Predicting the Quality of View Synthesis With Color-Depth Image Fusion

Leida Li, *Member, IEEE*, Yipo Huang, Jinjian Wu, Ke Gu, Yuming Fang, *Senior Member, IEEE*

***Abstract*—With the increasing prevalence of free-viewpoint video applications, virtual view synthesis has attracted extensive attention. In view synthesis, a new viewpoint is generated from the input color and depth images with a depth-image-based rendering (DIBR) algorithm. Current quality evaluation models for view synthesis typically operate on the synthesized images, i.e. after the DIBR process, which is computationally expensive. So a natural question is that can we infer the quality of DIBR-based synthesized images using the input color and depth images directly without performing the intricate DIBR operation. With this motivation, this paper presents a no-reference image quality prediction model for view synthesis via COlor-Depth Image Fusion, dubbed CODIF, where the actual DIBR is not needed. First, object boundary regions are detected from the color image, and a Wavelet-based image fusion method is proposed to imitate the interaction between color and depth images during the DIBR process. Then statistical features of the interactional regions and natural regions are extracted from the fused color-depth image to portray the influences of distortions in color/depth images on the quality of synthesized views. Finally, all statistical features are utilized to learn the quality prediction model for view synthesis. Extensive experiments on public view synthesis databases demonstrate the advantages of the proposed metric in predicting the quality of view synthesis, and it even suppresses the state-of-the-art post-DIBR view synthesis quality metrics.**

***Index Terms*—View synthesis, DIBR, color-depth fusion, interactional region, quality prediction.**

I. Introduction

NOWADAYS, view synthesis has attracted extensive attention owing to the increasing prevalence of free-viewpoint video applications [1]–[4]. In virtual view synthesis, a new viewpoint is generated using the input color and depth images jointly, where depth-image-based rendering (DIBR) is commonly used [5], [6]. In practice, various distortions could be introduced into the color and depth images, from acquisition, compression to transmission, which in turn impair the perceptual quality of the synthesized views. Quality assessment for view synthesis is of paramount importance in joint color/depth image coding and bit allocation [7]. Without an effective view synthesis quality index, benchmarking and optimization of DIBR algorithms is also troublesome [8].

In the literature, a mass of image quality assessment (IQA) methods have been proposed. Based on the availability of high-quality pristine image, the existing IQA models can be categorized into full-reference (FR), reduced-reference (RR) and no-reference (NR) [9]–[12]. Representative FR-IQA models include Structural Similarity (SSIM) [13], Feature Similarity (FSIM) [14] and Visual Information Fidelity (VIF) [15], which utilize both the test image and the corresponding perfect-quality pristine image to calculate the quality score. RR-IQA metrics utilize side information of the reference images, typically through feature extraction, to achieve quality evaluation [16], [17]. In contrast, NR-IQA models calculate image quality score based on the distorted image directly. Popular NR-IQA metrics include the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [18], Natural Image Quality Evaluator (NIQE) [19], Integrated Local NIQE (IL-NIQE) [20] and M3 [21], just to name a few. The above IQA models have achieved significant success in evaluating the quality of natural scene images. However, they are usually limited when used for view synthesis [8]. This is mainly because that distortions in view synthesis are much more complicated. Specifically, in addition to the common distortions in natural scenes, distortions of the input depth images and the DIBR operation easily introduce local geometric distortions to the synthesized views, which cannot be handled by traditional IQA metrics [8].

In this paper, we present a new no-reference quality prediction model for view synthesis based on COlor-Depth Image Fusion, called CODIF. The underlying idea is to imitate the interactions between color and depth images during the DIBR process by devising an effective Wavelet-based color-depth image fusion approach. In addition, taking into account the different functional mechanisms of color and depth images in view synthesis, interactional regions and natural regions are defined based on the fused color-depth image, and different sets of synthesis-aware features are devised for predicting the degradations in view synthesis. The support vector regression (SVR) is adopted for building the final quality prediction model. The advantage of the proposed metric is also demonstrated based on extensive experiments and comparisons with both

L. Li is with the Guangzhou Institute of Technology, Xidian University, Guangzhou 510555, China, and also with the Pazhou Lab, Guangzhou 510330, China (e-mail: ldli@xidian.edu.cn).

Y. Huang is with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China (e-mail: huangyipo@hotmail.com).

J. Wu is with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: jinjian.wu@mail.xidian.edu.cn).

K. Gu is with Beijing Key Laboratory of Computational Intelligence and Intelligent System, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (e-mail: guke.doctor@gmail.com).

Y. Fang is with the School of Information Management, Jiangxi University of Finance and Economics, Nanchang, Jiangxi 330032, China (e-mail: fa0001ng@e.ntu.edu.sg).

This work was supported in part by the National Natural Science Foundation of China under Grants 61771473, 61991451 and 61379143, the Key Project of Shaanxi Provincial Department of Education (Collaborative Innovation Center) under Grant 20JY024, the Science and Technology Plan of Xi'an under Grant 20191122015KYPT011JC013, the Natural Science Foundation of Jiangsu Province under Grant BK20181354, and the Six Talent Peaks High-level Talents in Jiangsu Province under Grant XYDXX-063. *(Corresponding author: Yipo Huang.)*

pre-DIBR and post-DIBR view synthesis quality models. The contributions of our work can be summarized as follows.

- We propose a novel NR quality prediction model for view synthesis, which can infer the quality of synthesized images using the input color and depth images directly without performing the computationally expensive DIBR process.
- A Wavelet-based color-depth image fusion approach is proposed to imitate the interactions between color and depth images in the DIBR process.
- We propose to predict the view synthesis quality based on the fused color-depth image from interactional regions and natural regions, with the purpose to portray the two categories of distortions in view synthesis. Furthermore, a Tchebichef moment-based statistical feature is devised to measure the geometric distortions in the interactional regions.

II. Related Work

In the past decade, several quality metrics for view synthesis have been proposed. Battisti *et al.* [22] proposed the 3D Synthesized view Image quality Metric (3DSwIM) by analyzing the similarities of statistical features between the distorted and reference synthesized images in the Wavelet domain. In [23], [24], the Morphological Wavelet PSNR (MW-PSNR) and Morphological Pyramid PSNR (MP-PSNR) metrics were proposed based on the Morphological Wavelet and Morphological Pyramids representations, respectively. In addition, their reduced versions, i.e. RMW-PSNR [23] and RMP-PSNR [25], were also proposed by discarding the coarse-scale subbands. Li *et al.* [8] employed the SIFT-flow-based geometric warping to localize the disoccluded areas and evaluated the local geometric distortions from the detected disoccluded areas. Tian *et al.* [26] proposed the NIQSV metric by quantifying the distortions in luminance, contrast and saturation based on the morphological operations. Further, they proposed to measure the blur regions, holes and the stretching distortions, producing the NIQSV+ metric [27]. Gu *et al.* [28] first employed the autoregression (AR) model to generate a reconstruction image, based on which the error between the AR-reconstructed image and the corresponding DIBR-synthesized image was quantified to measure the geometric distortions. In [29], a No-Reference Morphological Wavelet with Threshold (NR-MWT) metric was proposed to measure the perceptual quality of synthesized images and videos. First, the morphological Wavelet was adopted for extracting the high-frequency visual content. Then a threshold was introduced to portray the most significant regions in the high-frequency Wavelet transform. Finally, the quality score was generated by only using coefficients above the threshold. Jakhetiya *et al.* [30] measured the visual quality of synthesized images, where geometric and structural distortions were highlighted based on median filtering. In [31], a novel method was proposed with Multiscale Natural Scene Statistical analysis (MNSS). First, the self similarity-based model was employed for measuring the DIBR-introduced geometric distortions. Then the degradations in main structures were also evaluated by the proposed statistical model. The final quality score was calculated by integrating them using a straightforward multiplication. Zhou *et al.* [5] addressed a blind view synthesis quality index by using the Difference-of-Gaussian feature to measure edge degradation and textural unnaturalness. More recently, Wang *et al.* [32] decomposed the DIBR-synthesized images using the discrete Wavelet transform. Then the geometric distortions were captured using the edge similarities between the binarized low-frequency and high-frequency subbands. In addition, the sharpness of the DIBR-synthesized image was evaluated by the log-energies of wavelet subbands. The final quality score of the synthesized image was calculated by integrating the geometric distortions and overall sharpness.

The above view synthesis quality metrics follow the same pipeline to perform quality assessment using the synthesized images, typically with the help of color images of the original viewpoint. In practice, the DIBR process consists of warping and rendering operations, which are computationally expensive. In addition, evaluating the view synthesis quality using both the DIBR-synthesized image and the original color image typically needs to determine the dense correspondence between the two images, e.g. SIFT-flow-based approach in [8], so that the geometric distortions can be accurately located. This also incurs heavy computational burden. Therefore, a straightforward question is that can we develop an efficient model to predict the quality of view synthesis without performing the complicated DIBR operation, i.e. evaluating using the input color and depth images directly. By this means, view synthesis systems can be more flexible, considering that if the input color/depth images cannot generate satisfactory synthesized viewpoint (by prediction), their quality can be adjusted before sending to the time-consuming DIBR process.

