419	0.2367	-0.2225	0.0285	0.0465	0.3006	-0.2906	0.0313	0.6197	0.7923	0.6390	0.6830	0.7707	0.4091	0.3623	0.5292
ILNIQE [39]	-0.1300	0.0745	0.1351	0.0167	-0.1410	0.0992	0.1743	0.0325	0.7861	0.7976	0.6167	0.7397	0.5484	0.3807	0.4200	0.4545
HVS-MaxPol [40]	**0.4222**	0.3498	0.3273	0.3699	**0.5394**	0.4854	0.4204	0.4861	**0.9504**	0.8436	0.6747	0.8331	**0.2515**	0.3399	0.3625	0.3139
PCRL [41]	0.3563	0.4923	0.5673	0.4642	0.4520	0.6220	0.6719	0.5738	0.9407	0.9186	0.8066	0.8936	0.2584	0.2507	0.3006	0.2683
SR-metric [14]	0.3723	**0.5723**	**0.5714**	**0.4980**	0.4772	**0.6921**	**0.6829**	**0.6098**	0.9415	**0.9276**	0.8442	**0.9080**	0.2758	**0.2430**	0.2687	**0.2628**
KLTSRQA	**0.6125**	**0.6296**	**0.6416**	**0.6268**	**0.7282**	**0.7522**	**0.7486**	**0.7422**	**0.9604**	**0.9285**	**0.8753**	**0.9246**	**0.2288**	**0.2319**	**0.2159**	**0.2260**

V. EXPERIMENTAL RESULTS

A. Algorithm Performance Test

First, we test the performance of *KLTSRQA* on *RealSRQ*. For performance evaluation, all the SISR images in *RealSRQ* are randomly divided into two subsets, 80% are used for training the SVMrank model, and the remaining 20% are used as the testing samples. Note that these two subsets do not have any content overlapping to ensure there is no performance bias towards specific image contents. The dataset random partition process is repeated for 1,000 times. For each time, we calculate the Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-Order Correlation Coefficient (SROCC), Kendall Rank-Order Correlation Coefficient (KROCC), and Root Mean Square Error (RMSE) between the predicted scores and ground truth subjective B-T scores. The median values over 1000 times are calculated to measure the consistency between objective evaluation and subjective rating results. An ideal match between objective and subjective scores will have KROCC=SROCC=PLCC=1 and RMSE=0.

To demonstrate the superiority of our method, 15 existing NR-IQA metrics, including two most recent dedicated metrics for image restoration [41] and super-resolution [14], are implemented for comparison. These NR-IQA metrics include GM-LOG [28], BLIINDS-II [29], CurveletQA [30], BRISQUE [31], OG-IQA [32], SSEQ [33], DIIVINE [34], RISE [35], BMPRI [36], FRIQUEE [37], NIQE [38], ILNIQE [39], HVS-MaxPol [40], PCRL [41], and SR-metric [14]. Among these methods, NIQE [38], ILNIQE [39], and HVS-MaxPol [40] are training-free while the rests are all training-based. For all the training-based methods, the regression functions are all implemented by SVMrank for fair comparison. The performance results are shown in Table VI. As shown, *KLTSRQA* achieves the best performance in terms of all performance criteria at all three scaling factors. FRIQUEE [37] and SR-metric [14] also have relatively good performance than the others.

