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/TMM.2020.3038305, IEEE Transactions on Multimedia

Re-visiting Discriminator for Blind Free-Viewpoint Image Quality Assessment

Suiyi Ling, *Member, IEEE*, Jing Li, *Member, IEEE*, Zhaohui Che, Wei Zhou, Junle Wang, and Patrick Le Callet, *Fellow, IEEE*

***Abstract*—Accurate measurement of perceptual quality is important for various immersive multimedia, which demand real-time quality control or quality-based bench-marking for relevant algorithms. For instance, virtual views rendering in Free-Viewpoint (FV) navigation scenarios is a typical case that introduces challenging distortions, particularly the ones around dis-occluded regions. Existing quality metrics, most of which are targeting for impairments caused by compression or network condition, fail to quantify such non-uniform structure-related distortions. Moreover, the lack of quality databases for such distortions makes it even more challenging to develop robust quality metrics. In this work, a Generative Adversarial Networks based No-Reference (NR) quality Metric, namely GANs-NRM, is proposed. We first present an approach to create masks mimicking dis-occlusions/textureless regions, which is applicable on large-scale 2D image databases publicly available in the computer vision domain. Using these synthetic data, we then train a GANs-based context renderer with the capability of rendering those masked regions. Since the naturalness of the rendered dis-occluded regions strongly relates to the perceptual quality, we assume that the discriminator of the trained GANs has an intrinsic ability for quality assessment. We thus use the features extracted from the discriminator to learn a Bag-of-Distortion-Word (BDW) codebook. We show that a quality predictor can be then well trained using only a small amount of subjective quality data for the FV views rendering. Moreover, in the proposed framework, the discriminator is also adapted as a distortion-detector to locate possible distorted regions. According to the experimental results, the proposed model outperforms significantly the state-of-the-art quality metrics. The corresponding context renderer also shows appealing visualized results over other rendering algorithms.**

***Index Terms*—Generative adversarial networks, no-reference quality assessment, depth-image-based-rendering, free-viewpoint navigation, non-uniform structure-related distortions**

I. Introduction

Recent advances in immersive equipment have attracted greater users' interests and raised a novel revolution in the viewing experience. Most of the immersive scenarios or applications are based on rendering technologies, including 3D-TV [1], [2], Free-viewpoint TV (FTV) [3], Virtual Reality (VR) [4], and light field [5]. For instance, FTV offers a 'flying in the scene' experience to users by allowing them to change the viewpoints freely in applications like virtual conferences, broadcasting live concerts, remote surveillance, *etc*. To realize this functionality, new virtual views need to be rendered with limited reference views taken by calibrated anchor cameras with wide-baseline as compressing and transmitting massive numbers of views are expensive and inefficient.

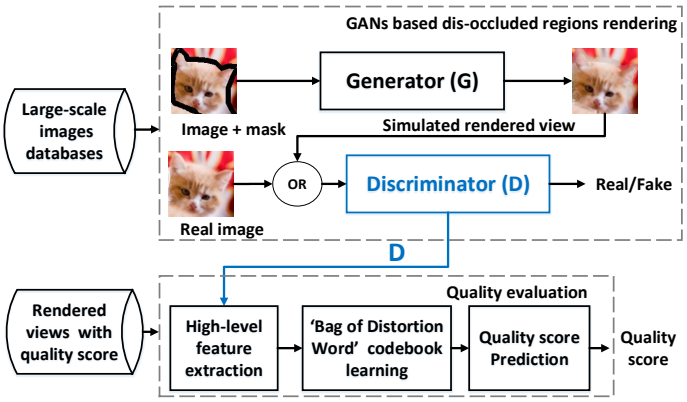


Fig. 1. The proposed No-Reference quality assessment framework. Step 1: A GANs is trained to render simulated *dis-occluded regions* with **large-scale 2D** databases; Step 2: High-level features related to rendering quality, *i.e.*, naturalness, are extracted from discriminator and used to learn a *Bag-of-Distortion-Word* codebook so that the quality could be then evaluated with **limited** real rendered views labeled with quality score.