To the best of our knowledge, only two relevant approaches have been reported towards blind pre-DIBR quality assessment of view synthesis using color and depth images [33], [34]. Wang *et al.* [33] proposed a novel FR quality model to predict the quality of synthesized views based on content-aware weighting of both color and depth images. The well-known SSIM [13] metric was utilized to calculate two quality indication maps between the degraded color/depth images and the corresponding reference images. An information content weighting map was also calculated by normalizing the depth similarity map using the original color image. Finally, an overall quality measure was generated by combining the color and depth quality maps. Shao *et al.* [34] reported a high-efficiency view synthesis quality prediction (HEVSQP) index using sparse representation of color and depth images. They first characterized the relationship between the synthesized image and the input color/depth distortions. Then, color-involved view synthesis quality prediction (CI-VSQP) and depth-involved view synthesis quality prediction (DI-VSQP) were achieved based on sparse representation features. Finally, an overall quality score was generated by performing a pooling between CI-VSQP and DI-VSQP. The effectiveness of the HEVSQP metric was verified based on experiments in a synthesized video quality database.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3024882, IEEE Transactions on Circuits and Systems for Video Technology

III. Proposed Pre-DIBR Quality Prediction Model

We first analyze the distortion characteristics in view synthesis. Then we detail the proposed view synthesis quality prediction model, including color-depth image fusion, statistical feature extraction from interactional and natural regions, and the regression-based quality model training.

A. Distortion Analysis in View Synthesis

In DIBR-based view synthesis, both the conventional distortions and geometric distortions are present in the synthesized images, which are caused by the distorted color and depth images [35]–[37]. In Fig. 1, we show two images synthesized with different combinations of input color and depth images, together with the SSIM [13] maps between the synthesized images and the corresponding pristine images. In the SSIM maps, darker region indicates heavier distortion. Fig. 1(c) shows the SSIM map of image (a), which is synthesized using distorted color image and undistorted depth image. Fig. 1(d) shows the SSIM map of image (b), which is synthesized using undistorted color image and distorted depth image. By comparing the two SSIM maps, it can be easily observed that the synthesis distortions caused by distorted color image mainly occur in natural regions of the synthesized image. In contrast, the synthesis distortions caused by distorted depth image mainly occur around object boundaries. This is because that distortions in color image will be straightforwardly transferred to the synthesized image. On the contrary, depth images are typically employed to assist the warping operation in DIBR, which typically influences object boundaries, because object boundaries in depth images represent different distances from the object to the camera [38]. In other words, distortions in depth images usually cause geometric distortions around object boundaries in the synthesized image, which is mainly because of the edge misalignment between color image and depth image in the DIBR process.

Inspired by the above observations, this paper presents a NR quality prediction model for view synthesis based on color-depth image fusion. Our objective is to predict the quality of DIBR-based view synthesis blindly based on the input color and depth images directly, without performing the computationally expensive DIBR operation. Fig. 2 shows a schematic diagram of the proposed metric.

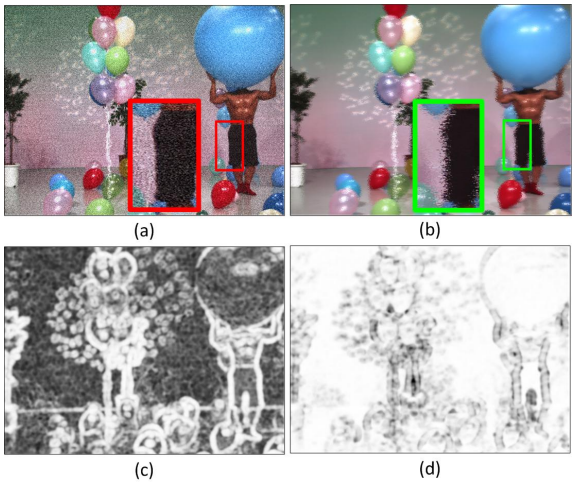

Fig. 1: SSIM maps of the synthesized images using different combinations of color and depth images. (a): synthesized image using distorted color image and undistorted depth image; (b): synthesized image using undistorted color image and distorted depth image; (c): SSIM map of image (a); (d): SSIM map of image (b).

B. Color-Depth Image Fusion

The DIBR process consists of a warping stage and a rendering stage. During warping, the input color image is first mapped into the 3D space under the guidance of the corresponding depth image. Then an inverse mapping is performed to obtain the target view. Rendering is mainly to fill the holes, a.k.a. disoccluded regions, introduced in the warping process, based on which the synthesized views can be obtained [8]. In this process, color and depth images interact to generate the synthesized view. In fact, the DIBR process can be viewed as a process to fuse the input color and depth information for generating the synthesized view. From this perspective, given a specific DIBR algorithm, degradations in the input color and depth images determines the final synthesized image quality. This motivates us to imitate the interactional mechanism between color and depth images by proposing a color-depth fusion approach, and further to predict the quality of view synthesis using the fused color-depth image.

Generally, edges in color images fall into different categories, such as object boundaries, shadows and color patterns [39]–[41]. By contrast, edges in depth images only represent object boundaries, because depth map measures the distance between an object and the camera. Therefore, depth edges constitute a subset of color image edges, and this subset of edges will interact between color and depth in the DIBR process. Ideally, depth and color edges should coexist and they should be aligned exactly around object boundaries. In other words, for high-quality color and depth images, the location of edges is expected to be coincident in object boundary regions. As a result, if we fuse the edge information of the high-quality color and depth images around object boundary regions, the structure of the fused edge should be consistent with the real object edge. On the contrary, if color and depth images are subject to distortions, there will be misalignment between the fused color-depth edges and the real object edges, which will in turn degrade the quality of the synthesized view. Therefore, color and depth images can be fused to imitate their mutual interactions during the actual DIBR process, and conversely the fused color-depth image can be utilized to predict the quality of the synthesized image with no need of the actual DIBR operation. This also illustrates our design philosophy.

Towards the above goal, we first propose a simple and effective Wavelet-based color-depth image fusion approach. This is mainly inspired by the fact that image edges are mainly present as detail information and the high-frequency coefficients of Discrete Wavelet Transform (DWT) are efficient in portraying image details [42]. Furthermore, DWT has the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3024882, IEEE Transactions on Circuits and Systems for Video Technology

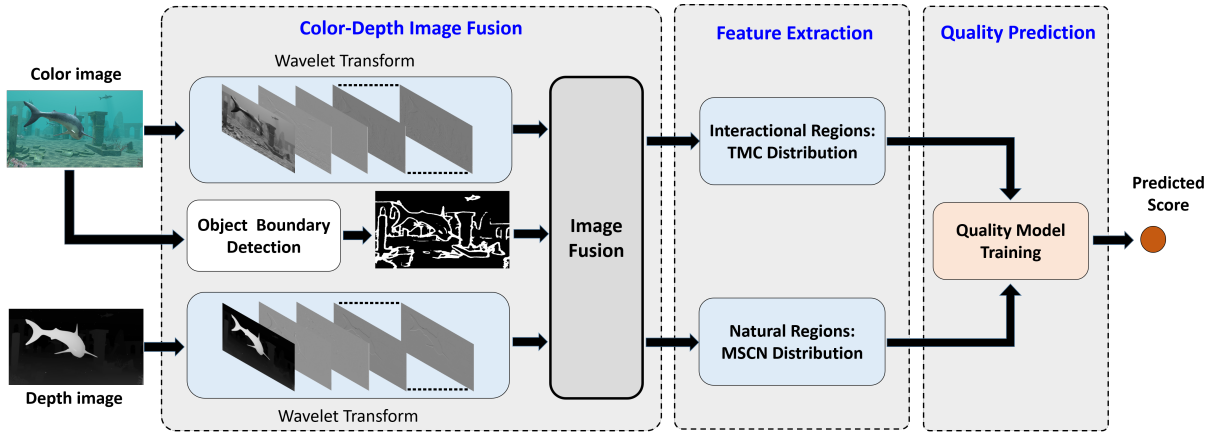

Fig. 2: Diagram of the proposed pre-DIBR quality prediction model for view synthesis.

advantage of multi-scale geometric analysis, which is similar to the hierarchical property Human Visual System (HVS) [43]. These merits make DWT an ideal choice for image fusion [44]–[46]. In this work, the Haar Wavelet is adopted [47].

As aforementioned, the interaction between color and depth images occurs around object boundary regions. Therefore, we first resort to an object boundary detection approach for locating these regions. For clarity, we use interactional regions (IRs) to denote the object boundary regions hereafter, while the other regions outside the IRs are called natural regions (NRs). It is worth to emphasize that the interaction between color and depth distortions mainly destroys the edges of salient objects, so traditional edge detection methods are not applicable here. The reason is that traditional edge detection methods tends to detect too many edges and many of them are textures inside objects [48], which are not useful in our problem. In this work, we employ the Holistically-nested Edge Detection (HED) model [49] for accomplishing this task, which is pre-trained on the BSDS database [50] and can achieve very competitive performance in salient object boundary detection. Since the object boundary regions, i.e. IRs, can be fully covered by the edges detected using the HED model, the HED edges are directly used to locate the interactional regions. Fig. 3 shows an example of object boundary detection and interactional region localization. It can be seen from the figure that the detected regions are all located at salient object boundaries, which are critical for view synthesis. In the proposed method, these regions are used to guide the subsequent color-depth image fusion.