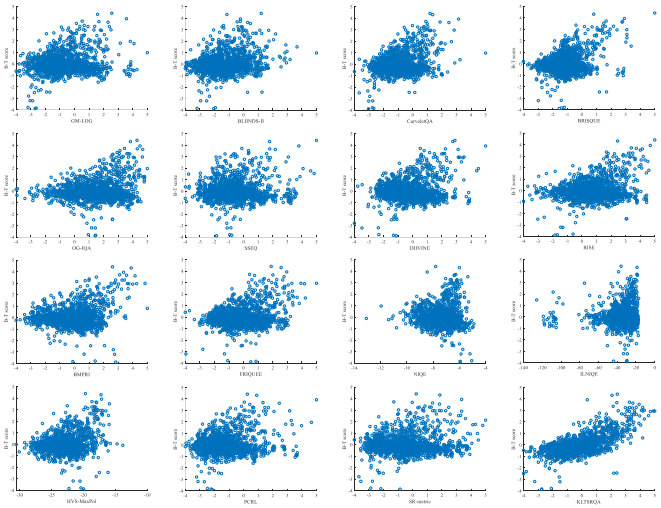

Fig. 11. Scatter plots of different NR-IQA metrics on RealSRQ.

Note that the SR-metric [14] is specifically designed for SR images, yet its performance is still worse than the *KLTSRQA*, implying the effectiveness of our KLT-based feature extraction methods. In addition, we draw the scatter plots between objective scores and subjective scores for better visualization of the performances of different NR-IQA metrics, as shown in Fig. 11. We can observe that the proposed *KLTSRQA* metric is more in line with subjective B-T scores. To further demonstrate the superiority of *KLTSRQA*, we also conduct statistical significance test. Specifically, the two sample t-test between the pair of SROCC values of 1,000 train-test loops at the 5% significance level is conducted. Fig. 12 presents the t-test results, where the value 1/-1 indicates that row algorithms perform statistically better/worse than the column algorithms while the value 0 indicates that row algorithms perform statistically competitive with the column algorithms.

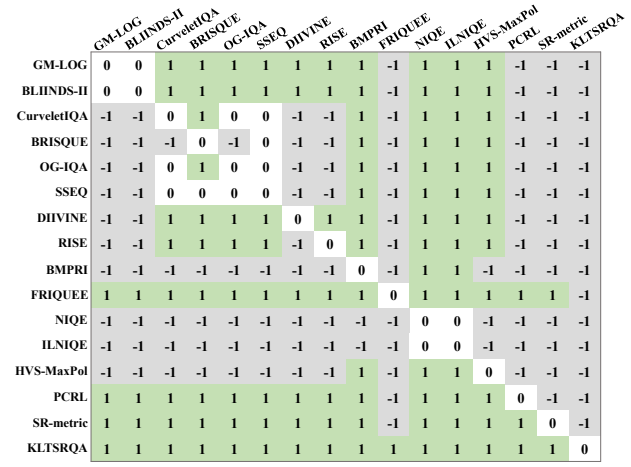

Fig. 12. T-test results of different NR-IQA metrics on RealSRQ.

TABLE VII
PERFORMANCE COMPARISON WITH DIFFERENT KERNEL SIZES.

Kernel Size	KROCC				SROCC			
	×2	×3	×4	All	×2	×3	×4	All
4	0.5260	0.5492	0.3695	0.4874	0.6446	0.6741	0.4686	0.6023
16	0.5406	0.5708	0.4801	0.5327	0.6580	0.6969	0.5944	0.6521
64	**0.6125**	**0.6296**	**0.6416**	**0.6268**	**0.7282**	**0.7522**	**0.7486**	**0.7422**
256	0.5491	0.5230	0.5618	0.5442	0.6749	0.6515	0.6831	0.6695
4+16	0.5358	0.5866	0.5025	0.5429	0.6536	0.7056	0.6175	0.6602
4+16+64	**0.6171**	**0.6357**	**0.6418**	**0.6306**	**0.7265**	**0.7566**	**0.7524**	**0.7442**
4+16+64+256	0.5497	0.5121	0.5585	0.5398	0.6667	0.6491	0.6735	0.6628

As we can see, *KLTSRQA* performs statistically better than all the competing NR-IQA metrics on *RealSRQ*, which further prove the superiority of *KLTSRQA*.