The Depth-Image-based Rendering (DIBR) [6] is one of the most widely adopted techniques for rendering virtual viewpoints in free-viewpoint navigation applications. One of the common challenging processes for DIBR techniques is to render *dis-occluded regions* (*dis-occlusions*, *textureless regions* or *holes*), which are the regions that can be seen in the virtual views but occluded in the reference/anchor views [7]. In most cases, DIBR techniques render *dis-occluded regions* based on inpainting algorithms. Nevertheless, DIBR techniques usually introduce similar challenging non-uniform structure-related distortions in rendered views mainly due to the occlusions. Examples of distortions introduced by DIBR algorithms [7]–[10] are shown in Figure 2. These structure-related distortions include (a) *blurry regions*; (b) *foreground bleeding into background* or vise versa; (c) *inconsistent structure/geometric distortions*; (d) *dark holes/textureless regions*; (e) *ghosting artifacts*.

These local non-uniform structure-related distortions introduced by rendering algorithms could be predominant in

Jing Li is the corresponding author, and she is with Alibaba Group, China (e-mail: jing.li.univ@gmail.com).

Suiyi Ling and Patrick Le Callet are with the Équipe Image, Perception et Interaction, Laboratoire des Sciences du Numérique de Nantes, Université de Nantes, France (e-mail: suiyi.ling@univ-nantes.fr; patrick.lecallet@univ-nantes.fr).

Zhaohui Che is with the Department of Electronic Engineering, Shanghai Jiao Tong University, China (e-mail: chezhaohui@sjtu.edu.cn).

Junle Wang is with Tencent, China (e-mail: wangjunle@gmail.com).

Wei Zhou is with University of Science and Technology of China, China (e-mail: weizhou@mail.ustc.edu.cn).

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/TMM.2020.3038305, IEEE Transactions on Multimedia

(a) (b) (c) (d) (e)

Fig. 2. Distortions introduced by DIBR techniques, which are mainly located around the *dis-occlusion regions* (dominant distortions are highlighted by red bounding boxes).

impacting the perceived quality, as they locate most of the time along the transition areas between foreground objects and background (*e.g., the dis-occluded regions*). In the *free-viewing system*, rendering process could therefore be the 'bottleneck' of delivering good quality of free viewing experience in most cases [11]. As thus, a robust quality assessment metric is necessary to benchmark rendering/synthesis algorithms, provide guidance for the optimization of the overall system performance and improve the Quality of Experience (QoE) of the users. However, these DIBR related artifacts, especially the ones introduced by *dis-occluded regions* filling/rendering, are challenging for existing commonly used quality metrics [12] designed for the uniformly distributed artifacts (*e.g.*, compression related artifacts).

Quality assessment metrics could be divided into three categories, including the Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) metric. In these three cases, complete, partial and no information of the reference image/video is required. In scenarios like free-viewpoint navigation, reference views are not accessible [11], therefore, NR metrics are of wider application values than the FR ones [11]. In 2D image quality assessment domain, there are already some NR metrics which achieve promising performances by employing the advanced deep learning techniques. Generally, the training of these metrics relies on 1) vast amounts of training data labeled with quality scores, and 2) assigning image-level quality scores to patches [13]. Nevertheless, there is a clear lack of multi-view contents and related databases that are equipped with quality scores in free-viewpoint applications. In addition, assigning image-level labels to patches becomes problematic in our cases since the dominant distortions within rendered contents are non-uniformly distributed. A specific NR quality model is thus in an urgent need.

Generative Adversarial Networks (GANs), first proposed by Goodfellow *et al.* [14], might be a solution to this question. GANs have been widely used in various domains, including 3D structure reconstruction [15], semantic inpainting (*i.e.*, *context renderer*) [16]–[20], realistic image synthesis [21], *etc*. The main idea of the adversarial nets framework is to train a generator (G) and a discriminator (D) simultaneously: a generative model that captures the data distribution and a discriminative model that is able to tell the 'real' image from the generated one. They are trained together so that the discriminator can keep pitting against the generator until the counterfeits generated by the generator cannot be distinguished by the discriminator. By doing so, both of them would be driven to improve their performance until convergence [14]. Taking *context renderer* as an example, the goal of G is to render the missing part of an image (with a mask map indicating the regions to be rendered), and 'trick' the discriminator to considered the rendered part as 'real/natural' image region; the goal of D is to distinguish between a rendered image and a real image, *i.e.*, ground truth image.