With the object boundary regions, the proposed color-depth image fusion approach operates as follows. First, the input color and depth images are both decomposed by two-level Wavelet transform, producing seven subbands including six high-frequency subbands and one low-frequency subband. Since low-frequency subband typically represents the average characteristic of an image and high-frequency subband represents image detail information, we only use high-frequency subbands of the color and depth images to conduct the fusion, where the IRs mask (denoted as $\mathbf{I}_{IRs}$) is used as guide information. Specifically, the high-frequency coefficients of the input color and depth images (denoted as $\mathbf{S}_{LH}$, $\mathbf{S}_{HL}$ and $\mathbf{S}_{HH}$) are averaged to obtain the high-frequency coefficients of the fused image, which is achieved by

$$\mathbf{S}^F_{LH_n}(i,j) = \begin{cases} \frac{\mathbf{S}^C_{LH_n}(i,j)+\mathbf{S}^D_{LH_n}(i,j)}{2}, & \mathbf{I}_{IRs}(i,j) = 1 \\ \mathbf{S}^C_{LH_n}(i,j), & \mathbf{I}_{IRs}(i,j) = 0 \end{cases} \quad (1)$$

$$\mathbf{S}^F_{HL_n}(i,j) = \begin{cases} \frac{\mathbf{S}^C_{HL_n}(i,j)+\mathbf{S}^D_{HL_n}(i,j)}{2}, & \mathbf{I}_{IRs}(i,j) = 1 \\ \mathbf{S}^C_{HL_n}(i,j), & \mathbf{I}_{IRs}(i,j) = 0 \end{cases} \quad (2)$$

$$\mathbf{S}^F_{HH_n}(i,j) = \begin{cases} \frac{\mathbf{S}^C_{HH_n}(i,j)+\mathbf{S}^D_{HH_n}(i,j)}{2}, & \mathbf{I}_{IRs}(i,j) = 1 \\ \mathbf{S}^C_{HH_n}(i,j), & \mathbf{I}_{IRs}(i,j) = 0 \end{cases} \quad (3)$$

where $\mathbf{S}^F$ denotes the fused high-frequency subband, $\mathbf{S}^C$ represents the high-frequency subband of the color image, $\mathbf{S}^D$ denotes the high-frequency subband of the depth image, and $n \in [1, 2]$ denotes the decomposition level in the Wavelet transform.

After fusing the high-frequency DWT coefficients, the fused color-depth image can be easily obtained from the fused high-frequency subbands and the low-frequency subband of the color image by performing an inverse DWT. To show the effectiveness of the proposed color-depth image fusion approach in imitating the interactions between color and depth images during DIBR, we show two examples in Fig. 4, where the color-depth fused images and the corresponding DIBR-synthesized images are shown together. Since distortions in color input are typically transferred to the DIBR-synthesized view directly, we only show examples where undistorted color image and distorted depth image are used in the DIBR process. In this case, the synthesis distortions are mainly located at object boundaries. By comparing the object boundary regions in the fused color-depth image and the corresponding DIBR-synthesized image in Fig. 4, it is easily observed that the distortion characteristics in the two images are quite similar. This confirms the effectiveness of the proposed color-depth image fusion approach.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3024882, IEEE Transactions on Circuits and Systems for Video Technology

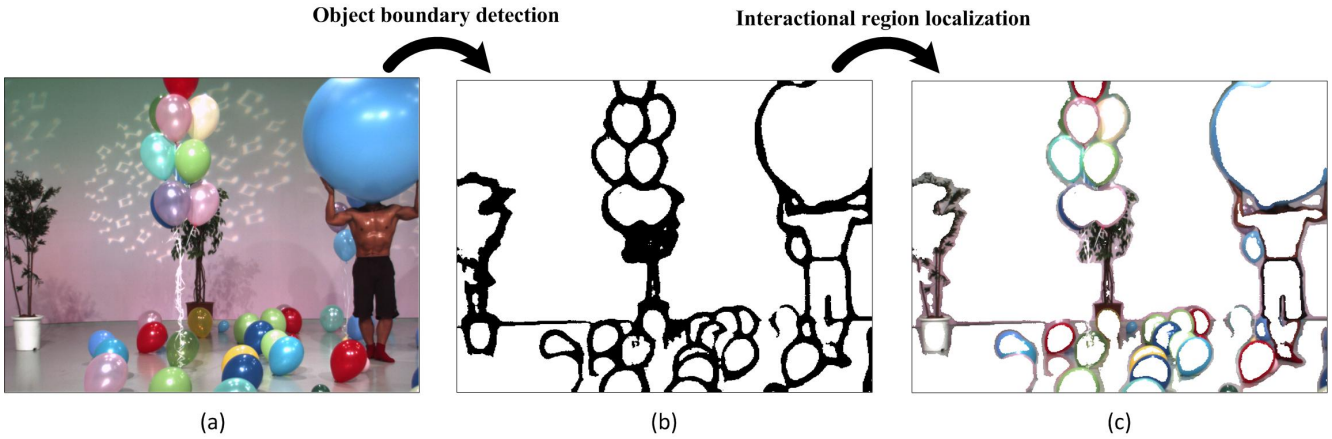

Fig. 3: Illustration of object boundary detection and interactional region localization. (a) color image, (b) detected object boundaries, and (c) interactional regions.

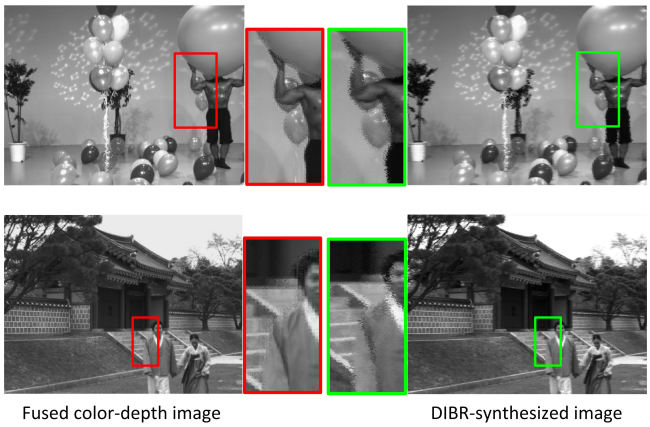

Fig. 4: Distortion characteristics in the fused color-depth image and the actual DIBR-synthesized image.

C. Statistical Feature Extraction

Following the color-depth image fusion process, we know that the interactional regions (IRs) of the fused image are generated by combining the color and depth images in object boundary regions, while the natural regions (NRs) maintain the distortion characteristics of the input color image. To be specific, distortions in interactional regions are mainly geometric distortions, while distortions in natural regions are mainly common distortions. For this reason, we propose to apply two different feature extraction strategies for IRs and NRs of the fused color-depth image.

For IRs, distortions of the fused image are mainly present as edge degradation, which are related to object shapes. It has been shown that Tchebichef moments are effective in portraying shapes in an image [9]. In this work, the Tchebichef moment is used to measure the edge degradation in the interactional regions of the fused color-depth image. In addition, the moment-based features are calculated in the gradient image, since gradient domain is more effective for shape description [51]. Therefore, for interactional regions $\mathbf{I}_{IRs}$, the gradient magnitude (GM) image $\mathbf{I}_{GM}$ is first calculated:

$$\mathbf{I}_{GM} = \frac{|\mathbf{G}_x| + |\mathbf{G}_y|}{2}, \tag{4}$$

$$\mathbf{G}_x = [-1\ 0\ 1] * \mathbf{I}_{IRs}, \quad \mathbf{G}_y = [-1\ 0\ 1]' * \mathbf{I}_{IRs}, \tag{5}$$

where $*$ denotes the convolution operation and $'$ is the transpose operation.