B. Determination of KLT Kernel Size

In order to determine the most appropriate KLT kernel size, we carry out experiments to manually select the KLT kernel size. In addition to the single size of KLT kernel, we also consider the combination of different sizes of KLT kernels. The experimental results are shown in Table VII. The two kernel sizes leading to the top two performance are shown in bold. As shown in Table VII, for single KLT kernel, size 64×64 achieves the best performance. For multiple KLT kernels, the combination of KLT kernels with size $4 \times 4, 16 \times 16, 64 \times 64$ achieves the best performance. However, compared to the single KLT kernel with size 64×64, the performance is only slightly improved but the complexity of the model greatly increased. Considering both algorithm performance and model simplicity, the single KLT kernel with size 64×64 is selected as the final one.

C. Ablation Study

Ablation study aims to test the contribution of each component in our model. Our ablation study includes four parts: (1) performance test of individual part of features, (2) performance test of individual channel in the opponent color space, (3) validity of using AGGD for parameter estimation, and (4) validity of using opponent color space.

TABLE VIII
PERFORMANCE TEST OF INDIVIDUAL PART OF FEATURES.

Feature	KROCC				SROCC			
	×2	×3	×4	All	×2	×3	×4	All
KLTSRQA-Energy	0.5305	0.5441	0.5557	0.5425	0.6498	0.6716	0.6664	0.6620
KLTSRQA-AGGD	0.6037	0.6136	0.6383	0.6173	0.7195	0.7342	0.7464	0.7324
KLTSRQA	**0.6125**	**0.6296**	**0.6416**	**0.6268**	**0.7282**	**0.7522**	**0.7486**	**0.7422**

TABLE IX
PERFORMANCE TEST OF EACH INDIVIDUAL CHANNEL.

Channel	KROCC				SROCC			
	×2	×3	×4	All	×2	×3	×4	All
KLTSRQA $-$ $\mathbf{O}_1$	0.5495	0.5991	0.5521	0.5668	0.6632	0.7224	0.6677	0.6843
KLTSRQA $-$ $\mathbf{O}_2$	0.5315	0.5806	0.5720	0.5599	0.6454	0.7039	0.6821	0.6758
KLTSRQA $-$ $\mathbf{O}_3$	0.5489	0.5741	0.6108	0.5756	0.6663	0.7024	0.7201	0.6943
KLTSRQA	**0.6125**	**0.6296**	**0.6416**	**0.6268**	**0.7282**	**0.7522**	**0.7486**	**0.7422**

TABLE X
PERFORMANCE RESULTS OF DIFFERENT COLOR SPACES. OC: OPPONENT COLOR SPACE.

Color Space	KROCC				SROCC			
	×2	×3	×4	All	×2	×3	×4	All
RGB	0.5621	**0.6333**	0.5787	0.5908	0.6788	0.7497	0.6951	0.7073
HSV	0.5750	0.6050	0.5549	0.5790	0.6835	0.7286	0.6684	0.6941
YCbCr	0.5906	0.6222	0.5988	0.6035	0.7059	0.7448	0.7100	0.7201
OC	**0.6125**	0.6296	**0.6416**	**0.6268**	**0.7282**	**0.7522**	**0.7486**	**0.7422**

TABLE XI
PERFORMANCE TEST OF INDIVIDUAL PART OF FEATURES.

Fitting Model	KROCC				SROCC			
	×2	×3	×4	All	×2	×3	×4	All
GGD	0.5903	0.5929	0.5764	0.5870	0.6989	0.7229	0.6872	0.7034
AGGD	**0.6125**	**0.6296**	**0.6416**	**0.6268**	**0.7282**	**0.7522**	**0.7486**	**0.7422**

Performance Test of Individual Part of Features: The extracted features in our model contains two types: (1) AGGD parameters of KLT coefficients and (2) KLT coefficient energy. We will analyze the contributions of these two types of features. The experimental results are shown in Table VIII where *KLTSRQA*-Energy represents KLT coefficients energy and *KLTSRQA*-AGGD represents the estimated AGGD parameters of KLT coefficients. As shown in Table VIII, both KROCC and SROCC of *KLTSRQA*-AGGD are higher than *KLTSRQA*-Energy at three scaling factors. Therefore, the contribution of the estimated AGGD parameters of KLT coefficients in our model plays a more important role than the KLT coefficients energy features. However, a combination of these two types of features can successfully lead to the best performance at all three scaling factors.