Considering the case of rendering virtual views using DIBR techniques, the most annoying non-uniform distortions usually are the non-continuous distortions introduced during the *dis-occlusion regions* rendering stage. If the *context inpainter/renderer* is trained to render *dis-occluded regions*, then the discriminator (trained along with the generator) is supposed to be able to tell whether the input image is inpainted/rendered or not. This idea is summarized in Fig. 1. Intuitively, the discriminator makes the decision based on the distribution of the input. In FV views rendering applications, the naturalness of the rendered regions determines the perceptual quality. Thus, we assume that the discriminator has an indirect relationship with the rendering-quality of the input. In this study, the discriminator is thus utilized to detect the local poorly rendered regions, and learn the perceptual quality of the rendered image. Furthermore, according to recent theoretical analysis of *perceptual loss* [22], high-level representations extracted from hidden layers of well-trained networks for certain task, *e.g,* classification, are beneficial to interpret the *task-related semantics* of the input [23]. Ideally, given the task of rendering *dis-occluded regions* to a GANs, high-level features extracted from its discriminator are supposed to have the capability to capture the perceptual information that reflects the rendering-quality.

Based on the discussion above, in this paper, a GANs based NR quality Metric (GANs-NRM) is proposed for evaluating the quality of rendered *free-viewpoint* virtual views. The overall idea is concluded in Fig. 1. In the first step, as there are limited available *free-viewpoint* contents that could be used for training, masks that mimic *dis-occluded regions* that appear in rendered virtual view are designed. Then, a generative adversarial network is trained with available 2D large-scale image dataset without quality score to render the simulated *dis-occlusions*. By doing so, its discriminator is trained to indicate whether those regions are well rendered. To tackle the problem of learning NR quality model using limited free-viewpoint data labeled with quality score, in the second step, a Bag-of-Distortion-Word codebook is then learned (separately after the first step) with the high-level features extracted from the discriminator. Finally, the quality model is trained using limited rendered views that are labeled with quality scores from the human. By doing so, we build the bridge between rendering and quality evaluation of the rendered views so that the latter process is more task-oriented and thus more robust. Overall, there are three main contributions:

- A novel quality framework that takes advantage of the characteristic of the GANs model. With this framework, large-scale common databases could be firstly used to train a *dis-occluded regions renderer*. After learning a BDW codebook based on the rendering-quality-related features extracted from the discriminator, perceptual quality could be then predicted using limited *free-viewpoint*

data with quality scores.
- A local non-uniform distortion region detection strategy is proposed based on the discriminator trained to distinguish rendered input from the real one.
- As a byproduct, the trained context *inpainter/renderer* (generator) could also be used in free-viewpoint framework for rendering dis-occluded regions.

II. Related Work

A. Commonly used image quality metrics

There are numerous image quality metrics [24] designed for evaluating the uniformly distributed distortions, for instance, blurriness and blockiness induced by different compression technologies. These metrics have achieved a big success, which show high consistency with human perception. However, most of the widely used quality metrics such as PSNR [25] fail to well predict the perceived quality of free-viewpoint synthesized/rendered views for the following reasons:

- Point-wise metrics like PSNR over-penalize the acceptable global uniform 'object shifting' artifact due to the mis-matched correspondences [26]. The global shifting of objects that may not be noticed by human observers could be penalized heavily by this type of metrics.
- Most of the existing quality metrics are not designed for local non-uniform artifacts. While observing an image, the artifacts located at the 'regions of interest' are more annoying than those located at inconspicuous areas [27]. The rendering related distortions are mainly located around these regions, which are easier to be perceived by human observers and prone to poor quality judgments. Meanwhile, regions of 'poor quality' are more likely to be perceived by humans with more severity. Thus, images with even a small number of 'poor quality' regions are penalized more gravely.
- In the wake of development in machine learning technologies, many quality assessment models have been proposed recently on the base of advanced deep learning schemes. However, most of them are proposed based on the assumption that the perceived quality of local regions is the same as the perceptual quality of the entire image [2], [28]–[34]. This assumption may work for images containing uniformly distributed distortions, but does not stand for those which contain non-uniformly distributed distortions, as shown in Fig. 2.