Next, the GM image of the interactional regions is first partitioned into equal-size blocks with $K \times K$ pixels. Then, for the n^{th} block, the Tchebichef moments from zero order to the $[(K-1)+(K-1)]$ order are computed as:

$$\mathbf{T}_n = \begin{pmatrix} t_{00} & t_{01} & \cdots & t_{0(K-1)} \\ t_{10} & t_{11} & \cdots & t_{1(K-1)} \\ \vdots & \vdots & \ddots & \vdots \\ t_{(K-1)0} & t_{(K-1)1} & \cdots & t_{(K-1)(K-1)} \end{pmatrix}. \tag{6}$$

Fig. 5 shows an example of the IRs and NRs of a color-depth fused image, where image (d) further shows the histogram distribution of Tchebichef moment coefficients (TMC) collected from all blocks in IRs. To extract quality-aware features, the Asymmetric Generalized Gaussian Distribution (AGGD) is adopted to fit the distribution of TMC coefficients [19], which is defined by

$$f(x;\theta,\sigma_l^2,\sigma_r^2) = \begin{cases} \frac{\theta}{(\beta_l+\beta_r)\Gamma(\frac{1}{\theta})} exp\left(-\left(\frac{-x}{\beta_r}\right)^{\theta}\right), & x \geq 0, \\ \frac{\theta}{(\beta_l+\beta_r)\Gamma(\frac{1}{\theta})} exp\left(-\left(\frac{-x}{\beta_l}\right)^{\theta}\right), & x < 0, \end{cases} \tag{7}$$

where θ donates the shape of the distribution, σ_l^2, σ_r^2 donate the left side and right side spreads of the TMC distribution respectively, and $\Gamma(\cdot)$ is the gamma function:

$$\Gamma(a) = \int_0^{\infty} t^{a-1} e^{-t} dt. \tag{8}$$

In addition to the above three features, the mean value is also calculated:

$$\mu = (\sigma_r - \sigma_l)\frac{\Gamma(\frac{2}{\theta})}{\Gamma(\frac{1}{\theta})}. \tag{9}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3024882, IEEE Transactions on Circuits and Systems for Video Technology

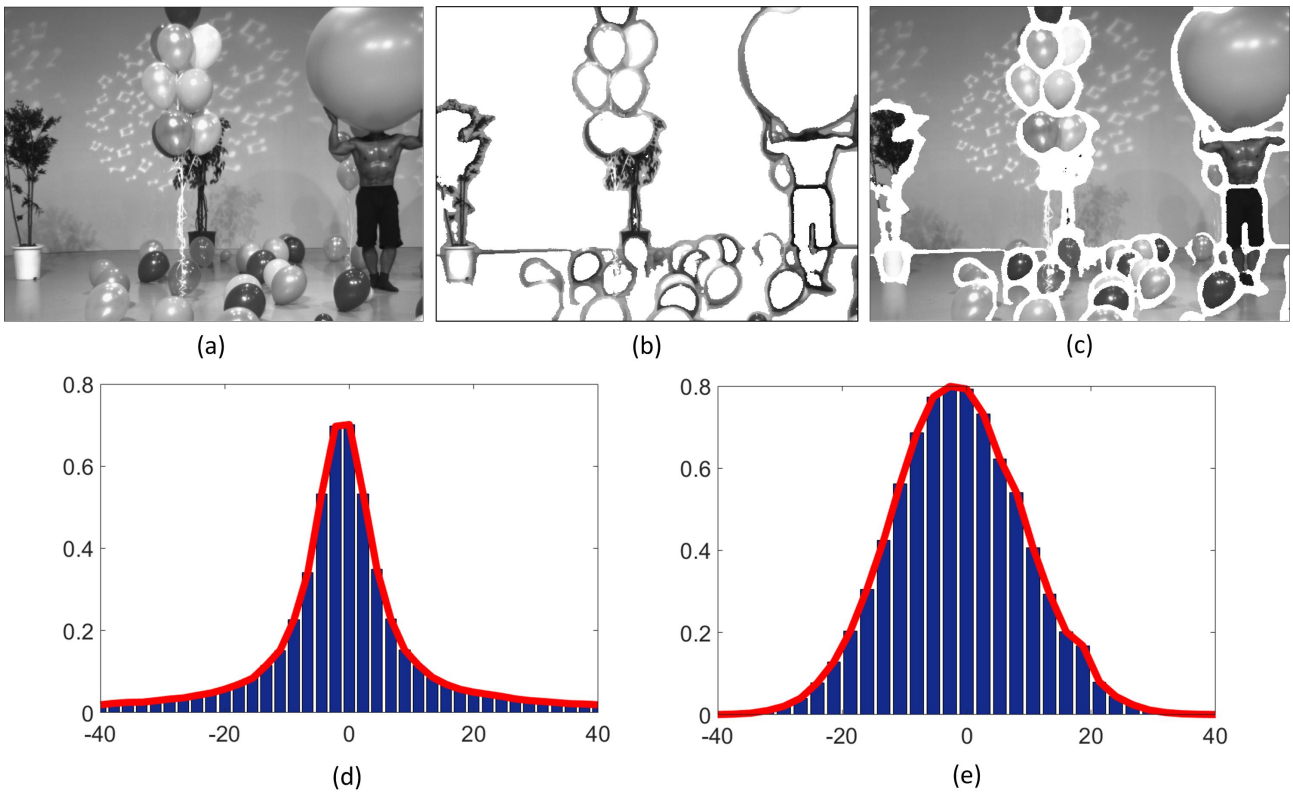

Fig. 5: Illustration of NRs and IRs as well as the corresponding distributions. (a) fused color-depth image; (b) IRs of image (a); (c) NRs of image (a); (d) TMC distribution of image (b); (e): MSCN distribution of image (c).

Four parameters $(\theta, \sigma_l^2, \sigma_r^2, \mu)$ can be calculated from the TMC (IRs) distributions, constituting the first set of quality-aware features.

For NRs, as the distortions in the input color image will be straightforwardly transferred to the synthesized images, the natural scene statistics (NSS) features can be used for NRs. In this work, the mean subtracted contrast normalized (MSCN) coefficients are first calculated [52], and the distribution parameters are employed for measuring the distortions in NRs.

Given the natural regions $\mathbf{I}_{NRs}$ of the color-depth fused image, the local divisive normalization at each pixel $\mathbf{I}_{NRs}(i,j)$ is performed, which serves as decorrelation to the brightness values of neighboring pixels. Then, the MSCN coefficients are computed as:

$$\hat{\mathbf{I}} = \frac{\mathbf{I}_{NRs}(i,j) - \mu(i,j)}{\sigma(i,j) + C}, \tag{10}$$

where $\mu(i,j)$ and $\sigma(i,j)$ denote the local mean and standard deviation, which are further defined as [18]:

$$\mu(i,j) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} \omega_{k,l} I_{k,l}(i,j), \tag{11}$$

$$\sigma(i,j) = \sqrt{\sum_{k=-K}^{K} \sum_{l=-L}^{L} \omega_{k,l} (I_{k,l}(i,j) - \mu(i,j))^2}, \tag{12}$$

where ω denotes the Gaussian function, and $K = L = 3$.

In Fig. 5(e), we show an example of the histogram distributions of MSCN coefficients in the NRs of the color-depth fused image (a). Similarly, we also employ the AGGD function to fit the distribution of the MSCN coefficients, producing the second set of quality-aware features.

It has been demonstrated that the human visual system possesses the multi-scale characteristics when perceiving the visual scenes [53]–[56]. To adapt to this property, in this work the Gaussian low-pass filtering [57] is utilized to generate a five-scale representation space of the fused image, and the above two sets of features are extracted accordingly. Finally, 40 statistical features are extracted for each fused color-depth image, including 20 IRs-TMC features and 20 NRs-MSCN features.

Fig. 6 shows an example of fitted curves of AGGD functions for IRs and NRs, where color/depth images are contaminated by various distortions. From the figure, it is obvious that different distortions generate distributions with different shapes. In accordance with the AGGD fitting curves, different sets of parameters could be obtained. As a result, the statistical features extracted in this paper are sensitive to distortion characteristics, which in turn confirms the validity of the proposed features.

D. Quality Prediction

In order to map the above statistical features into an overall score for predicting the quality of view synthesis, we employ the SVR [58] to learn the quality prediction model. In real-world applications, given a new pair of input color and depth images, the trained SVR model can be utilized to predict the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3024882, IEEE Transactions on Circuits and Systems for Video Technology

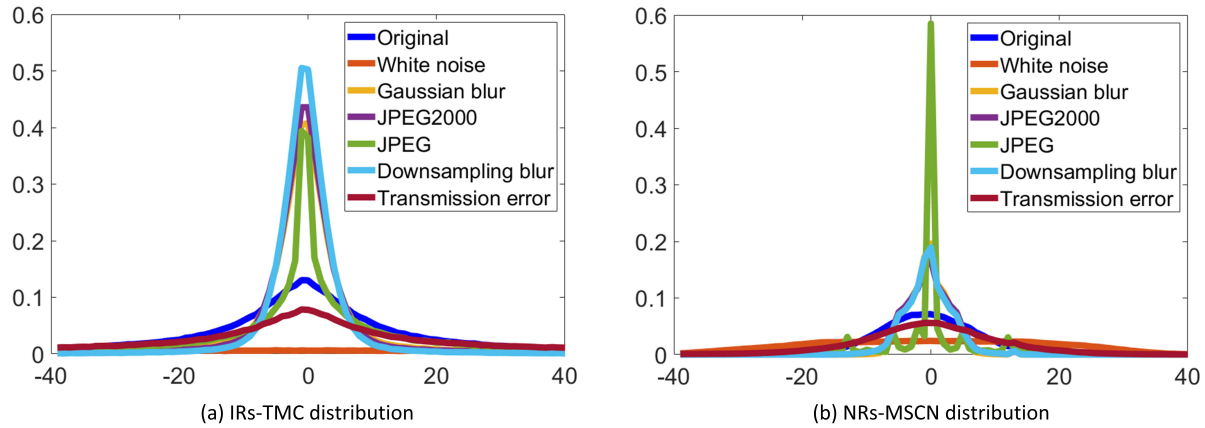

Fig. 6: Fitted curves of TMC distribution in interactional regions and MSCN distribution in natural regions of fused color-depth image with different types of distortions. For simplicity, color and depth images are subject to the same type of distortion in this example.

quality score of the target synthesized image. In this process, the actual DIBR operation is not performed, and the quality prediction is achieved solely using the distorted color and depth images directly in a blind (no-reference) manner. In this work, the Radial Basis Function (RBF) [58] kernel is adopted for training the SVR model.