Performance Test of Individual Channel: Our method is implemented in a perceptual quality relevant opponent color space including three channels, i.e., $\mathbf{O}_1$, $\mathbf{O}_2$, and $\mathbf{O}_3$. Now, we analyze the contribution of each individual channel. The experimental results are shown in Table IX. As shown in Table IX, the features extracted from $\mathbf{O}_1$ and $\mathbf{O}_2$ own the similar performance. At scaling factors $\times 2$ and $\times 3$, the features extracted from $\mathbf{O}_3$ own the similar performance with $\mathbf{O}_1$ and $\mathbf{O}_2$. At scaling factor $\times 4$, the features extracted from $\mathbf{O}_3$ achieve better performance than $\mathbf{O}_1$ and $\mathbf{O}_2$. In general, the

TABLE XII
PERFORMANCE COMPARISON ON QADS.

Metric	KROCC	SROCC	PLCC	RMSE
GM-LOG [28]	0.6479	0.8208	0.8433	0.1467
BLIINDS-II [29]	0.7095	0.8711	0.8884	0.1256
CurveletQA [30]	0.6994	0.8685	0.8744	0.1329
BRISQUE [31]	0.7957	0.9373	0.9427	0.0920
ILNIQE [39]	0.6794	0.8644	0.8628	0.1386
NIQE [38]	0.3393	0.4902	0.5581	0.2277
OG-IQA [32]	0.6905	0.8678	0.8848	0.1278
SSEQ [33]	0.6998	0.8645	0.8786	0.1315
DIIVINE [34]	0.7366	0.8903	0.9180	0.1100
RISE [35]	0.7066	0.8744	0.8868	0.1265
BMPRI [36]	0.5212	0.6865	0.7238	0.1877
FRIQUEE [37]	0.8021	0.9347	0.9425	0.0914
HVS-MaxPol [40]	0.6233	0.7914	0.8060	0.1630
PCRL [41]	0.7610	0.9059	0.9355	0.0971
SR-metric [14]	0.7567	0.9068	0.8973	0.1206
KLTSRQA	**0.8312**	**0.9564**	**0.9514**	**0.0846**

contributions of these three channels are complementary to each other for different scaling factors and the best performance is achieved by considering all these three color channels simultaneously.

Validity of Using Opponent Color Space: Our proposed *KLTSRQA* metric is implemented in the opponent color space. Compared with those normal color spaces such as RGB, HSV, and YCbCr, the opponent color space is more perceptually relevant. Here, we conduct experiments to demonstrate the effectiveness of the opponent color space in our proposed *KLTSRQA* metric. The experimental results are shown in Table X. We can find that using the opponent color space can lead to the best performance in most cases and the best overall performance.

Validity of Using AGGD for Parameter Estimation: We finally compare the performance by using AGGD or GGD for parameter estimation. Specifically, we have conducted experiments by replacing AGGD with GGD for parameter fitting, and adopted SVMrank to learn the quality regression model. The results are listed in Table XI. It can be seen that using AGGD can obtain better performance than GGD in terms of all performance criteria. Since in theory AGGD is a generalized version of GGD, its superior performance is expectable.

