IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 15, 2022, accepted May 24, 2022, date of publication June 2, 2022, date of current version June 8, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3179693

Unifying Structural and Semantic Similarities for Quality Assessment of DIBR-Synthesized Views

SAEED MAHMOUDPOUR, (Member, IEEE), AND PETER SCHELKENS, (Member, IEEE)
Department of Electronics and Informatics, Vrije Universiteit Brussel, 1050 Brussels, Belgium
IMEC, 3001 Leuven, Belgium

Corresponding author: Saeed Mahmoudpour (saeed.mahmoudpour@vub.be)

This work was supported by the Research Foundation—Flanders (FWO) under Grant G0B3521N.

ABSTRACT Multi-view 3D content is subject to distortions during the process of depth image-based rendering (DIBR). Studies have shown the unreliable performance of the well-established image quality assessment (IQA) models for evaluation of DIBR-synthesized views which surge the need for more effective IQA methods. Existing objective methods generally rely on the pixel-wise correspondences between the reference and distorted images, while view synthesis can introduce pixel shifts. DIBR distortions such as stretching and local hole-filling errors have different visual impacts from conventional distortions, challenging the existing IQA models. Here, we developed a Full-Reference (FR) objective IQA metric for synthesized views that significantly outperforms 2D IQA and the state-of-the-art DIBR IQA approaches. While the pixel misalignment between the reference and synthesized views is a big challenge for quality assessment, we deployed a Convolutional Neural Network (CNN) model to acquire a feature representation that inherently offers resilience to the imperceptible pixel shift between the compared images. Therefore, our model does not need accurate shift compensation. We deployed a set of quality-aware CNN features representing high-order statistics, to measure the structural similarity which is combined with a semantic similarity measure for accurate quality assessment. Moreover, prediction accuracy is improved by incorporating a visual saliency model acquired using the activations of the higher CNN layers. Experimental results indicate a significant performance gain (14.6% in terms of Spearman's rank-order correlation) compared to the top existing IQA model. The source code of the proposed metric is available at: https://gitlab.com/saeedmp/sequss.

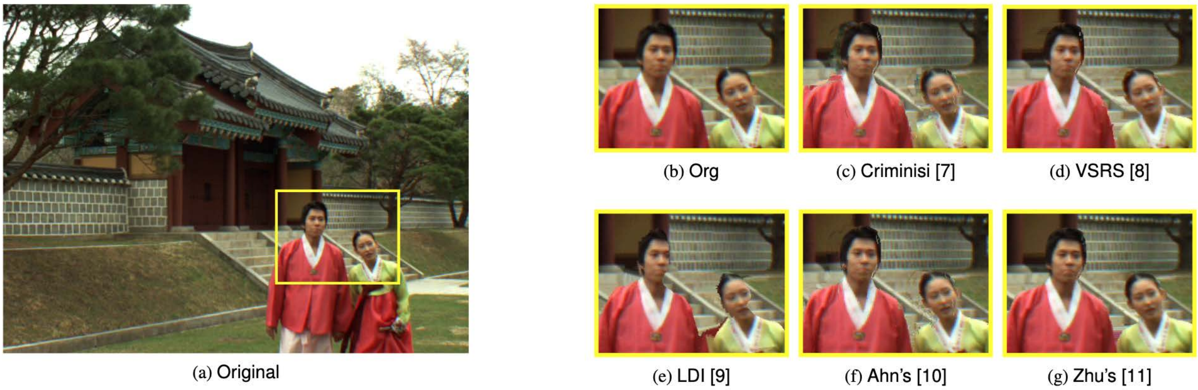
INDEX TERMS Deep neural networks, depth image-based rendering, image semantics, saliency map, visual quality assessment.

I. INTRODUCTION

With the rapid advances in virtual reality and 3D applications, new multimedia modalities such as free-viewpoint video (FVV) [1], light fields, point clouds, and holography [2], [3] have attracted significant attention. These emerging multimedia formats promise to enable immersive experience to end-users by delivering a richer visualization with full-parallax properties. Multi-view representation requires handling a tremendous volume of data captured from different viewpoints, therefore, effective data representation, storage, and transmission methods are key factors to promote the application of immersive multimedia. Multi-View texture plus Depth (MVD) [4], [5] is a widespread immersive media format that aims to represent a full 3D scene using a subset of texture views accompanied with the corresponding disparity information. A technique called Depth Image-Based Rendering (DIBR) [6] can be deployed to synthesize the missing virtual views using the texture and depth information captured from the adjacent camera locations.

Despite the remarkable advantages of the DIBR-based approaches, synthesized views often suffer from multiple distortion types caused by imperfect synthesis, occlusion, and depth data errors. Therefore, reliable Image Quality Assessment (IQA) methods are essential to evaluate and monitor the quality of the reconstructed views. DIBR distortions that occur in disoccluded regions have different visual effects from the conventional blur, noise, and blocking artifacts. Moreover, view synthesis errors can introduce geometric distortion and stretching effects which lead to

The associate editor coordinating the review of this manuscript and approving it for publication was Marco Giannelli.


(a) Original

(b) Org

(c) Criminisi [7]

(d) VSRS [8]

(e) LDI [9]

(f) Ahn's [10]

(g) Zhu's [11]

FIGURE 1. (a) The original 'Lovebird1' image from the IETR data set. (b) cropped original, and the synthesized images obtained using five DIBR techniques with subjective scores (c) DMOS = 67.91 (d) DMOS = 23.94 (e) DMOS = 84.84 (f) DMOS = 51.94 (g) DMOS = 25.04.

misalignment between the reference and the synthesized view. These distinct characteristics of DIBR distortions challenge the well-established objective quality assessment algorithms (such as Structural SIMilarity (SSIM) [7] and Visual Information Fidelity (VIF) [8] indices). The 2D IQA models highly rely on pixel-wise alignment of the comparing images and they are principally designed to evaluate the visual impacts of the conventional distortions on 2D images. Hence, they fail on assessing the quality of DIBR-synthesized views, and specific IQA models are required for MVD content evaluation.

The existing objective quality assessment methods for synthesized views can be classified into two main categories of Full-Reference (FR) or No-Reference (NR), based on the availability of the reference image. NR methods aim to assess the quality of synthesized views independent of a reference image. Tian *et al.* [9] proposed an NR algorithm that deploys morphological operations to compute quality scores in luminance and chrominance components. Then, the individual quality scores are combined and a generated edge image is used to weight the pixel-wise quality values. The authors further expanded their work in [10] by adding black hole and stretching detection strategies to improve the quality prediction task. Gu *et al.* [11] deployed a multi-scale scheme to design two DIBR-specific Natural Scene Statistics (NSS) models based on self-similarity and structural consistency characteristics. Thereafter, the quality scores obtained from the two NSS models were combined to pool the final score. Jakhetiya *et al.* [12] developed a computationally efficient NR method that uses simple median filtering to detect geometric and structural distortions for quality assessment.