B. Quality metrics for free-viewpoint rendered views

To resolve the issues mentioned above, objective quality assessment metrics designed for rendered/synthesized views are developed. Among the existing FR metrics, one of the very first ones is View Synthesis Quality Assessment (VSQA) [35], which improves SSIM with three visibility maps by characterizing the complexity of the images. The 3DswIM is proposed by Battisti *et al.* [36] based on statistical features of wavelet sub-bands. Stanković [37] first employs morphological wavelet decomposition for quality assessment of synthesized images, namely MW-PSNR. Later, another metric, which devises PSNR with morphological pyramids decomposition (MP-PSNR), is proposed in [38]. Targeting the problem that global shifting artifacts are normally over-penalized by point-wise metrics, CT-IQM [39] is proposed using a context tree based encoding scheme. To quantify the change of contours' categories from a higher level, ST-IQM and ST-VQM are proposed in [40], [41] using *Sketch Token* descriptor. To quantify the deformations of curves in synthesized views, EM-IQM is proposed in [26] based on an elastic metric, which is extended for video in [42] based on motion trajectory. One of the most recent FR metrics is the LOGs [43] that considers both the geometric distortions as well as the sharpness of the images.

Nevertheless, since the references of the synthesized views are generally not available, NR metric is more desirable. Compared to FR metrics mentioned above, only a few NR metrics are designed for rendered virtual views. In [44], NIQSV is proposed based on a strong hypothesis that high-quality images consist of flat areas separated by edges. Later on, NIQSV+ is introduced in [45] to improve NIQSV by taking *textureless regions* into account. Recently, a novel NR quality metric APT is proposed in [11] using the auto-regression (AR) based local image description. Unfortunately, according to a recent subjective test reported in [46], even the best performing NR metric APT correlates poorly with the mean opinion score provided by human observers (with Pearson linear correlation coefficients of 0.422). Convolutional sparse coding was adopted in [47], [48] to quantify the non-natural local structure distortions. In [49], a blind video quality metric was proposed based on high-high wavelet subband based metric. Wang *et al.* has presented a blind metric in [50], where relevant distortions were measured in the discrete wavelet transform domain. Later, another metric was proposed by Gu *et al.* utilizing multiscale natural scene statistical analysis.

III. The Proposed Model

In this section, the proposed GANs based No-Reference quality Metric (GANs-NRM) is described. The diagram of the proposed method is depicted in Fig. 3. Notations of relevant variables and functions are summarized in TABLE I.

Step 1: To mitigate the issue of having limited publicly available multi-view contents, we design special masks that mimic *dis-occluded regions* or areas where distortions may arise, and train a GANs model to render the regions indicated by these masks. In such a manner, existing large-scale image datasets (*e.g., ImageNet* [51], *CIFAR10/100* [52], *PASCAL VOC* [53], *Places challenge* [54]) could be fully utilized. More importantly, the discriminator is thus capable of capturing the rendering-related quality information of the synthesized views.

Step 2: Limited databases with human quality labels to some extent restrict the usage of advanced deep learning based model to develop referenceless quality metric, especially for novel immersive or computer graphic applications. To provide a solution to this thorny issue, we take advantage of the rendering-related discriminator trained in the first step by using it to indicate possible distorted regions and generate rendering-related representations. A BDW codebook, where each item

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/TMM.2020.3038305, IEEE Transactions on Multimedia