D. Validation on Synthetic Dataset

In addition to *RealSRQ*, we also conduct performance test on another SISR dataset *QADS* [16] to more comprehensively validate the performance of *KLTSRQA*. As described in [16], QADS is a recently published synthetic SISR image quality dataset where the LR images are generated by simulating a simple and uniform degradation on their HR versions. Also, the same 15 NR-IQA metrics are included for comparison on *QADS* [16]. The numerical performance results are shown in Table XII and the scatter plots are shown in Fig. 13. As shown, *KLTSRQA* still achieves the best numerical performance results (i.e., the highest PLCC, KROCC, and SROCC values, while the lowest RMSE value) on *QADS* [16]. As observed from Fig. 13, the scatter plot of *KLTSRQA* is also highly in line with the subjective scores, which further demonstrates its superiority and good robustness on different datasets.

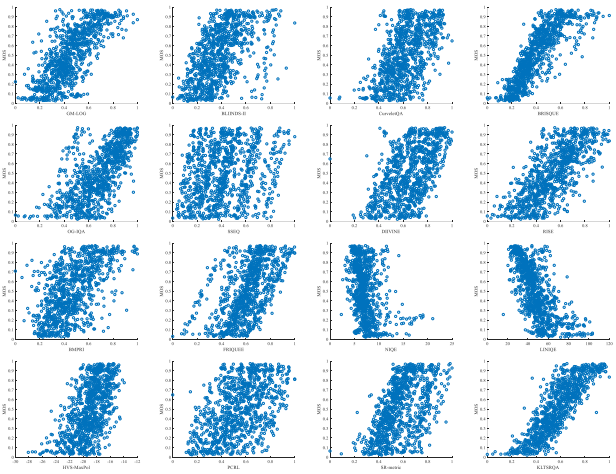

Fig. 13. Scatter plots of different NR-IQA metrics on QADS.

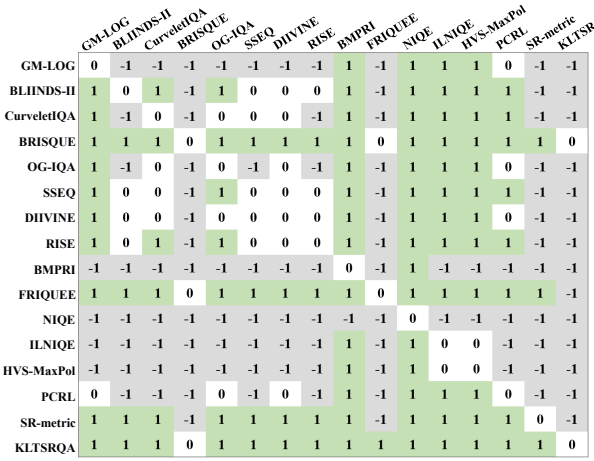

Fig. 14. T-test results of different NR-IQA metrics on QADS.

In Fig. 14, we present the two sample t-test results on *QADS*. It is seen that *KLTSRQA* performs statistically better than almost all the competing NR-IQA metrics except the BRISQUE metric [31], which is actually equivalent with *KLTSRQA*. It means that even with the traditional BRISQUE metric [31], a highly accurate quality assessment of synthetic SISR images can be achieved. However, we notice that the SROCC and KROCC values on *QADS* are much higher than those on *RealSRQ*, implying that it is generally more difficult to accurately evaluate the visual quality of SISR images in the real-world case and also demonstrating the necessity of more research efforts on real-world SISR image quality evaluation in the near future.

E. Running Time Comparison

Besides the high prediction accuracy, an excellent NR-IQA metric should also be computationally efficient. We test the running time of different NR-IQA metrics with the same setting and platform. The testing image is a 1200×800 color image. The experiments are all conducted on a PC with an

TABLE XIII
RUNNING TIME COMPARISON (IN SECOND) OF DIFFERENT NR-IQA METRICS FOR PROCESSING A 1200×800 COLOR IMAGE.