IV. Experimental Results

A. Evaluation Protocols

In this part, we perform a series of experiments to verify the performance of the proposed quality prediction metric. Two publicly available databases for view synthesis quality evaluation are used, including MCL-3D [59] and IST [60].

MCL-3D [59]. This database consists of 684 synthesized image pairs. Among them, 648 image pairs are generated by the View Synthesis Reference Software (VSRS) [59] using the color-depth image pairs. Six kinds of distortions are applied to the input color and/or depth images, namely Gaussian blur, JPEG compression, down-sampling blurring, additive white noise, JPEG2000 and transmission error. Further, four different distortion levels are applied for each distortion type. Moreover, there are three configurations for view synthesis in the database, including 1) undistorted color image with distorted depth image, 2) distorted color image with undistorted depth image, and 3) distorted color image with distorted depth image. In the proposed metric, we employ the input color-depth image pairs to predict the quality of synthesized views, so the Mean Opinion Score (MOS) of the synthesized image is used as the ground truth of the input color-depth image pair.

IST database [60]. This database contains 180 synthesized image pairs. Among them, 60 pairs are obtained based on the VSRS algorithm [59], and the other 120 pairs are obtained using the VSIM algorithm [61]. The input color and depth images are both degraded by different degrees of compression artifacts. Since the proposed metric is designed for quality prediction of synthesized views using input color-depth images, the DIBR algorithm needs to be fixed during test. So we conduct two sets of experiments respectively for the DIBR algorithms VSRS and VSIM. Similarly, the MOS value of the synthesized image is used as the ground truth of the input color-depth image pair. Moreover, confidence intervals of the individual MOS values are also provided in the database.

We adopt four widely used criteria for performance evaluation, including Pearson Linear Correlation Coefficient (PLCC), Spearman Rank order Correlation Coefficient (SRCC), Kendalls Rank Correlation Coefficient (KRCC) and Root Mean Square Error (RMSE). Further, considering the uncertainty of the subjective scores, we also calculate the epsilon-insensitive RMSE (RMSE*) using confidence intervals on IST database. This criterion is recommended by ITU-T P.1401 [62]. Since the MCL-3D database does not provide the individual MOS values or confidence intervals, the corresponding RMSE* values can not be calculated on this database. Among the mentioned criteria, PLCC, RMSE and RMSE* are used to evaluate the prediction accuracy, while SRCC and KRCC are used to evaluate the prediction monotonicity. A better quality metric should achieve higher SRCC, KRCC and PLCC values, as well as lower RMSE and RMSE* values. To compute PLCC, RMSE and RMSE*, the following five-parameter nonlinear mapping is first performed:

$$f(x) = \xi_1 \left(0.5 - \frac{1}{1 + e^{\xi_2 (x - \xi_3)}} \right) + \xi_4 x + \xi_5, \qquad (13)$$

where x donates the prediction score, $f(x)$ denotes the mapped score, and ξ_i, i=1, 2, ..., 5, are the fitting parameters.

B. Performance Evaluation

We compare the performance of the proposed CODIF metric with the relevant state-of-the-arts. Four popular general-purpose NR-IQA metrics are compared, including NIQE [19], BRISQUE [18], IL-NIQE [20] and M3 [21]. The compared quality metrics for synthesized images (post-DIBR) include MW-PSNR [23], MP-PSNR [24], LOGS [8], SET [5], Ref. [30] and NIQSV [26]. Furthermore, the pioneer pre-DIBR quality index [33] is also compared. For the learning-based quality metrics, 80% of the images are randomly chosen for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3024882, IEEE Transactions on Circuits and Systems for Video Technology

TABLE I: Performances of view synthesis quality metrics on MCL-3D and IST databases.

Category	Metric	Type	VSRS on MCL-3D Database				VSRS on IST Database					VSIM on IST Database				
			PLCC	SRCC	KRCC	RMSE	PLCC	SRCC	KRCC	RMSE	RMSE*	PLCC	SRCC	KRCC	RMSE	RMSE*
Post-DIBR	NIQE [19]	GNR	0.7543	0.7104	0.5325	1.6980	0.6395	0.6202	0.4494	0.7472	0.5453	0.5836	0.5858	0.4122	0.7273	0.5359
	BRISQUE [18]	GNR	0.6944	0.6646	0.4679	1.8722	0.7454	0.7108	0.5426	0.5946	0.3985	0.6512	0.5877	0.4242	0.6208	0.4395
	IL-NIQE [20]	GNR	0.7155	0.6450	0.4684	1.7993	0.6131	0.5989	0.4517	0.7679	0.5764	0.3959	0.3792	0.2647	0.8225	0.5855
	M3 [21]	GNR	0.5610	0.4634	0.3177	2.0733	0.7129	**0.7192**	**0.5496**	0.6819	0.4281	0.6623	0.6120	0.4574	0.5937	0.4323
	MW-PSNR [23]	SRR	0.8012	0.8099	0.6063	1.5568	0.5722	0.5638	0.3891	0.7971	0.6011	0.6843	0.6772	0.4897	0.6532	0.4611
	MP-PSNR [24]	SRR	0.8169	0.8231	0.6206	1.5007	0.5520	0.5353	0.3675	0.8105	0.6198	0.7221	0.7271	0.5302	0.6196	0.4244
	LOGS [8]	SRR	0.7263	0.6607	0.4826	1.7885	0.6335	0.6081	0.4512	0.7520	0.5716	0.6298	0.6265	0.4439	0.6957	0.5097
	SET [5]	SNR	**0.9179**	0.9171	0.7473	**1.0279**	**0.7533**	0.7098	0.5386	**0.5802**	**0.3784**	**0.8152**	**0.8027**	**0.6277**	**0.4884**	**0.3163**
	Ref. [30]	SNR	0.4906	0.4763	0.3298	2.2670	0.5043	0.3433	0.2253	0.8393	0.6567	0.3570	0.3674	0.2500	0.8366	0.6290
	NIQSV [26]	SNR	0.6780	0.6216	0.4392	1.9123	0.5209	0.4546	0.3299	0.8296	0.6499	0.3765	0.3585	0.2340	0.8298	0.6244
Pre-DIBR	Ref. [33]	SFR C+D	0.9064	**0.9175**	**0.7481**	1.0993	–	–	–	–	–	–	–	–	–	–
	CODIF	SNR C+D	**0.9352**	**0.9290**	**0.7783**	**0.9262**	**0.7851**	**0.7222**	**0.5649**	**0.5711**	**0.3434**	**0.8249**	**0.7950**	**0.6095**	**0.4709**	**0.3009**

Fig. 7: Five color-depth image pairs with different distortions, the ground truth scores (MOS) and the objective scores predicted by CODIF. Top row shows color images, and bottom row shows the associated depth images.

training, and the other 20% are used for test. To avoid bias, we repeat the above process by 1,000 times and report the median values. Table I lists the experimental results on MCL-3D and IST databases, where we highlight the top two metrics in boldface. In the table, 'Post-DIBR' indicates that the metric uses DIBR-synthesized images in the quality evaluation, while 'Pre-DIBR' indicates that the metric uses the input color and depth images to predict the quality of view synthesis. In addition, 'GNR' denotes the general-purpose no-reference IQA metric, 'SFR/SRR/SNR' denotes the full-reference/reduced-reference/no-reference view synthesis quality metric. 'C+D' indicates that the metric uses input color-depth image pair for quality prediction. Kindly note that the metric in [33] needs the undistorted color and depth images during the quality evaluation, which are not provided in the IST database. So these results are not available.

It is known from Table I that the CODIF metric delivers the best performance in MCL-3D, in terms of both prediction accuracy and monotonicity. The pre-DIBR metric [33] delivers the second best prediction monotonicity, while the post-DIBR metric SET [5] produces the second best prediction accuracy. For VSRS on IST database, CODIF also delivers the best performance. The SET metric [5] again performs the second best in prediction accuracy. The general-purpose natural image quality metric M3 [21] obtains the second best prediction monotonicity. For VSIM on IST database, the proposed metric delivers the best accuracy criterion as well as the second best monotonicity criterion (only slightly worse than SET [5]). From these results, it is evident that the proposed metric achieves the best overall performance in predicting the quality of synthesized views. In addition, as a pre-DIBR metric, CODIF also outperforms the post-DIBR metrics.

To know the performance of the proposed metric more intuitively, Fig. 7 shows five color/depth image pairs with different scenes and distortions, as well as the ground truth MOS values of the synthesized image and the predicted scores by the proposed CODIF metric. From Figs. 7(a)-7(e), it is clearly observed that, with their MOS values decreasing, the predicted scores of our proposed metric also decrease accordingly. Moreover, the predicted scores are very close to the MOS values. These results further demonstrate that the predicted scores are consistent with the human ratings.