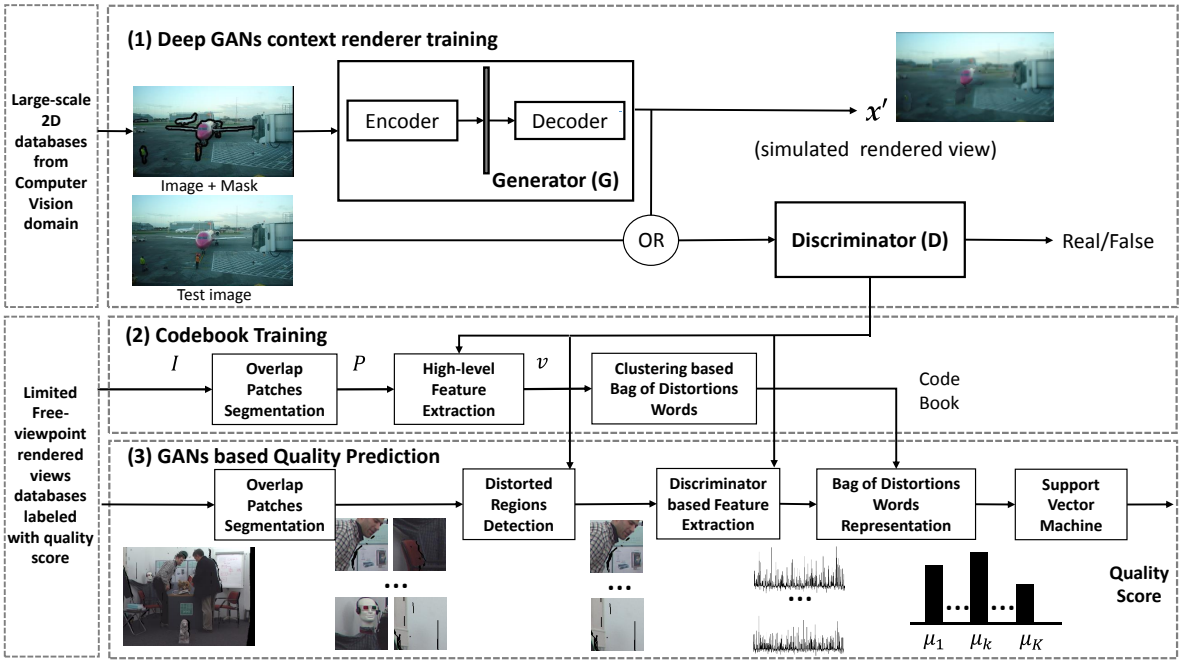


Fig. 3. Diagram of the proposed model: (1) Deep GANs context renderer pre-training; (2) Distortion codebook training; (3) Quality prediction.

TABLE I
SUMMARY OF NOTATIONS.

$\mathcal{L}_{rec}$	the content reconstruction loss
M	the binary mask indicating the missing regions need to be rendered
$\odot$	the element-wise multiplication
$G(\cdot)$	the generator
$\mathcal{L}_{adv}$	the adversarial loss
λ	the hyper-parameter balancing $\mathcal{L}_{rec}$ and $\mathcal{L}_{adv}$
$D(\cdot)$	the discriminator
P_i	the set of overlapping patches
p_{ij}	a patch within an image
n_p	the total number of patches
I	the dataset used for codebook training
l	the l_{th} layer of the discriminator
v_{ij}	the corresponding feature vector of a patch p_{ij} extracted from the l_{th} layer in the discriminator
K	the number of clusters gathered during codebook training
c_i	the i_{th} codeword (cluster) in the codebook
h_{adv}	the histogram representing each image
μ_i	the frequency of patches belong to the i_{th} cluster within an image
$\mathbf{1}(\cdot)$	an indicator function that equals to 1 if the specified binary clause is true
$D_{BS}(\cdot)$	the output of the last convolutional layer in the discriminator
ε	a threshold for poor-quality patches selection

represents a group of patches with a certain level of quality, is first learned based on high-level features extracted from the discriminator. Then, the quality predictor could be then trained by quantifying the number of 'good/bad quality' patches within the under-test rendered view with limited databases equipped with quality scores.

Our key novelty is to employ the trained discriminator in the target task (*e.g.* rendering) to obtain high-level, task-oriented, and quality-related representations for quality assessment. Details of each procedure are given in the following subsections.

A. Simulating the process of rendering dis-occluded regions using GANs

1) Semantic inpainting using GANs: Recent years, semantic inpainting/rendering is one of the hot research topics in the field of computer vision, where the goal is to render the missing regions within an image according to its semantics. Unlike traditional inpainting or texture synthesis methodologies, semantic inpainting [16]–[20] aims at filling the missing parts by using statistical information from external dataset instead of making only use of the internal property of the image needed to be rendered.