Whereas an accurate NR quality assessment of DIBR-synthesized views is very challenging due to the lack of a reference view, several FR methods have been developed for providing more reliable quality predictions. Sandic-Stankovic *et al.* [13] computed Peak Signal-to-Noise Ratio (PSNR) on multi-scale images obtained from Morphological Wavelet (MW) decomposition. The authors further improved the IQA accuracy in [14] by using Morphological Pyramids (MP) filters for image decomposition. In [15], reference and synthesized views were used to detect error-prone disoccluded regions. The size and strength of these regions were then considered to compute a quality score which is later combined with a global sharpness assessment score. More recently, some works showed the effectiveness of shift compensation to align reference and synthesized views for FR quality assessment. In [16], a two-step shift compensation approach was proposed based on the SURF and a multi-scale block matching, and finally, a quality score is obtained by computing pixel-wise mean squared error between the reference and shift-compensated distorted views. In our recent work [17], we proposed a quality index called SSPD that uses feature matching and superpixel difference to pool a quality score. The corresponding interest point features are compared in reference and synthesized views for local quality assessment. In addition, a global quality score is pooled by computing the gradient magnitude difference of superpixels in reference and shift-compensated synthesized images. The SSPD could outperform the competing approaches on both conventional and new DIBR data sets.

Current DIBR IQA models are generally designed to perform well on the conventional DIBR distortions. Data sets such as IRCCyN/IVC [18] include DIBR algorithms that only resemble old-fashioned DIBR distortions including blurring, black holes, and geometric distortions. However, recent DIBR techniques have been improved significantly and can better address the DIBR errors. The black hole errors are almost resolved, geometric distortions are better handled, and sophisticated inpainting methods are proposed to better compensate for the errors in the disoccluded regions. In 2019, a new public DIBR data set namely IETR [19] was released that covers the new DIBR techniques. Fig. 1 visualizes an original image and the synthesized versions obtained using different DIBR algorithms from the IETR data set. As the figure illustrates, local distortions on object boundaries, induced by different DIBR algorithms, present diverse visual impacts on image structures. Moreover, techniques such as LDI [20] can deform important visual cues

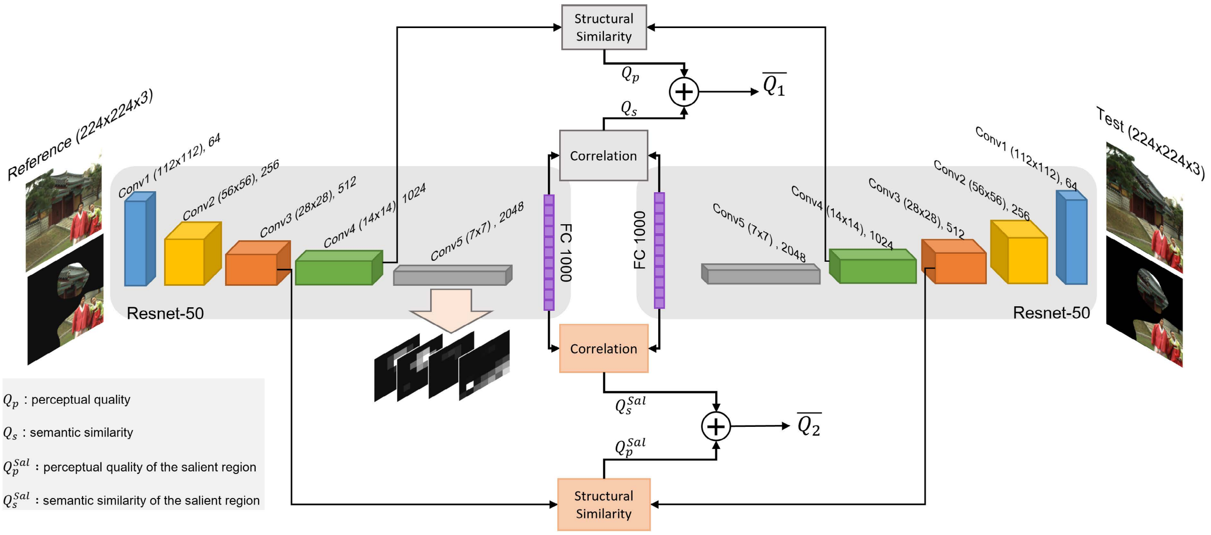


FIGURE 2. Framework of the proposed objective quality assessment metric. The perceptual quality and semantic similarity scores are computed and combined in each of the two channels. The final SEQUSS score is then computed as a weighted sum of the two channel scores ($\overline{Q_1}$ and $\overline{Q_2}$).

such as faces which affect image semantics. Performance evaluation on the IETR data set has revealed that DIBR IQA models fail to deliver high accuracy for new DIBR algorithms [19]. The highest performance is achieved using our SSPD model on the IETR data set with Spearman's Rank-order Correlation Coefficient (SRCC) of 0.685 which suggests a noticeable room for improvement [17]. In this paper, we proposed a new FR objective quality assessment metric based on SEmantic- and QUality-aware feature Similarity measures plus a Salient-region detection (SEQUSS). The proposed metric achieves a large performance gain over the existing IQA approaches.

The state-of-the-art DIBR IQA methods often follow a shift-compensation strategy as a pre-processing step to compensate for the misalignment between the reference and synthesized views; however, the proposed algorithm circumvents the need for shift compensation using deep features of a Convolutional Neural Network (CNN) model. Although the alignments of the images being compared can lead to more accurate quality assessment, the shift compensation process is not always flawless and often comes with warping errors that can influence the quality evaluation task. Using the pre-trained ResNet50 CNN [21], we transform images to a multi-scale representation with perceptual features that are more tolerable to shifts. Though a pair of reference and test images that are different in terms of their precise pixel locations might look rather similar for the Human Visual System (HVS) [22], they are often judged to be different when using FR objective quality assessment methods. The deep perceptual features obtained from the CNN model are better aligned with the perceptual preferences and show good tolerance to imperceptible pixel shifts. Here we developed two measures by extracting both quality- and semantic-aware features from the ResNet50 layers. Moreover, the feature activation maps from the last CNN block were used to highlight the visually-salient regions for more effective perceptual evaluation. The main contributions of the proposed method are summarized as follows:

- The proposed method utilizes perceptual features of deep CNN layers for quality assessment. These features better adhere to the HVS behavior and are less sensitive to small pixel-wise shifts between the comparing images. This allows more reliable quality assessment free of the error-prone shift-compensation methods. The proposed metric achieves a substantial accuracy gain over the state-of-the-art approaches.
- We proposed to incorporate the semantic content features in quality assessment since DIBR local distortions and stretching artifacts can influence image semantics.
- The visual attention processing behavior of HVS is considered in the design of our IQA model by generating visual saliency masks using the last CNN block. The saliency map can suitably highlight the regions of interest that are visually important for quality assessment.

The rest of the paper is organized as follows: Section II elaborates the proposed quality assessment algorithm. The experimental results are summarized and discussed in Section III. Finally, section IV concludes the paper.

II. PROPOSED OBJECTIVE QUALITY ASSESSMENT METHOD

Unlike traditional distortion types that rather uniformly affect the entire image, DIBR distortions include several local and global artifacts that alter the structural and semantic characteristics of the scene and degrades the overall visual Quality of Experience (QoE). This urges sophisticated models that can better comply with the complex properties of HVS. Data processing in the HVS follows a hierarchical mechanism in which the sensitivity to complex stimulus characteristics is increasing along the ventral visual pathway. Early visual processing areas encode low-level frequency components while higher visual areas are more sensitive to complex textures and semantic shapes [23], [24] [25]. CNN architectures – initially designed for computer vision tasks - also follow a hierarchical multi-scale data processing mechanism and they can roughly approximate the data processing of the HVS [26], [27]. Recent studies reveal the effectiveness of the pre-trained deep CNNs for the task of quality assessment [28].