C. Analysis of Statistical Significance

To estimate the statistical significance of the proposed metric in contrast to the compared metrics, we perform a statistical

TABLE II: Summary of statistical performances between the proposed metric and state-of-the-art quality metrics. (a) VSRS on MCL-3D Database, (b) VSRS on IST Database, and (c) VSIM on IST Database

Metric	RMSE			RMSE*		PLCC		
	(a)	(b)	(c)	(b)	(c)	(a)	(b)	(c)
NIQE [19]	1	1	1	1	1	1	0	0
BRISQUE [18]	1	0	1	0	1	1	0	0
IL-NIQE [20]	1	1	1	1	1	1	0	1
M3 [21]	1	0	1	1	1	1	0	0
MW-PSNR [23]	1	1	1	1	1	1	0	0
MP-PSNR [24]	1	1	1	1	1	1	0	0
LOGS [8]	1	1	1	1	1	1	0	0
SET [5]	1	0	0	0	0	0	0	0
Ref. [30]	1	1	1	1	1	1	0	1
NIQSV [26]	1	1	1	1	1	1	0	1
Ref. [33]	1	–	–	–	–	0	–	–

TABLE III: Performances of the proposed metric on six kinds of distortions in MCL-3D.

Distortion	PLCC	SRCC	KRCC	RMSE
White noise	0.9618	0.9462	0.8248	0.6749
JPEG	0.9551	0.8806	0.7293	0.6355
JPEG2000	0.9811	0.9376	0.8161	0.4680
Gaussian blur	0.9810	0.9609	0.8572	0.5178
Downsampling blur	0.9822	0.9670	0.8739	0.5280
Transmission error	0.9029	0.8794	0.7295	0.8170

TABLE IV: Feature ablation study on MCL-3D database.

Feature	PLCC	SRCC	KRCC	RMSE
IRs-TMC	0.9059	0.8903	0.7166	1.1393
NRs-MSCN	0.8927	0.8877	0.7217	1.1632
Overall	0.9352	0.9290	0.7783	0.9262

analysis using the F test [63], which is commonly used to determine whether a metric performs statistically better/worse than another one. In implementation, the F measure is first computed using the RMSE values of a metric X and CODIF as:

$$F = \left(\frac{\mathrm{RMSE_X}}{\mathrm{RMSE_{CODIF}}} \right)^2. \tag{14}$$

Then, a threshold $F_{critical}$ is calculated based on the image number in each database with confidence level 95%. If $F > F_{critical}$, CODIF performs significantly better than metric X. If $F < 1/F_{critical}$, CODIF performs significantly worse than metric X. Otherwise, the two metrics have competitive performance. In this work, the thresholds $F_{critical}$ for VSRS on MCL-3D database,VSRS on IST database, and VSIM on IST database are 1.1381, 1.5343 and 1.3519, respectively. Similarly, we also calculate the F measure using RMSE* values of a metric X and CODIF. Fig. 8 shows the bar plots of the F-statistics of the compared metrics against the proposed metric. The corresponding statistical analysis results are listed in Table II. In the table, "1/0" indicates that the proposed metric performs significantly better/competitive than the compared metric.

It is observed from Fig. 8 that the F values of all compared metrics are bigger than 1, indicating that CODIF predictions are more accurate than the other metrics. From Table II, we know that the proposed CODIF significantly outperforms all the compared metrics in MCL-3D database. In the IST database, only one of the ten compared metrics can obtain statistically competitive performance with CODIF on both VSRS and VSIM. These results also demonstrate that the CODIF outperforms the existing metrics by a large margin.

Moreover, we conduct a statistical significance test for the difference between PLCC values of the proposed CODIF and the compared metrics, where more experimental details can be found in ITU-T P.1401 [62]. The statistical analysis results are summarized in Table II. From the results, we can see that the proposed CODIF achieves a very promising performance. Besides, among the existing quality metrics, only SET [5] have competitive statistical performance with the proposed metric on both databases.

D. Performance on Different Distortions

The MCL-3D database contains six kinds of distortions for color/depth images. To further investigate the performance of the proposed CODIF metric on individual distortion types, we further test it on the six distortions respectively. In Table III, we summarize the corresponding experimental results.

From the results, it is easily observed that CODIF also delivers very satisfactory results on individual distortion types. Especially for JPEG2000, Gaussian blur and Downsampling blur, the prediction accuracy values are all higher than 0.98. These results demonstrate that CODIF can predict the view synthesis quality when the color and/or depth images are subject to diversified distortions, which is a property highly needed in real applications.

E. Ablation study

In the proposed metric, two sets of features are extracted from interactional regions and natural regions of the fused color-depth image for building the quality prediction model. To further know the relative importance of the two groups of features, we further perform an ablation experiment based on the MCL-3D database. Specifically, the IRs-TMC features and NRs-MSCN features are separately fed into the SVR model for training and test with the same setting as before. Table IV summarizes the corresponding experimental results.

It is known from Table IV that the two groups of features can both achieve very encouraging results when they are used separately. Even a single group of feature is used, the performances are better than most state-of-the-art post-DIBR metrics. In addition, when the two sets of features

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3024882, IEEE Transactions on Circuits and Systems for Video Technology

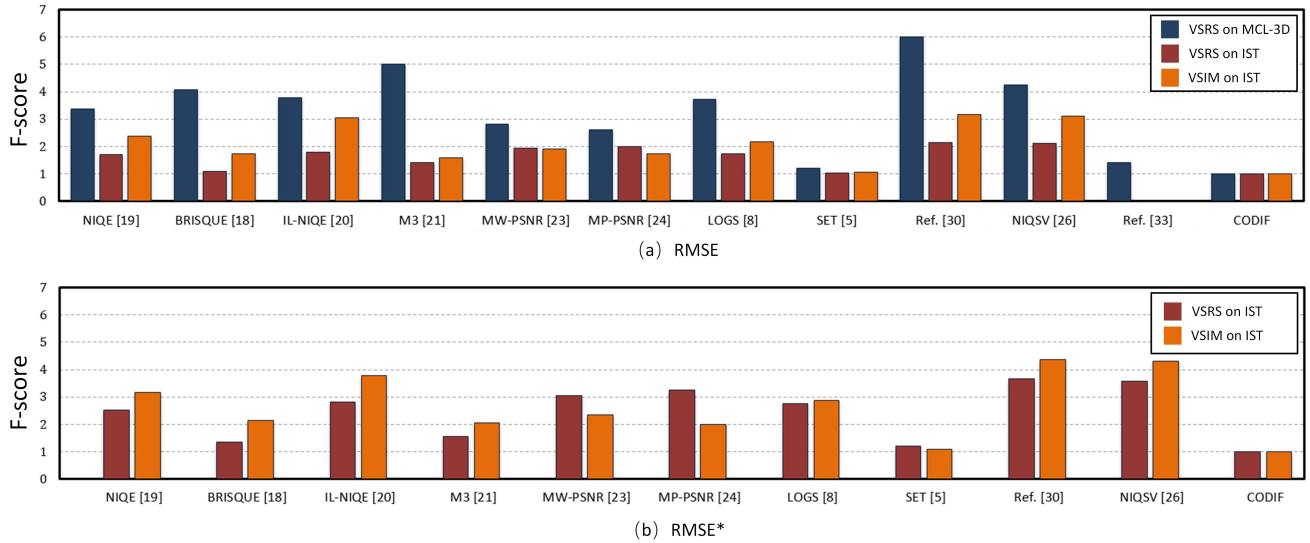

Fig. 8: Plot of F scores of CODIF vs. the state-of-the-art metrics.

TABLE V: Performances of the proposed metric using different regression models.

Regression Model	PLCC	SRCC	KRCC	RMSE
BP	0.9238	0.9146	0.7553	0.9325
RF	0.9109	0.9211	0.7697	0.9408
SVR (Proposed)	**0.9352**	**0.9290**	**0.7783**	**0.9262**

are combined to train the prediction model, the performance further improves significantly. This also demonstrates the necessity and reasonability of integrating the features from both interactional regions and natural regions for building an advanced view synthesis quality metric.

F. Evaluation of Quality Regression Model

In this part, we first compare the performance of the proposed CODIF using SVR-based quality regression with those of using the back propagation (BP) neural network [64] and random forest (RF) regression model [5], where 80% image pairs are adopted for model training, and 20% image pairs are used for the test on MCL-3D database. In implementation, the training-test process is repeated 1000 times, and the median values are summarized in Table V. It can be seen that the proposed CODIF using SVR achieves better performance than the metric using BP or RF. Therefore, we leverage the SVR to train the quality regression model in this work.

Further, we also investigate the performance of the proposed metric when different percentages of image pairs are used for model training. Table VI lists the experimental results on MCL-3D database. From Table VI, it is clearly observed that the proposed metric is not that sensitive to the various ratios of training set. Even with 40% image pairs to train the model, the performance of the proposed CODIF is still better than the majority of the existing quality metrics, and PLCC value exceeds 0.85. These results demonstrate that the proposed metric is not very dependent on the number of training image pairs, which is crucial for real-world applications.

TABLE VI: Performance of the proposed metric when different percentages of image pairs are used for model training and test.