GM-LOG	BLIINDS-II	CurveletQA	BRISQUE	ILNIQE	NIQE	OG-IQA	SSEQ	DIIVINE	RISE	BMPRI	FRIQUEE	HVS-MaxPol	PCRL	SR-metric	*KLTSRQA*
0.1114	112.3120	4.2238	0.1310	45.7343	0.3351	0.1107	1.5638	28.6201	2.1553	5.0127	38.1595	0.3006	1.6948	57.0650	0.4908

AMD Ryzen 7 4800H@2.9GHZ CPU and 16GB RAM. The software platform is MATLAB R2018a. The running time of different NR-IQA metrics can be found in Table XIII. It is observed that the proposed KLTSRQA is highly efficient, i.e., it only requires less than 0.5 second to process a 1200×800 color image.

F. Discussions

One important point should be noted is that KLT is similar with pyramid decomposition and wavelet transform in function. Therefore, pyramid decomposition and wavelet transform can also be potentially useful for SISR image quality evaluation. Gaussian pyramid provides a representation of the same image at multiple scales, using simple low-pass filtering and decimation techniques. The Laplacian pyramid provides a coarse representation of the image as well as a set of detailed images at different scales. Unlike the Gaussian and Laplacian pyramids, Wavelet decomposition provides a complete image representation and perform the image decomposition according to both scale and orientation. Since these image decomposition techniques also decompose an image into basic and detail components, they can be potentially applied to SISR image quality evaluation. However, the challenging problem is that how to extract effective quality-aware features from the decomposed components. This can be a future work that deserves further investigations.

VI. CONCLUSION

This paper focuses on the problem of perceptual quality assessment of real-world SISR. The first contribution is that we construct a real-world SISR quality dataset (i.e., *RealSRQ*) and conduct comparative human subjective studies with 10 representative SISR algorithms. Comprehensive analyses on the results from the subjective studies are also presented. Through subjective studies, we find that traditional SISR algorithms (e.g., ASDS) can perform much better than the deep learning-based algorithms on real-world LR images. The second contribution is that we propose a new objective metric *KLTSRQA* to evaluate the quality of SISR images in a NR manner. Experiments on both real-world and synthetic SISR quality datasets have demonstrated the superiority of *KLTSRQA*. In addition, we find that it is much more challenging to accurately evaluate the quality of real-world SISR images than the synthetic ones. Overall, our *RealSRQ* dataset creates a reliable platform to fairly compare the performance of different image quality metrics on SISR images and our *KLTSRQA* metric offers a more accurate solution to address the challenging problem.

REFERENCES

[1] W. Yang, X. Zhang, Y. Tian, W. Wang, J. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019.

[2] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L. Alberi-Morel, "Low-complexity single image super-resolution based on nonnegative neighbor embedding," in *BMVC*, 2012.

[3] R. Zeydeand Michael Elad and M. Protter, "On single image scale-up using sparse-representations," in *Lecture Notes Comput. Sci*, 2010, pp. 711–730.

[4] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 898–916, 2011.

[5] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[6] J. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5197–5206.

[7] R. Timofte et al., "Ntire 2017 challenge on single image super-resolution: Methods and results," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1110–1121.

[8] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1122–1131.

[9] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3086–3095.

[10] X. Zhang, Q. Chen, R. Ng, and V. Koltun, "Zoom to learn, learn to zoom," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3757–3765.

[11] P. Wei, Z. Xie, H. Lu, Z. Zhan, Q. Ye, W. Zuo, and L. Lin, "Component divide-and-conquer for real-world image super-resolution," in *Computer Vision – ECCV 2020*, 2020, pp. 101–117.

[12] H. R. Vaezi Joze, I. Zharkov, K. Powell, C. Ringler, L. Liang, A. Roulston, M. Lutz, and V. Pradeep, "Imagepairs: Realistic super resolution dataset via beam splitter camera rig," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 2190–2200.

[13] C.-Y. Yang, C. Ma, and M.-H. Yang, "Single-image super-resolution: A benchmark," 09 2014, pp. 372–386.

[14] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017.

[15] G. Shi, W. Wan, J. Wu, X. Xie, W. Dong, and H. Wu, "Sisrset: Single image super-resolution subjective evaluation test and objective quality assessment," *Neurocomputing*, vol. 360, 2019.