Here we developed a new DIBR IQA metric using the features extracted from different layers of the ResNet-50 CNN. This CNN model is trained on more than a million images of the ImageNet data set [29] and consists of five main convolutional stages (L_1 to L_5) followed by a Fully Connected (FC) layer at the end of the network. The number of filters for the five CNN layers is L_1: 64, L_2: 256, L_3: 512, L_4: 1024, and L_5: 2048. Moving toward deeper convolutional layers, the size of CNN filters (feature maps) shrinks while the number of filters is increased.

Fig. 2 presents the framework of the proposed IQA method. Deep features of the ResNet-50 are deployed to effectively quantify both the structural and semantic degradations. In addition, we incorporated saliency maps – obtained using the features of the fifth CNN block – in quality prediction to account for the visual attention processing of HVS. The proposed method consists of two computational streams for quality pooling. The first stream uses intact reference and synthesized views as inputs to the network while the second stream applies saliency masks on input images as a pre-processing step to extract features only from perceptually-salient regions. For each computational stream, a perceptual and a semantic score are acquired using deep quality-aware and sematic-aware feature similarity measures, respectively. Finally, the scores from two channels are combined to obtain the final DIBR quality score.

A. QUALITY-AWARE FEATURE COMPARISON

The proposed quality assessment model is based on a non-linear transformation of the reference and synthesized images to a new over-complete feature space representation. We used the features of the ReLU layers available at the end of each of the five CNN blocks. Similar to data processing in the visual cortex, early CNN layers include smaller receptive fields and capture low-level features using a larger number of neurons in each feature map while higher CNN layers are more sensitive to high-order statistics and complex edge features. Fig. 3 presents some feature maps selected from the five CNN layers of the ResNet50 model. As shown in the figure, structural information has been encoded in five layers of the CNN model at different levels of frequency details.

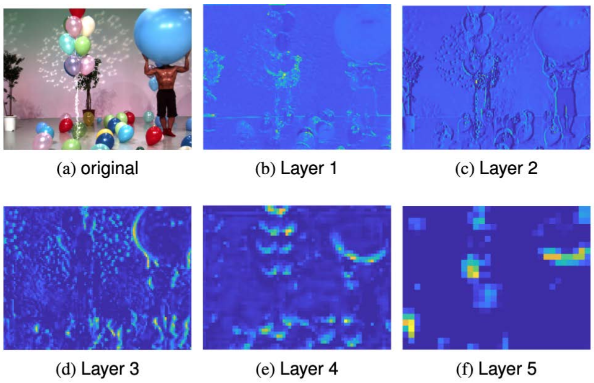

(a) original
(b) Layer 1
(c) Layer 2
(d) Layer 3
(e) Layer 4
(f) Layer 5

FIGURE 3. (a) Original 'balloons' image and the selected feature maps from the five CNN layers. feature map No.: (b) 50 (c) 250, (d) 470, (e) 180, (f) 1980.

Due to the misalignment between reference and synthesized views, it can be expected that early CNN layers deliver lower IQA performance when precise frequency components are compared; however, deeper levels might perform better since the comparison is performed in a higher level of visual appearance and the shift is better tolerated. This assumption was confirmed by examining all five CNN blocks for quality assessment in which the best performance was achieved using the fourth layer, thus, we used this layer for quality-aware feature extraction. Using the features in the fourth layer, we ensure high sensitivity to the structural degradation while preserving a good tolerance against spatial imperceptible misalignments. Please note that although quality assessment in higher CNN layers could suitably mitigate the impact of misalignment without the need for shift compensation, severe geometrical distortions can still affect the algorithm accuracy. However, such large displacements does not appear in the reconstruction of the new view synthesis algorithms. Let ψ_r and ψ_d are the resized $N \times 1$ feature vectors extracted from the feature maps of the reference and distorted views respectively, the structural similarity of the features in the lth layer of the ResNet50 is computed as:

$$Q_p^l = \frac{2\sigma^l_{\psi_r,\psi_d} + c_1}{(\sigma^l_{\psi_r})^2 + (\sigma^l_{\psi_d})^2 + c_1} \tag{1}$$

where σ_{ψ_r} and σ_{ψ_d} are the global variance of the features in reference and distorted views and σ_{ψ_r,ψ_d} denotes the global covariance of the features. The parameter c_1 is a small positive value to ensure numerical stability of the measurement.

B. SEMANTIC-AWARE FEATURE COMPARISON

Image semantic information describes the content appearance in the image and distortions can deviate the semantic characteristics. Researchers have shown that semantic image category has a noticeable impact on subjective quality ratings [30] and modifications in image semantics can impact the impression of the overall perceptual quality [31], [32]. Since DIBR impairments – induced by stretching and faulty reconstructions on shape borders – can degrade the semantic understanding of images, integrating quality- and semantic-aware features can help to quantify the impact of visual artifacts on content recognition and the final QoE. Here, we propose to use semantic features to further improve

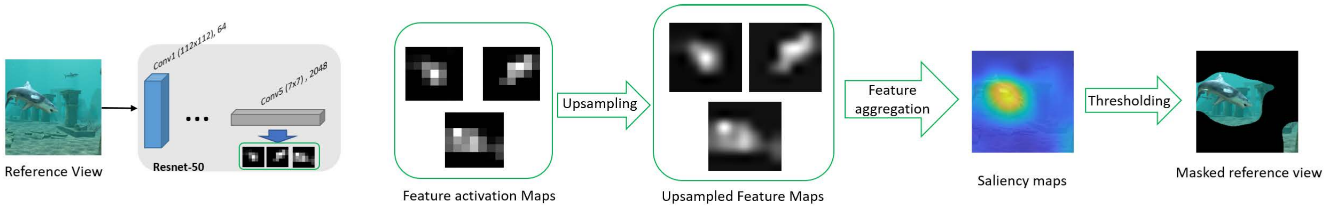


FIGURE 4. Illustration of the saliency map generation method and the input image masking.