Training-test	PLCC	SRCC	KRCC	RMSE
80%-20%	0.9352	0.9290	0.7783	0.9262
70%-30%	0.9162	0.9108	0.7535	1.0351
60%-40%	0.8965	0.8913	0.7271	1.1469
50%-50%	0.8762	0.8760	0.7025	1.2497
40%-60%	0.8579	0.8555	0.6775	1.3356

G. Generalization Ability

To investigate the generalization ability of the proposed metric, we further conduct a cross-database test. Specifically, we train the proposed metric on MCL-3D and IST databases, and then we test the metric performance in a recently released view synthesis quality database IETR [65]. The IETR database is intended for distortions solely introduced by DIBR algorithms, so undistorted color/depth images are provided. In other words, distortions in the synthesized images are only caused by the imperfect DIBR operation. Table VII summarizes the experiments results of the proposed CODIF model and the existing post-DIBR quality models, which are based on the VSRS synthesis method. It should be noted that SET [5] is a learning-based model while others are learning-free. For the learning-free models, we test their performance on the IETR database directly.

It is seen from Table VII that CODIF achieves the top two prediction monotonicity (SRCC and KRCC) when trained on MCL-3D and IST databases. For the prediction accuracy (PLCC), the proposed metric achieves the second best result

TABLE VII: Generalization performances of view synthesis quality models in IETR database. Experiments are conducted for the VSRS synthesis method.

Model	PLCC	SRCC	KRCC	RMSE
MW-PSNR [23]	0.4361	0.2632	0.1789	0.1478
MP-PSNR [24]	0.4203	0.3729	0.2737	0.1490
LOGS [8]	**0.7890**	0.4633	0.3595	**0.1039**
Ref. [30]	0.3802	0.4256	0.3263	0.1519
NIQSV [26]	0.4594	0.5293	0.3789	0.1459
SET [5] (Training on MCL-3D)	0.3955	0.3910	0.2421	0.1509
SET [5] (Training on IST)	0.4655	0.3368	0.2316	0.1454
CODIF (Training on MCL-3D)	**0.7260**	**0.6904**	**0.5033**	**0.1163**
CODIF (Training on IST)	0.6044	**0.5562**	**0.4248**	0.1348

when trained on MCL-3D, which is slightly worse than LOGS [8]. It is worth noting that LOGS [8] is specifically designed for measuring the rendering distortions in DIBR, so it is not surprising that it delivers the best prediction accuracy in IETR database. However, LOGS does not perform very well under diversified distortions, which can be seen from Table I. As a pre-DIBR metric, CODIF outperforms most of the post-DIBR metrics in terms of generalization ability, which further confirms the advantages of the proposed metric. Moreover, it can be observed that CODIF achieves better performance when the quality model is trained on the MCL-3D database than IST database. The main reason is that the images in the IST database only contain compression distortion, while the images in the MCL-3D database are generated by six different kinds of distortions. On the other hand, the MCL-3D database can provide more training images than IST database.

V. Conclusion

In DIBR-based virtual view synthesis, distortions in the input color/depth signals lead to the degraded synthesized view. Most of the current quality models operate on the synthesized images, after performing the computationally expensive DIBR process. To tackle the problem, we have presented a NR quality prediction model for view synthesis using the input color and depth images directly without performing the DIBR process. Based on the proposed Wavelet-based color-depth image fusion approach, the interactions between color and depth images during the DIBR process can be simulated, so that the quality of the synthesized image can be predicted using the fused color-depth image. We have also proposed to predict the overall quality from the interactional regions and natural regions simultaneously. We have conducted extensive experiments and compared the performance of the proposed metric with the state-of-the-arts. The results have demonstrated that, in spite of a pre-DIBR quality prediction metric, the proposed model even outperforms the current post-DIBR quality models. As an application, the proposed view synthesis quality prediction metric is expected to benefit joint color/depth coding and bit allocation optimization for improving view synthesis quality.

References

[1] Y. Yuan, G. Cheung, P. Le Callet, P. Frossard, and H. V. Zhao, "Object shape approximation and contour adaptive depth image coding for virtual view synthesis," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 28, no. 12, pp. 3437-3451, Dec. 2018.

[2] A. I. Purica, E. G. Mora, B. Pesquet-Popescu, M. Cagnazzo, and B. Ionescu, "Multiview plus depth video coding with temporal prediction view synthesis," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 26, no. 2, pp. 360-374, Feb. 2016.

[3] S. Li, C. Zhu, and M. Sun, "Hole filling with multiple reference views in DIBR view synthesis," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 1948-1959, Aug. 2018.

[4] H. Hsu, C. Chiang, and S. Lai, "Spatio-temporally consistent view synthesis from video-plus-depth data with global optimization," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 24, no. 1, pp. 74-84, Jan. 2014.

[5] Y. Zhou, L. Li, S. Wang, J. Wu, Y. Fang, and X. Gao, "No-reference quality assessment for view synthesis using DoG-based edge statistics and texture naturalness," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4566-4579, Sep. 2019.

[6] W. Li, J. Zhou, B. Li, and M. I. Sezan, "Virtual view specification and synthesis for free viewpoint television," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 19, no. 4, pp. 533-546, Apr. 2009.

[7] F. Zou, D. Tian, A. Vetro, H. Sun, O. C. Au, and S. Shimizu, "View synthesis prediction in the 3-D video coding extensions of AVC and HEVC," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 24, no. 10, pp. 1696-1708, Oct. 2014.

[8] L. Li, Y. Zhou, K. Gu, W. Lin, and S. Wang, "Quality assessment of DIBR-synthesized images by measuring local geometric distortions and global sharpness," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 914-926, Apr. 2018.

[9] L. Li, W. Lin, X. Wang, G. Yang, K. Bahrami, and A. C. Kot, "No-reference image blur assessment based on discrete orthogonal moments," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 39-50, Jan. 2016.

[10] L. Po, M. Liu, W. Y. F. Yuen, Y. Li, X. Xu, C. Zhou, P. H. W. Wong, K. W. Lau, and H. Luk, "A novel patch variance biased convolutional neural network for no-reference image quality assessment," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 29, no. 4, pp. 1223-1229, Apr. 2019.

[11] S. Wang, K. Gu, X. Zhang, W. Lin, S. Ma, and W. Gao, "Reduced-reference quality assessment of screen content images," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 28, no. 1, pp. 1-14, Jan. 2018.

[12] S. Ryu and K. Sohn, "No-reference quality assessment for stereoscopic images based on binocular quality perception," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 24, no. 4, pp. 591-602, Apr. 2014.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[14] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378-2386, Aug. 2011.

[15] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430-444, Feb. 2006.

[16] F. Qi, D. Zhao, and W. Gao, "Reduced reference stereoscopic image quality assessment based on binocular perceptual information," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2338-2344, Dec. 2015.

[17] W. Zhu, G. Zhai, X. Min, M. Hu, J. Liu, G. Guo, and X. Yang, "Multi-channel decomposition in tandem with free-energy principle for reduced-reference image quality assessment," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2334-2346, Sep. 2019.

[18] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695-4708, Dec. 2012.

[19] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209-212, Mar. 2013.

[20] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579-2591, Aug. 2015.

[21] 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 Trans. Image Process.*, vol. 23, no. 11, pp. 4850-4862, Nov. 2014.

[22] F. Battisti, E. Bosc, M. Carli, P. Le Callet, and S. Perugia, "Objective image quality assessment of 3D synthesized views," *Signal Process. Image Commun.*, vol. 30, pp. 78-88, Jan. 2015.

[23] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "DIBR synthesized image quality assessment based on morphological wavelets," in *Proc. IEEE Int. Workshop Quality Multimedia Exper. (QoMEX)*, May 2015, pp. 1-6.

[24] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "DIBR synthesized image quality assessment based on morphological pyramids," in *Proc. 3DTV-Conf., True Vis.-Capture, Transmiss. Display 3D Video (3DTV-CON)*, Jul. 2015, pp. 1-4.

[25] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "Multi-scale synthesized view assessment based on morphological pyramids," *J. Elect. Eng.*, vol. 67, no. 1, pp. 3-11, 2016.

[26] S. Tian, L. Zhang, L. Morin, and O. Deforges, "NIQSV: A no reference image quality assessment metric for 3D synthesized views," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2017, pp. 1248-1252.

[27] S. Tian, L. Zhang, L. Morin, and O. Dforges, "NIQSV+: A no-reference synthesized view quality assessment metric," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1652-1664, Apr. 2018.

[28] K. Gu, V. Jakhetiya, J. Qiao, X. Li, W. Lin, and D. Thalmann, "Model-based referenceless quality metric of 3D synthesized images using local image description," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 394-405, Jan. 2018.

[29] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "Fast blind quality assessment of DIBR-synthesized video based on high-high wavelet sub-band," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5524-5536, Nov. 2019.

[30] V. Jakhetiya, K. Gu, T. Singhal, S. C. Guntuku, Z. Xia, and W. Lin, "A highly efficient blind image quality assessment metric of 3-D synthesized images using outlier detection," *IEEE Trans. Ind. Inform.*, vol. 15, no. 7, pp. 4120-4128, Jul. 2019.

[31] K. Gu, J. Qiao, S. Lee, H. Liu, W. Lin, and P. Le Callet, "Multiscale natural scene statistical analysis for no-reference quality evaluation of DIBR-synthesized views," *IEEE Trans. Broadcast.*, pp. 1-13, May 2019.