[16] F. Zhou, R. Yao, B. Liu, and G. Qiu, "Visual quality assessment for super-resolved images: Database and method," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3528–3541, 2019.

[17] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error measurement to structural similarity," *IEEE Trans. Imgage Process.*, vol. 13, 11 2004.

[18] 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, 2003, pp. 1398–1402 Vol.2.

[19] 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, 2011.

[20] 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.

[21] 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.

[22] 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.

[23] 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, 2014.

[24] J. Wu, W. Lin, G. Shi, and A. Liu, "Perceptual quality metric with internal generative mechanism," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 43–54, 2013.

[25] 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.

[26] Z. Ni, L. Ma, H. Zeng, J. Chen, C. Cai, and K.-K. Ma, "Esim: Edge similarity for screen content image quality assessment," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4818–4831, 2017.

[27] Z. Ni, H. Zeng, L. Ma, J. Hou, J. Chen, and K.-K. Ma, "A gabor feature-based quality assessment model for the screen content images," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4516–4528, 2018.

[28] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4850–4862, 2014.

[29] 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.

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

[31] 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.

[32] L. Liu, Y. Hua, Q. Zhao, H. Huang, and A. C. Bovik, "Blind image quality assessment by relative gradient statistics and adaboosting neural network," *Signal Processing: Image Communication*, vol. 40, pp. 1–15, 2016.

[33] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Processing: Image Communication*, vol. 29, no. 8, pp. 856–863, 2014.

[34] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[35] L. Li, W. Xia, W. Lin, Y. Fang, and S. Wang, "No-reference and robust image sharpness evaluation based on multiscale spatial and spectral features," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 1030–1040, 2017.

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

[37] D. Ghadiyaram and A. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17, no. 1, 01 2017.

[38] 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, 2013.

[39] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.

[40] M. S. Hosseini, Y. Zhang, and K. N. Plataniotis, "Encoding visual sensitivity by maxpol convolution filters for image sharpness assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4510–4525, 2019.

[41] 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.

[42] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 1838–1857, 2011.

[43] T. Peleg and M. Elad, "A statistical prediction model based on sparse representations for single image super-resolution," *IEEE Transactions on Image Processing*, vol. 23, no. 6, pp. 2569–2582, 2014.

[44] R. Timofte, V. Desmet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian Conference on Computer Vision*, 2014, pp. 111–126.

[45] E. Pérez-Pellitero, J. Salvador, J. Ruiz-Hidalgo, and B. Rosenhahn, "Antipodally invariant metrics for fast regression-based super-resolution," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2456–2468, 2016.

[46] C. Dong, C. L. Chen, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*, 2014, pp. 184–199.

[47] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 370–378.

[48] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1646–1654.

[49] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.

[50] K. Zhang, L. Van Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3214–3223.

[51] 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.

[52] J.-M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and A. Dev, "Color and scale: The spatial structure of color images," in *Computer Vision - ECCV 2000*, 2000, pp. 331–341.

[53] J. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts, "Color invariance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 12, pp. 1338–1350, 2001.

[54] N. Lasmar, Y. Stitou, and Y. Berthoumieu, "Multiscale skewed heavy tailed model for texture analysis," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 2281–2284.

[55] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 1, pp. 52–56, 1995.

[56] T. Joachims, "Training linear svms in linear time," *ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.

Qiuping Jiang (M'18) is currently an Associate Professor with Ningbo University, Ningbo, China. He received the Ph.D. degree in Signal and Information Processing from Ningbo University in 2018. From Jan. 2017 to May 2018, he was a visiting student with Nanyang Technological University, Singapore. His research interests include image processing, visual perception, and computer vision. He received the Distinguished Youth Scholar Funding of Zhejiang Natural Science Foundation (ZJNSF), the Best Paper Honorable Mention Award of the *Journal of Visual Communication and Image Representation*, and the Excellent Doctoral Dissertation Award of Zhejiang Province. He also serves as the Associate Editor of *Journal of Electronic Imaging* and *APSIPA Trans. on Information and Signal Processing*, and the Area Chair/Session Chair/PC member for IJCAI/AAAI/ACM-MM/ICME/ICIP/APSIPA-ASC.