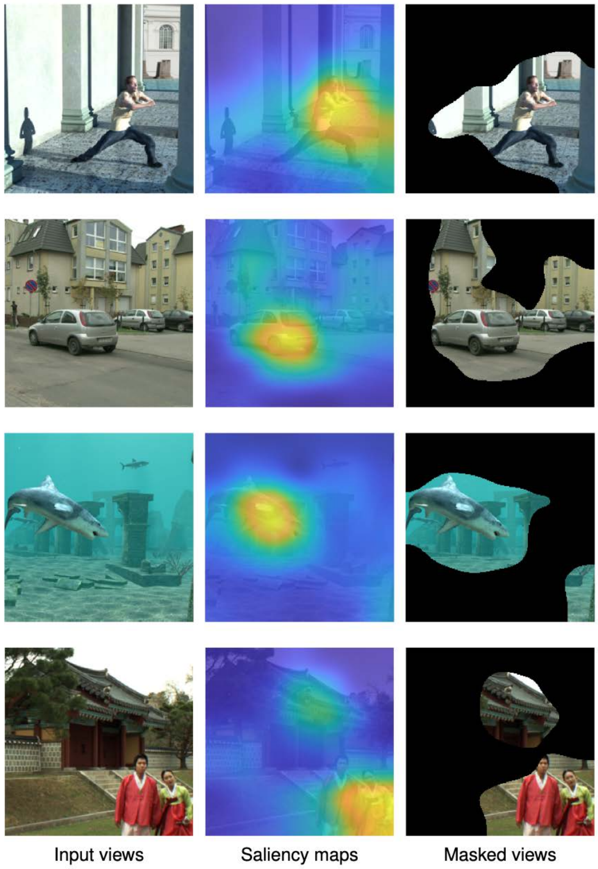


FIGURE 5. Saliency maps and the masked images of four reference images in the IETR data set.

the accuracy of DIBR quality assessment. We used the features of the FC layer for semantic comparison. The FC layer comes at the end of the network after the CNN blocks and before the softmax layer used for the classification task, thus, the FC layer is supposed to contain coarse features that represent scene semantics. Using the 1000 features of the FC layer in the ResNet-50, a semantic measure is acquired by computing the degree of consistency between two feature vectors. We computed the Spearman Rank-order Correlation Coefficient (SRCC) between the features of the FC layer in reference and synthesized views ψ_r^{FC} and ψ_d^{FC}:

$$Q_s = SRCC(\psi_r^{FC}, \psi_d^{FC}) = \frac{6\sum_{i=1}^{N}\delta_i^2}{N(N^2-1)} \tag{2}$$

where δ_i is the difference between the ranks of pairs of FC features in reference and synthesized views.

Using the quality and semantic measures, Q_p and Q_s, the overall quality score of the first computational stream is computed as:

$$\overline{Q_1} = \frac{Q_p + Q_s}{2} \tag{3}$$

The two measures in (3) are combined with equal weights. According to our experiments, a small increase in the weight of the quality measure Q_p can yield a minor gain, while larger weights (i.e., a significant decrease in the influence of the semantic measure) has a negative impact on performance. Thus, the equally weighted measures can provide high performance while avoiding extra parameter tuning.

C. SALIENT-REGION QUALITY MEASURE

The HVS is more attracted to visually-salient regions in the scene and thus the quality degradation in such regions of interest (RoI) is more critical. Various visual saliency models have been developed in the literature that are inspired by the human visual attention processing behavior and the generated saliency maps have been incorporated in the design of objective quality assessment algorithms to better replicate the HVS characteristics in quality assessment. In the FR scenario, saliency values are often used to weight the pixel-wise quality difference between the comparing reference and distorted images. Zhang *et al.* [33] studied the added value of 20 different saliency models for the task of 2D IQA through a comprehensive statistical analysis. The outcomes revealed a statistically significant gain in the performance of objective IQA models when incorporating the saliency models. Here, we benefit from saliency detection to improve the quality assessment of the 3D synthesized views.

Saliency models aim to capture biologically-inspired features by considering image intensity, color, edge, and texture. As mentioned earlier, higher visual areas (such as V3, and V4) can be characterized by sensitivity to natural textures and the neurons in these areas are more selective for complex textures and shapes of the stimuli [34], [35]. Assuming that the higher layers of CNN multi-scale architectures can roughly replicate the complex responses of the higher visual areas of the human visual cortex, we utilized the activation maps of deep CNN layers to generate visual saliency maps for quality assessment. Instead of using an off-the-shelf saliency model, we take advantage of the feature maps of the last CNN layer (L_5) of ResNet50 to highlight the RoI.

We constructed feature maps of the reference image from the 2048 feature activation maps in the fifth CNN layer. Thus, a set of n 2D feature maps in the lth layer is defined as:

$$F^l = \left\{f_1^l, f_2^l, \ldots, f_n^l\right\} \tag{4}$$

The set of feature maps is then upsampled ($\widehat{F}^l$) to the input image size (224 × 224) using the bicubic interpolation. Finally, the pixel values of the 2048 upsampled maps are aggregated to obtain the visual importance probability map (Sal):

$$Sal = \sum_{i=1}^{n} \widehat{F}_i^l \tag{5}$$

Due to the misalignment between the reference and synthesized views, it is not straightforward to directly use saliency maps for pixel-wise weighting. Instead, saliency-masked images are generated to compare RoI features in reference and test images. Using the obtained saliency maps, new masked inputs are constructed by discarding the non-salient regions as follows:

$$I_m(i,j) = \begin{cases} I(i,j) & Sal(i,j) > \mu \\ 0 & otherwise \end{cases} \tag{6}$$

where μ is the mean pixel value of the saliency map Sal, and I_m is the saliency-masked input image. Fig. 4 depicts the procedure of the saliency map generation and masking. The saliency maps obtained from different reference images of the IETR data set are presented in Fig. 5 which implies that the deployed method can effectively highlight the visually important objects and regions of the scene.

The saliency-masked reference and synthesized views are feed forwarded to the network to obtain the features of the salient regions. Finally, quality-aware and semantic-aware similarity measures are computed as in (1) and (2) and the overall quality score for the saliency channel $\overline{Q_2}$ is obtained by averaging two scores as described in (3). To compute the quality-aware measure of the salient region (Q_p^{Sal}), the third CNN layer was used for feature extraction as it showed the highest performance among the five convolutional layers.

D. FINAL SEQUSS SCORE

The objective quality scores from the two computational streams are combined to obtain the final quality score. Although the saliency-based IQA is important, it does not account for severe distortions that might appear in non-salient regions. Therefore, we integrated the saliency score $\overline{Q_2}$ with the global score from the first computational stream $\overline{Q_1}$ by computing the weighted sum of the two scores as follows.

$$SEQUSS = \beta\overline{Q_1} + (1-\beta)\overline{Q_2} \tag{7}$$

where β is a non-zero constant parameter set to 0.6 to slightly increase the weight of the global score.

III. EXPERIMENTAL RESULTS

This section describes the performance evaluation results of the proposed SEQUSS metric. The IETR [19] is a new data set of 140 DIBR synthesized views that is used to benchmark our quality model and it is the only subjectively-annotated public data set that include both the conventional and new DIBR approaches. Conventional DIBR distortions such as severe black holes and large geometric shifts are not considered in this data set since such artifacts are not the main visual impairments in the new DIBR algorithms. Subjective scores are gathered from 42 naive participants and a Differential Mean Opinion Scores (DMOS) is acquired for each test stimuli.

Besides geometric and stretching distortions, the performance of the inpainting approach used in the DIBR plays an important role in the quality of the synthesized views. Therefore, recent efforts are mostly focused on proposing more accurate inpainting approaches to improve the hole-filling in the disoccluded regions. The IETR encompasses 10 MVD sequences processed by 7 DIBR algorithms including: Criminisi [41], View Synthesis Reference Software (VSRS) from MPEG 3D video group [42], Layered Depth Image (LDI) DIBR approach [20], Hierarchical Hole-Filling (HHF) method [43], Ahn's [44], Luo's [45], and Zhu's [46] hole-filling methods. The VSRS algorithm is deployed in two scenarios for single-view (VSRS1) and multi-view (VSRS2) synthesis applications. The performance of the proposed methods is compared against 15 objective IQA methods including five FR methods (PSNR, SSIM [7], VIF [8], GMSD [36], and FSIM [37]), five NR DIBR methods (NIQSV [9], NIQSV+ [10], MNSS [11], NR-MWT [38], and Jakhetia's [12]), and six FR DIBR models (MW-PSNR [13], MP-PSNR [14], LOGS [15], SC-IQA [16], Peng *et al.* [39], Sui *et al.* [40], and SSPD [17]).