[32] G. Wang, Z. Wang, K. Gu, L. Li, Z. Xia, and L. Wu, "Blind quality metric of DIBR-synthesized images in the discrete wavelet transform domain," *IEEE Trans. Image Process.*, vol. 29, pp. 1802-1814, 2020.

[33] J. Wang, S. Wang, K. Zeng, and Z. Wang, "Quality assessment of multi-view-plus-depth images," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 85-90.

[34] F. Shao, Q. Yuan, W. Lin, and G. Jiang, "No-reference view synthesis quality prediction for 3-D videos based on color-depth interactions," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 659-674, Mar. 2018.

[35] X. Liu, Y. Zhang, S. Hu, S. Kwong, C. J. Kuo, and Q. Peng, "Subjective and objective video quality assessment of 3D synthesized views with texture/depth compression distortion," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4847-4861, Dec. 2015.

[36] G. Wang, Z. Wang, K. Gu, and Z. Xia, "Blind quality assessment for 3D-synthesized images by measuring geometric distortions and image complexity," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 4040-4044.

[37] M. B. Angeles, F. Yilmaz, G. Boato, M. Carli, E. Dumic, M. Gelautz, M. Gelautz, C. Hewage, D. Kukolj, P. L. Callet, A. Liotta, C. Pasquini, A. P. Banos, C. Politis, D. Sandic, M. T. Vega, and V. Zlokolica, "Quality of experience and quality of service metrics for 3d content," *3D Visual Content Creation, Coding and Delivery*, pp. 267-297, July 2018.

[38] E. Cemiloglu and G. N. Yilmaz, "Blind video quality assessment via spatiotemporal statistical analysis of adaptive cube size 3D-DCT coefficients," *IET Image Process.*, vol. 14, no. 5, pp. 845-852, Apr. 2020.

[39] L. Wang, S. Xiang, G. Meng, H. Wu, and C. Pan, "Edge-directed single-image super-resolution via adaptive gradient magnitude self-interpolation," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 23, no. 8, pp. 1289-1299, Aug. 2013.

[40] G. N. Yilmaz, "A no reference depth perception assessment metric for 3D video," *Multimedia Tools Appl.*, vol. 74, pp. 6937-6950, Mar. 2014.

[41] K. Gu, J. Zhou, J. Qiao, G. Zhai, W. Lin, and A. C. Bovik, "No-reference quality assessment of screen content pictures," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 4005-4018, Aug. 2017.

[42] Q. Jiang, F. Shao, W. Lin, and G. Jiang, "BLIQUE-TMI: Blind quality evaluator for tone-mapped images based on local and global feature analyses," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 29, no. 2, pp. 323-335, Feb. 2019.

[43] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, "Assessing visual quality of omnidirectional videos," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 29, no. 12, pp. 3516-3530, Dec. 2019.

[44] A. Ellmauthaler, C. L. Pagliari, and E. A. B. da Silva, "Multiscale image fusion using the undecimated wavelet transform with spectral factorization and nonorthogonal filter banks," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 1005-1017, Mar. 2013.

[45] E. Daniel, "Optimum wavelet-based homomorphic medical image fusion using hybrid geneticcgrey wolf optimization algorithm," *IEEE Sensors J.*, vol. 18, no. 16, pp. 6804-6811, Aug. 2018.

[46] L. Ye and Z. Hou, "Memory efficient multilevel discrete wavelet transform schemes for JPEG2000," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 25, no. 11, pp. 1773-1785, Nov. 2015.

[47] X. Wang, "Moving window-based double haar wavelet transform for image processing," *IEEE Trans. Image Process.*, vol. 15, no. 9, pp. 2771-2779, Sep. 2006.

[48] G. Ding, Y. Guo, K. Chen, C. Chu, J. Han, and Q. Dai, "DECODE: Deep confidence network for robust image classification," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3752-3765, Aug. 2019.

[49] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. on Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395-1403.

[50] P. Arbelez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898-916, May 2011.

[51] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500-1512, Apr. 2012.

[52] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695-4708, Dec. 2012.

[53] Y. Fu, H. Zeng, L. Ma, Z. Ni, J. Zhu, and K. Ma, "Screen content image quality assessment using multi-scale difference of gaussian," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 28, no. 9, pp. 2428-2432, Sep. 2018.

[54] G. Ding, W. Chen, S. Zhao, J. Han, and Q. Liu, "Real-time scalable visual tracking via quadrangle kernelized correlation filters," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 140-150, Jan. 2018.

[55] B. Konuk, E. Zerman, G. N. Yilmaz, and G. B. Akar, "A spatiotemporal no-reference video quality assessment model," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2013, pp. 54-58.

[56] Y. Guo, G. Ding, and J. Han, "Robust quantization for general similarity search," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 949-963, Feb. 2018.

[57] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91-110, 2004.

[58] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1-27, Jan. 2011.

[59] R. Song, H. Ko, and C. C. Kuo, "MCL-3D: A database for stereoscopic image quality assessment using 2D-image-plus-depth source," *J. Inf. Sci. Eng.*, vol. 31, no. 5, pp. 1593-1611, 2015.

[60] F. Rodrigues, J. M. Ascenso, A. Rodrigues, and P. Queluz, "Blind quality assessment of 3D synthesized views based on hybrid feature classes," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1737-1749, July 2019.

[61] M. S. Farid, M. Lucenteforte, and M. Grangetto, "Depth image based rendering with inverse mapping," in *Proc. Int. Conf. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2013, pp. 135-140.

[62] Recommendation ITU-T P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," [Online]. Available: https://www.itu.int/rec/T-RECP.1401-202001-I/env, accessed: Jan. 2020.

[63] L. Li, D. Wu, J. Wu, H. Li, W. Lin, and A. C. Kot, "Image sharpness assessment by sparse representation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1085-1097, Jun. 2016.

[64] 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 Process. Image Commun.*, vol. 40, pp. 1-15, Jan. 2016.

[65] S. Tian, L. Zhang, L. Morin, and O. Deforges, "A benchmark of DIBR synthesized view quality assessment metrics on a new database for immersive media applications," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1235-1247, May 2019.

Leida Li (M'14) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2004 and 2009, respectively. In 2008, he was a Research Assistant with the Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan. From 2014 to 2015, he was a Visiting Research Fellow with the Rapid-Rich Object Search Laboratory, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he was a Senior Research Fellow from 2016 to 2017. He is currently a Professor with the Guangzhou Institute of Technology, Xidian University, China. His research interests include multimedia quality assessment, affective computing, information hiding, and image forensics. He has served as an SPC for IJCAI 2019-2020, the Session Chair for ICMR in 2019 and PCM in 2015, and the TPC for AAAI in 2019, ACM MM 2019-2020, ACM MM-Asia in 2019, ACII in 2019, and PCM in 2016. He is currently an Associate Editor of the *Journal of Visual Communication and Image Representation* and the *EURASIP Journal on Image and Video Processing*.

Yipo Huang received the B.S. degree from Zhengzhou University, Zhengzhou, China, in 2017. He is currently pursuing the M.S. degree with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China. His research interests include multimedia quality assessment and perceptual image processing.

Jinjian Wu received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2008 and 2013, respectively. From 2011 to 2013, he was a Research Assistant with Nanyang Technological University, Singapore, where he was a Postdoctoral Research Fellow from 2013 to 2014. From 2015 to 2019, he was an Associate Professor with Xidian University, where he has been a Professor since 2019. His research interests include visual perceptual modeling, biomimetic imaging, quality evaluation, and object detection. Prof. Wu received the Best Student Paper Award at ISCAS 2013. He has served as an Associate Editor for the *Journal of Circuits, Systems and Signal Processing*, the Special Section Chair for IEEE Visual Communications and Image Processing in 2017, and the Section Chair/Organizer/TPC Member for ICME 2014-2015, PCM 2015-2016, ICIP 2015, VCIP 2018, and AAAI 2019.

Ke Gu (M'19) received the B.S. and Ph.D. degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009 and 2015, respectively. He is a Professor with 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), the Best Student Paper Award at the IEEE International Conference on Multimedia and Expo (ICME) in 2016, and the Excellent Ph.D. Thesis Award from the Chinese Institute of Electronics in 2016. He was the Leading Special Session Organizer in the VCIP 2016 and the ICIP 2017, and serves as a Guest Editor for Digital Signal Processing (DSP). He is currently an Associate Editor for IEEE ACCESS and IET Image Processing (IET-IPR), and an Area Editor for Signal Processing Image Communication (SPIC). He is a Reviewer for 20 top SCI journals.

Yuming Fang (M'13-SM'17) received the B.E. degree from Sichuan University, Chengdu, China, the M.S. degree from the Beijing University of Technology, Beijing, China, and the Ph.D. degree from Nanyang Technological University, Singapore. He is currently a Professor with the School of Information Management, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include visual attention modeling, visual quality assessment, computer vision, and 3D image/video processing. He serves as an Associate Editor for IEEE ACCESS. He serves on the Editorial Board of Signal Processing: Image Communication.