Zhentao Liu is currently an forth-year undergraduate student major in Communication Engineering with the School of Information Science and Engineering, Ningbo University, China. His research interests include image processing, image quality assessment, and visual perception modeling.

Ke Gu is a Professor at the Beijing University of Technology, Beijing, China. His research interests include environmental perception, image processing, quality assessment, and machine learning. He received the Best Paper Award from the IEEE Transactions on Multimedia (T-MM) and the Best Student Paper Award at the IEEE International Conference on Multimedia and Expo (ICME) in 2016. He was the Leading Special Session Organizer in the VCIP 2016 and the ICIP 2017. He is currently an Associate Editor for Computer Animation and Virtual Worlds (CAVW) and IET Image Processing (IET-IPR), an Area Editor for Signal Processing Image Communication (SPIC), and an Editor for Applied Sciences, Displays, and Entropy. He is a Reviewer for 20 top SCI journals.

Feng Shao received the B.S. and Ph.D. degrees in Electronic Science and Technology from Zhejiang University, Hangzhou, China, in 2002 and 2007, respectively. He is currently a Professor with the Faculty of Information Science and Engineering, Ningbo University, Ningbo, China. In 2012, he was a visiting scholar with the School of Computer Engineering, Nanyang Technological University, Singapore. He was the receipt of the Excellent Youth Scholar Funding of Natural Science Foundation of China. His research interests include image processing, image quality assessment, and immersive media computing.

Xinfeng Zhang (M'19) (M'16-SM'20) received the B.S. degree in computer science from the Hebei University of Technology, Tianjin, China, in 2007, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2014. From 2014 to 2017, he was a Research Fellow with the Rapid-Rich Object Search Lab, Nanyang Technological University, Singapore. From Oct. 2017 to Oct. 2018, he was a Post-Doctoral Fellow with the School of Electrical Engineering System, University of Southern California, Los Angeles, CA, USA. From Dec. 2018 to Aug. 2019, he was a Research Fellow with the department of Computer Science, City University of Hong Kong.

He currently is an Assistant Professor with the School of Computer Science and Technology, University of Chinese Academy of Sciences. He authored more than 100 refereed journal/conference papers and received the Best Paper Award of IEEE Multimedia 2018, the Best Paper Award at the 2017 Pacific-Rim Conference on Multimedia (PCM) and the Best Student Paper Award in IEEE International Conference on Image Processing 2018. His research interests include video compression and processing, image/video quality assessment, and 3D point cloud processing.

Hantao Liu received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands in 2011. He is currently an Associate Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. He is an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology and the IEEE Signal Processing Letters.

Weisi Lin (F'16) is currently a Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He received the Bachelor's degree in Electronics and then a Master's degree in Digital Signal Processing from Sun Yat-Sen University, Guangzhou, China, and the Ph.D. degree in Computer Vision from King's College, London University, UK. His research interests include image processing, perceptual modeling, video compression, multimedia communication, and computer vision.

He is a Fellow of the IEEE and IET, an Honorary Fellow of the Singapore Institute of Engineering Technologists, and a Chartered Engineer in U.K. He was the Chair of the IEEE MMTC Special Interest Group on Quality of Experience. He was awarded as the Distinguished Lecturer for IEEE Circuits and Systems Society in 2016-2017. He served as a Lead Guest Editor for a Special Issue on Perceptual Signal Processing of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING in 2012. He also has served or serves as an Associate Editor for IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Multimedia, IEEE Signal Processing Letters, and Journal of Visual Communication and Image Representation.