Spearman's Rank-order Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), Kendall's Rank Correlation Coefficient (KRCC), and Root Mean Square Error (RMSE) are four evaluation criteria that were deployed to compare the performance of IQA models against the subjective scores. A higher value of SRCC, PLCC, and KRCC indicates a higher consistency of the objective scores with human opinions and better performance. We applied the following nonlinear fitting function to the objective scores x prior to the computation of the evaluation indices:

$$F(x) = \lambda_1\left(\frac{1}{2} - \frac{1}{1+e^{\lambda_2(x-\lambda_3)}}\right) + \lambda_4 x + \lambda_5 \tag{8}$$

where λ_1 to λ_5 denote the fitting parameters.

A. PERFORMANCE EVALUATION ON THE IETR DATA SET

Table 1 compares the efficacy of the proposed method with the competing approaches. Our IQA model achieves a substantial improvement of the prediction accuracy in terms of all four evaluation indices. Compared to the SSPD which is in the second rank, the SEQUSS model delivers a performance gain of 14.6% in SRCC, 14.4% in PLCC, 22.1%

TABLE 1. Performance comparison of the proposed IQA metric with 15 objective IQA models on the IETR data set.

IQA methods	Type	SRCC	PLCC	KRCC	RMSE
PSNR	FR (2D)	0.536	0.605	0.376	0.197
SSIM [7]	FR (2D)	0.229	0.424	0.156	0.224
VIF [8]	FR (2D)	0.259	0.269	0.173	0.238
GMSD [36]	FR (2D)	0.478	0.542	0.325	0.208
FSIM [37]	FR (2D)	0.483	0.498	0.328	0.215
NIQSV [9]	NR (DIBR)	0.155	0.182	0.108	0.243
NIQSV+ [10]	NR (DIBR)	0.230	0.230	0.156	0.241
MNSS [11]	NR (DIBR)	0.296	0.293	0.194	0.237
NR-MWT [38]	NR (DIBR)	0.456	0.472	0.317	0.218
Jakhetia et al. [12])	NR (DIBR)	0.165	0.164	0.112	0.244
MW-PSNR [13]	FR (DIBR)	0.487	0.537	0.340	0.209
MP-PSNR [14]	FR (DIBR)	0.548	0.576	0.385	0.202
LOGS [15]	FR (DIBR)	0.616	0.624	0.442	0.193
SC-IQA [16]	FR (DIBR)	0.596	0.663	0.423	0.185
Peng et al. [39]	FR (DIBR)	0.658	0.665	-	0.193
Sui et al. [40]	FR (DIBR)	0.593	0.673	-	0.183
SSPD [17]	FR (DIBR)	0.685	0.702	0.467	0.179
SEQUSS	FR (DIBR)	**0.802**	**0.803**	**0.600**	**0.147**

TABLE 2. Performance of the individual quality measures used in the SEQUSS metric.

IQA measures	SRCC	PLCC	KRCC	RMSE
Q_p	0.755	0.760	0.567	0.166
Q_s	0.703	0.716	0.515	0.170
$\overline{Q_1}$	0.778	0.781	0.581	0.157
Q_p^{Sal}	0.688	0.695	0.503	0.172
Q_s^{Sal}	0.595	0.594	0.436	0.206
$\overline{Q_2}$	0.700	0.703	0.512	0.165
$SEQUSS$	0.802	0.803	0.600	0.148

TABLE 3. Performance of the features in five CNN layers to compute the global quality-aware measure (Q_p), and the saliency-guided quality-aware measure (Q_p^{Sal}).

ResNet Layers		SRCC	PLCC	KRCC	RMSE
Q_p	L1	0.593	0.646	0.419	0.189
	L2	0.679	0.711	0.492	0.174
	L3	0.753	0.763	0.562	0.160
	L4	**0.764**	**0.770**	**0.570**	**0.158**
	L5	0.736	0.764	0.536	0.160
Q_p^{Sal}	L1	0.600	0.552	0.386	0.198
	L2	0.636	0.665	0.455	0.185
	L3	**0.696**	**0.704**	**0.508**	**0.176**
	L4	0.686	0.687	0.501	0.180
	L5	0.601	0.615	0.434	0.195

in KRCC, and 17.8% in RMSE. As shown in the table, the FR algorithms for 2D images and NR DIBR methods fail to provide a meaningful correlation with the ground-truth subjective scores and the performance of the competing FR DIBR methods is low. While some DIBR IQA models have shown promising results on the conventional DIBR data sets such as IRCCyN/IVC [18], the outcomes reveal the failure of the existing DIBR IQA models on the new IETR data set.

The proposed SEQUSS model consists of several computational units that deliver the quality-aware and semantic-aware measures using the full synthesized view as well as the salient regions. Table 2 presents the performance of each individual quality measure as well as the overall performance on the IETR data set. The quality-aware measures (Q_p, Q_p^{Sal}) have a higher correlation with DMOS when compared to the semantic measures (Q_s, Q_s^{Sal}) while the integration of these two measures could further improve the performance of the quality assessment. The second computational stream measures the quality by focusing on the visual salient regions and the integration of the quality scores of this channel ($\overline{Q_2}$) with the global score of the first stream ($\overline{Q_1}$) lead to better overall estimation accuracy.

As mentioned in section II, we used the fourth CNN Layer for feature extraction in the first computational stream (global), and the second stream (salient region) deploys the features of the third CNN layer. In Table 3, we reported the performance of all five CNN layers. As shown in the table, Layer 3 and 4 always have the highest performance compared to other layers. Comparing the SRCC in two streams, the accuracy is slightly shifted toward lower layers when deploying salient regions. Lower layers allow a more detailed comparison of the frequency components, although the tolerance to geometric distortions and misalignment issues are diminished when moving toward the lower layers.

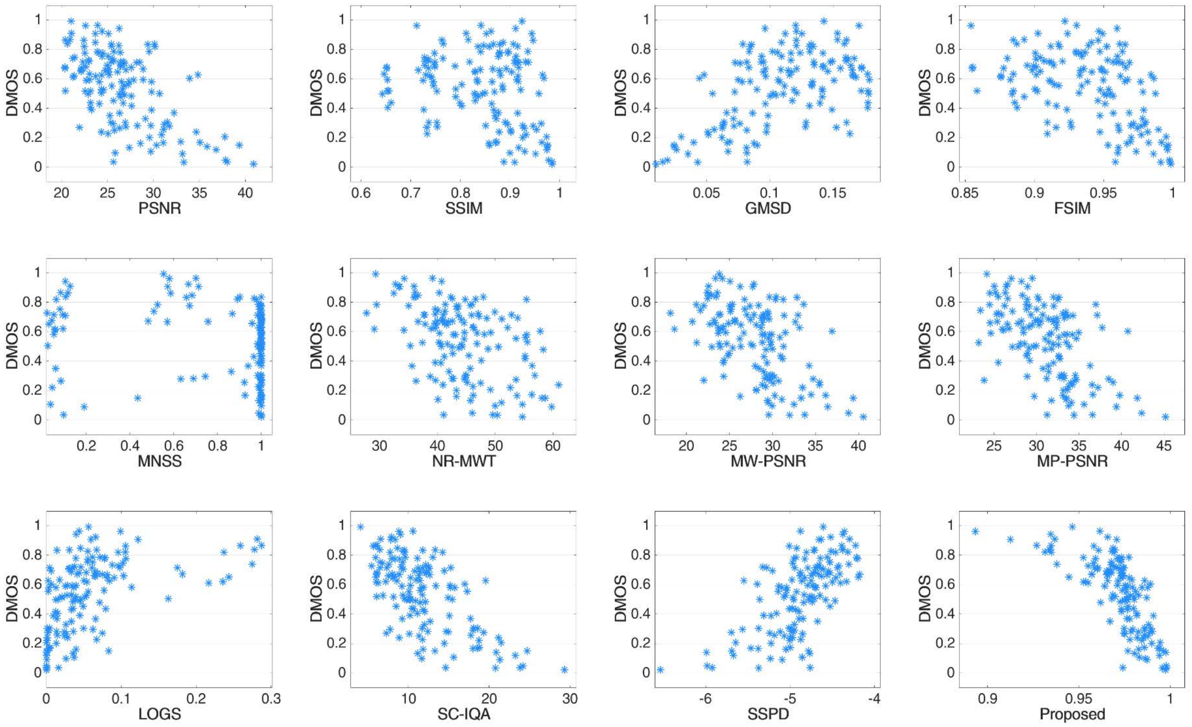


FIGURE 6. Scatter plots of the objective scores versus DMOS on the IETR data set.

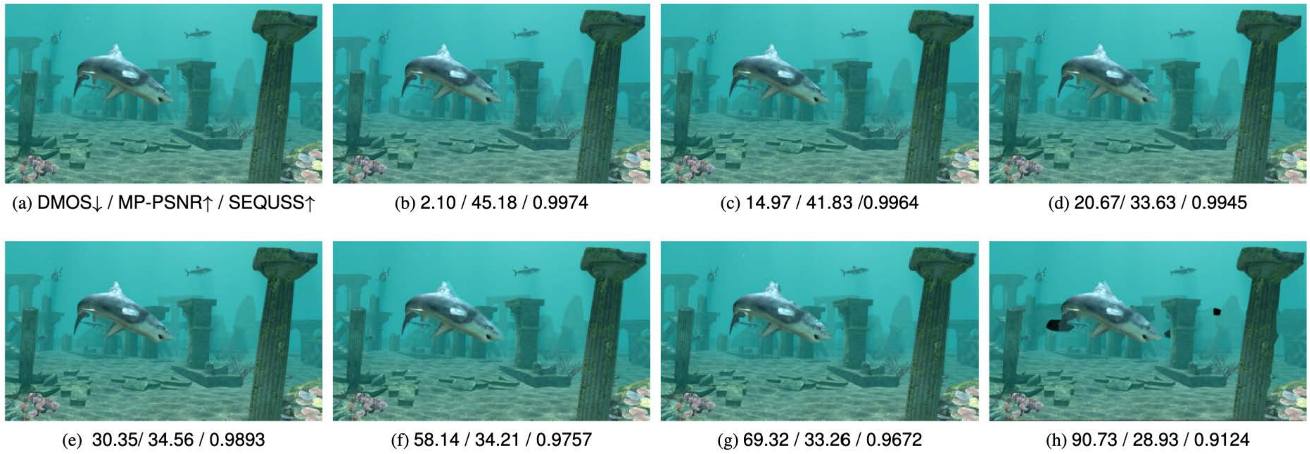
Fig. 6 illustrates the scatter plots of the objective scores versus the DMOS on the IETR data set. The plot presents a better convergence of the data points when using the SEQUSS which implies higher agreement of the proposed metric with the subjective opinions compared to other competing approaches. A higher SEQUSS score and lower DMOS indicate better quality. The plots indicate severe failures (rather random distribution of scatter points) for several FR methods including SSIM and FSIM, suggesting that these general FR methods are unreliable for quality assessments of the synthesized views. Fig. 7 visualizes an example of the SEQUSS scores assigned to the synthesized views of a reference image ('Shark') in the IETR data set. The synthesized views (b-h) are arranged from highest to lowest DMOS values. As the figure shows, the SEQUSS is performing quite well and the objective scores (higher is better) are highly consistent with the human subjective scores (lower is better), however, the other competing method (MP-PSNR) indicates some disagreements with the subjective DMOS.

B. STATISTICAL SIGNIFICANCE TEST

We performed a statistical significance test according to the ITU-T Recommendation P.1401 [47]. We applied a two-sample t-test with 95% confidence level on the SRCC values of all the metric pairs under the null hypothesis that there is no significant difference between the two metrics. The null hypothesis is rejected at 5% significance level. The outcomes of the significance test are presented in Table 4 in which the symbols '1' implies that an IQA model in the row axis is superior to a metric in the column axis, and '-1' indicates the inferior performance of the row metric compared to the metrics in the column axis. The symbol '0' denotes that the difference is not significant. The table confirms that the proposed method performs significantly better than all other approaches including SSPD as the second-best method.

C. PERFORMANCE EVALUATION ON THE IRCCyN/IVC DATA SET

Performance evaluation experiments confirm the superiority of the proposed methods compared to other DIBR methods that are mainly devised to quantify conventional DIBR distortions. The existing state-of-the-art DIBR techniques often suffer from moderate geometric distortions and a flawed inpainting process to fill the holes in the disoccluded regions. However, most available objective metrics tried to address conventional artifacts by deploying shift compensation strategies or methods to quantify the visual impact of the black holes. Although the proposed method is devised to address the new DIBR quality assessment requirements, in this section we reported the performance of the proposed SEQUSS on the traditional IRCCyN/IVC DIBR data set [18]. This data set contains 12 reference and 84 test views synthesized using 7 traditional DIBR algorithms developed between the years 2003 to 2010. We excluded 12 test images synthesized by Fehn's algorithm due to the severe shifts applied to the synthesized views. The quality



(a) DMOS↓ / MP-PSNR↑ / SEQUSS↑ (b) 2.10 / 45.18 / 0.9974 (c) 14.97 / 41.83 /0.9964 (d) 20.67/ 33.63 / 0.9945

(e) 30.35/ 34.56 / 0.9893 (f) 58.14 / 34.21 / 0.9757 (g) 69.32 / 33.26 / 0.9672 (h) 90.73 / 28.93 / 0.9124

FIGURE 7. Example to visually show the performance of the proposed method. (a) original 'Shark' image from the IETR data set. (b) VSRS2, (c) Zhu's, (d) VSRS1, (e) Ahn's, (f) HHF, (g) Criminisi, and (h) LDI. The DMOS human opinion scores, MP-PSNR, and SEQUSS scores are presented below each image. The SEQUSS scores (a higher is better ↑) are consistent with the DMOS (a lower value is better ↓).

TABLE 4. Statistical significance test in terms of SRCC on the IETR data sets.

	PSNR	FSIM	NR-MWT	MW-PSNR	MP-PSNR	LOGS	SC-IQA	SSPD	Proposed
PSNR	0	0	0	0	0	0	0	-1	-1
FSIM	0	0	0	0	0	0	0	-1	-1
NR-MWT	0	0	0	0	0	0	0	-1	-1
MW-PSNR	0	0	0	0	0	0	0	-1	-1
MP-PSNR	0	0	0	0	0	0	0	0	-1
LOGS	0	0	0	0	0	0	0	0	-1
SC-IQA	0	0	0	0	0	0	0	0	-1
SSPD	1	1	1	1	0	0	0	0	-1
Proposed	1	1	1	1	1	1	1	1	0

comparison was performed on the remaining 72 synthesized views. Table 5 compares the objective IQA performance on the IRCCyN/IVC data set in terms of the four evaluation indices. The results show that the SEQUSS model can still perform well and it is among the top 3 methods in terms of SRCC on the IRCCyN/IVC data set. The highest SRCC of 0.865 is achieved by SSPD, our recently proposed method, which is developed based on an accurate shift compensation and black hole effect measurement.

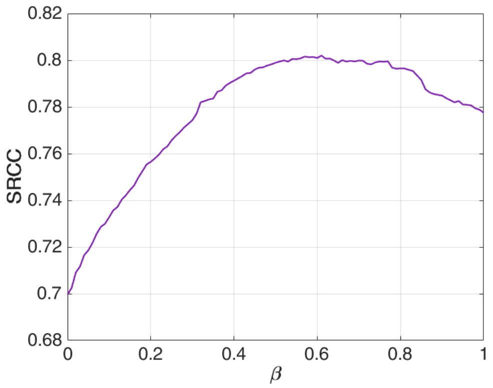


FIGURE 8. Performance of the proposed method (in terms of SRCC) for various values of the weight parameter β.

D. SENSITIVITY TO THE WEIGHTING PARAMETER

The proposed model has only one parameter to adjust to obtain optimal performance. The weight parameter β in (7) specifies the weight of the scores from two computational streams (i.e. an score computed using the entire image $\overline{Q_1}$ versus an score of the salient region $\overline{Q_2}$). Although the quality assessment of salient region is important, it does not consider the quality loss on other non-salient areas that might attract attention especially when severe distortion is occurring. Therefore, the global quality score and the saliency score are combined using a weighting function. Fig. 8 shows how the SRCC varies by selecting different values of β. The highest performance is achieved around $\beta = 0.6$ whereas there is no abrupt change around the selected value which shows the performance stability of the metric.

The SEQUSS is more computationally efficient compared to SSPD. We performed a test on an image of size 1024 × 768 on a Windows laptop with 16 GB RAM and a Core i7-2.7 GHz CPU in which an average processing time of 2.11 seconds over 10 repetitions is achieved using the

TABLE 5. Performance comparison of the proposed IQA metric with 15 objective IQA models on the IRCCyN/IVC data set.

IQA methods	Type	SRCC	PLCC	KRCC	RMSE
PSNR	FR (2D)	0.544	0.610	0.386	0.466
SSIM	FR (2D)	0.447	0.595	0.300	0.536
VIF	FR (2D)	0.207	0.258	0.143	0.645
GMSD	FR (2D)	0.560	0.614	0.393	0.467
FSIM	FR (2D)	0.569	0.622	0.396	0.450
NIQSV	NR (DIBR)	0.190	0.382	0.122	0.523
NIQSV+	NR (DIBR)	0.488	0.580	0.347	0.483
MNSS	NR (DIBR)	0.660	0.615	0.466	0.526
NR-MWT	NR (DIBR)	0.562	0.600	0.391	0.437
Jakhetia et al.	NR (DIBR)	0.733	0.786	0.525	0.412
MW-PSNR	FR (DIBR)	0.776	0.856	0.583	0.344
MP-PSNR	FR (DIBR)	0.818	0.889	0.644	0.305
LOGS	FR (DIBR)	0.754	0.771	0.538	0.424
SC-IQA	FR (DIBR)	0.772	0.850	0.599	0.351
SSPD	FR (DIBR)	0.865	0.896	0.700	0.295
SEQUSS	FR (DIBR)	0.797	0.859	0.603	0.338

SEQUSS which is more than 10x faster than SSPD with the execution time of 21.67 seconds.

IV. CONCLUSION

This paper proposed SEQUSS, a full-reference IQA metric to predict the quality of DIBR-synthesized views. The method takes advantage of ResNet50, a pre-trained deep CNN model, to transform reference and synthesized views to a representation that better complies with the HVS characteristics. The features of the CNN were used to compute the structural and semantic similarities between the reference and synthesized images. The two similarity measures were then unified to compute an overall quality score. The selected feature space can effectively quantify the visual impact of challenging distortion types of DIBR images while it is also robust to the geometrical shifts. To incorporate the human visual attention properties in quality assessment, we produced saliency maps using the feature activations of the CNN layers. Quality assessment on the selected regions of interest could further improve the performance accuracy of the proposed model. While none of the competing IQA models could perform well on the new IETR data set, our SEQUSS model improved the IQA of DIBR images by a large margin.

The proposed metric deploys the features from the deeper layers of the ResNet50 to estimate the visual quality. In future work, strategies for selecting perceptually-important features from the CNN layers should be developed. Considering the multi-scale representation of features in CNNs, another possible direction is to design a multi-scale quality metric that takes advantage of the information of all layers, however, the issue of high sensitivity to geometric distortions and pixel shifts in lower layers must be addressed.

REFERENCES

[1] M. Tanimoto, "FTV (free-viewpoint TV)," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 67–76.

[2] R. K. Muhamad, T. Birnbaum, A. Gilles, S. Mahmoudpour, K.-J. Oh, M. Pereira, C. Perra, A. Pinheiro, and P. Schelkens, "JPEG Pleno holography: Scope and technology validation procedures," *Appl. Opt.*, vol. 60, no. 3, p. 641, Jan. 2021.

[3] T. Ebrahimi, S. Foessel, F. Pereira, and P. Schelkens, "JPEG Pleno: Toward an efficient representation of visual reality," *IEEE Multimedia*, vol. 23, no. 4, pp. 14–20, Oct./Dec. 2016.

[4] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2007, pp. 1–4.

[5] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, *Reference Softwares for Depth Estimation and View Synthesis*, ISO/IEC JTC1/SC29/WG11 MPEG Standard 20081, 2008.

[6] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," *Proc. SPIE*, vol. 5291, pp. 93–104, May 2004.

[7] 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.

[8] 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.

[9] 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)*, Mar. 2017, pp. 1248–1252.

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

[11] 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.*, vol. 66, no. 1, pp. 127–139, Mar. 2020.

[12] 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. Informat.*, vol. 15, no. 7, pp. 4120–4128, Jul. 2019.

[13] D. Sandic-Stankovic, D. Kukolj, and P. Le Callet, "DIBR synthesized image quality assessment based on morphological wavelets," in *Proc. 7th Int. Workshop Quality Multimedia Exper. (QoMEX)*, May 2015, pp. 1–6.

[14] D. Sandic-Stankovic, D. Kukolj, and P. Le Callet, "DIBR synthesized image quality assessment based on morphological pyramids," in *Proc. 3DTV-Conf., True Vision—Capture, Transmiss. Display 3D Video (3DTV-CON)*, 2015, pp. 1–4, doi: 10.1109/3DTV.2015.7169368.

[15] L. D. Li, Y. Zhou, K. Gu, W. S. Lin, and S. Q. 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.

[16] S. Tian, L. Zhang, L. Morin, and O. Deforges, ''SC-IQA: Shift compensation based image quality assessment for DIBR-synthesized views,'' in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2018, pp. 1–4.

[17] S. Mahmoudpour and P. Schelkens, ''Synthesized view quality assessment using feature matching and superpixel difference,'' *IEEE Signal Process. Lett.*, vol. 27, pp. 1650–1654, 2020.

[18] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, ''Towards a new quality metric for 3-D synthesized view assessment,'' *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 7, pp. 1332–1343, Nov. 2011.

[19] 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, Oct. 2019.

[20] V. Jantet, C. Guillemot, and L. Morin, ''Object-based layered depth images for improved virtual view synthesis in rate-constrained context,'' in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 125–128.

[21] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[22] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, ''Image quality assessment: Unifying structure and texture similarity,'' 2020, *arXiv:2004.07728*.

[23] D. H. Hubel and T. N. Wiesel, ''Receptive fields and functional architecture of monkey striate cortex,'' *J. Physiol.*, vol. 195, no. 1, pp. 215–243, 1968.

[24] Y. El-Shamayleh and A. Pasupathy, ''Contour curvature as an invariant code for objects in visual area V4,'' *J. Neurosci.*, vol. 36, no. 20, pp. 5532–5543, 2016.

[25] T. Kim, W. Bair, and A. Pasupathy, ''Neural coding for shape and texture in macaque area V4,'' *J. Neurosci.*, vol. 39, no. 24, pp. 4760–4774, Jun. 2019.

[26] C. Zhuang, Y. Wang, D. Yamins, and X. Hu, ''Deep learning predicts correlation between a functional signature of higher visual areas and sparse firing of neurons,'' *Frontiers Comput. Neurosci.*, vol. 11, p. 100, Oct. 2017.

[27] M. N. U. Laskar, L. G. S. Giraldo, and O. Schwartz, ''Deep neural networks capture texture sensitivity in V2,'' *J. Vision*, vol. 20, no. 7, pp. 1–21, 2020.

[28] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, ''The unreasonable effectiveness of deep features as a perceptual metric,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, ''ImageNet large scale visual recognition challenge,'' *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[30] E. Siahaan, A. Hanjalic, and J. A. Redi, ''Augmenting blind image quality assessment using image semantics,'' in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2016, pp. 307–312.

[31] E. Siahaan, A. Hanjalic, and J. A. Redi, ''Does visual quality depend on semantics? A study on the relationship between impairment annoyance and image semantics at early attentive stages,'' *Electron. Imag.*, vol. 28, no. 16, pp. 1–9, Feb. 2016.

[32] K. Sim, J. Yang, W. Lu, and X. Gao, ''Blind stereoscopic image quality evaluator based on binocular semantic and quality channels,'' *IEEE Trans. Multimedia*, vol. 24, pp. 1389–1398, 2022.

[33] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, ''The application of visual saliency models in objective image quality assessment: A statistical evaluation,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1266–1278, Jun. 2016.

[34] J. Freeman, C. M. Ziemba, D. J. Heeger, E. P. Simoncelli, and J. A. Movshon, ''A functional and perceptual signature of the second visual area in primates,'' *Nature Neurosci.*, vol. 16, no. 7, pp. 974–981, Jul. 2013.

[35] G. Okazawa, S. Tajima, and H. Komatsu, ''Gradual development of visual texture-selective properties between macaque areas V2 and V4,'' *Cerebral Cortex*, vol. 27, pp. 4867–4880, Sep. 2016.

[36] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, ''Gradient magnitude similarity deviation: A highly efficient perceptual image quality index,'' *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[37] 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.

[38] D. D. Sandic-Stankovic, D. D. Kukolj, and P. Le Callet, ''Fast blind quality assessment of DIBR-synthesized video based on high-high wavelet subband,'' *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5524–5536, Nov. 2019.

[39] Z. Peng, Q. Jiang, F. Shao, W. Gao, and W. Lin, ''LGGD+: Image retargeting quality assessment by measuring local and global geometric distortions,'' *IEEE Trans. Circuits Syst. Video Technol.*, early access, Sep. 14, 2021, doi: 10.1109/TCSVT.2021.3112933.

[40] X. Sui, M. Ding, J. Yan, Y. Fang, Y. Zuo, and Z. Tan, ''Objective quality assessment of synthesized images by local variation measurement,'' *Signal Process., Image Commun.*, vol. 92, Mar. 2021, Art. no. 116096.

[41] A. Criminisi, P. Pérez, and K. Toyama, ''Region filling and object removal by exemplar-based image inpainting,'' *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[42] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, ''*Reference Softwares for Depth Estimation and View Synthesis*, ISO/IEC JTC1/SC29/WG11 MPEG Standard 20081, 2008.

[43] M. Solh and G. AlRegib, ''Hierarchical hole-filling for depth-based view synthesis in FTV and 3D video,'' *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 5, pp. 495–504, Sep. 2012.

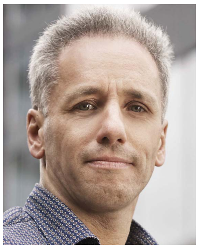
[44] I. Ahn and C. Kim, ''A novel depth-based virtual view synthesis method for free viewpoint video,'' *IEEE Trans. Broadcast.*, vol. 59, no. 4, pp. 614–626, Dec. 2013.

[45] G. Luo, Y. Zhu, Z. Li, and L. Zhang, ''A hole filling approach based on background reconstruction for view synthesis in 3D video,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1781–1789.

[46] C. Zhu and S. Li, ''Depth image based view synthesis: New insights and perspectives on hole generation and filling,'' *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 82–93, Mar. 2016.

[47] *Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*, Recommendation ITU-T P.1401, Geneva, Switzerland, 2012. [Online]. Available: https://www.itu.int/rec/RREC-BT.500

SAEED MAHMOUDPOUR (Member, IEEE) received the Ph.D. degree in computer and communications engineering, in 2017. Afterwards, he joined imec, Belgium, as a Senior Research Scientist. He is currently a Postdoctoral Fellow and a Lecturer with imec and the Department of Electronics and Informatics, Vrije Universiteit Brussel (VUB), Brussels, Belgium. His research interests include visual quality assessment, machine learning, and high dynamic range imaging. He is a member of the ISO/IEC JTC1/SC29/WG1 (JPEG) standardization committee and the Co-Chair of the JPEG Pleno Light Field subgroup.



PETER SCHELKENS (Member, IEEE) holds a professorship with the Department of Electronics and Informatics, Vrije Universiteit Brussel, Belgium, and is a Research Group Leader at imec, Belgium. In 2013, he received an EU ERC Consolidator Grant focusing on digital holography. He is the co-editor of the books *The JPEG 2000 Suite* (Wiley, 2009) and *Optical and Digital Image Processing* (Wiley, 2011). He is the Chair of the Plenoptic Coding and Quality subgroup of the ISO/IEC JTC1/SC29/WG1 (JPEG) standardization committee.