Among the existing state-of-the-art semantic context inpainters, the ones proposed in [17], [18] that are based on Generative Adversarial Networks (GANs) provide the most appealing performance. In both works, the proposed *context renderer* (generator) is designed as an auto-encoder with an unfilled image as input. In detail, a content reconstruction term is defined in Equation (1) to regress the missing parts to the ground truth content:

$$\mathcal{L}_{rec}(x) = |M \odot (x - G((1 - M) \odot x))|, \tag{1}$$

where M denotes the binary mask indicating the missing regions that need to be rendered; $\odot$ is the element-wise multiplication; $G(\cdot)$ indicates the generator. In this study, we employ L_1 norm instead of L_2 norm as done in [18] to overcome the blurry preference problem aroused by L_2 loss, *i.e.*, it tends to predict the mean of the distribution and thus results in an averaged blurrier image. Then, the adversarial loss

is introduced to jointly optimize both G and D as formalized in equation (2):

$$\mathcal{L}_{joint} = \lambda \mathcal{L}_{rec} + (1-\lambda)\mathcal{L}_{adv}, \tag{2}$$

where λ is a hyper-parameter balancing the weights between the two losses; $\mathcal{L}_{adv}$ is further defined in (3) by customizing GANs for the *context renderer* task with the mask M:

$$\mathcal{L}_{adv} = \max_{D} \mathbb{E}_{x\sim p_{\mathbf{x}}}[log(D(x)) + log(1 - D(G((1-M)\odot x)))], \tag{3}$$

where $D(\cdot)$ denotes the discriminator. In this paper, based on a similar recipe, we design our own masks M, which mimics the *dis-occluded regions* appearing during DIBR process, to train a new *context renderer*. Then, we explore to use the trained discriminator to evaluate the quality of *free-viewpoint* synthesized images.

2) Design of masks: *Dis-occluded/textureless regions/dark holes* are commonly introduced during the DIBR synthesis process, as introduced before. There are mainly two types of those regions that need to be filled: 1) edge-like regions that are located along the boundaries of the foreground objects as shown in Fig. 4a, and 2) small or medium size of textureless regions (introduced by inaccurate depth map/occlusion edges predictions in DIBR framework) that are distributed throughout the entire images, as shown in Fig. 4b. The shapes of these regions are normally related to the shapes of the objects. These regions can be filled with certain inpainting/rendering algorithms. However, inpainting-related artifacts may also be introduced.

Generally, *dis-occluded regions* located along the border between the foreground and the background are challenging for existing inpainting/rendering algorithms. It is often to see that foreground regions are rendered with background pixels or vice versa. As a result, the structures of objects are disrupted. Structure-related degradations around foreground objects, accompanying with inter-view inconsistency on depth, might then cause binocular rivalry, binocular suppression, or binocular superposition [55], [56], which eventually lead to visual discomfort. Concerning the issues above, and to train a new *context renderer* that is capable of better rendering the missing regions mentioned above (similar to the dis-occluded regions shown in Fig. 4a), two types of masks M are designed:

(a) (b)

Fig. 4. Examples of typical dis-occluded/missing regions introduced during the process of DIBR based views synthesis. (a) Examples of dis-occluded regions that are around foreground objects' boundaries (bounded by green boxes); (b) Examples of small and medium size of textureless regions (bounded by red and blue bounding boxes correspondingly) that distributed throughout the image.

- **Mask I**: to mimic the dis-occluded regions around foreground objects' boundaries. The mask is designed as the dilated object boundaries. An example is shown in Fig.5c.
- **Mask II**: to mimic the shifted objects' boundaries in the synthesized views induced by compression on depth map [12] or inaccurate geometric reconstruction [57], we generate the second type of masks by simply shifting the first type of mask with certain pixels as shown in Fig. 5d.

Generally, it is easier to inpaint/render smooth regions with homogeneous textures than the complicated regions with non-homogeneous textures as the context around a smoother region is more 'copyable'. If one wants to train a more powerful *context renderer*, the selected masks should contain contents/structures that can not be replicated from the surroundings. In addition, *dis-occluded or textureless regions* in a virtual view are generally disconnected, and the shapes of these regions are always related to the foreground objects (*i.e.*, related to the depth map). With these two concerns, the third mask is proposed:

- **Mask III**: The SLIC super-pixels algorithm [58] is used to select regions where masks should be located for later training. More specifically, an image is first segmented into a set of super-pixels, as shown in Fig. 6a and 6d. Then, **two mask sizes are considered**. Super-pixels that occupy less than 0.05% of the whole image area are considered as small-size mask (similar to the red box region shown in Fig.4b), while super-pixels occupy 0.1% to 0.5% of the whole image area are considered as medium-size mask (similar to the blue box region in Fig.4b). Examples are presented in Fig. 6. By doing so, 1) the selected masks are separately distributed in the entire image; 2) the shape of the masks are related to contents of the images, *e.g.*, objects.

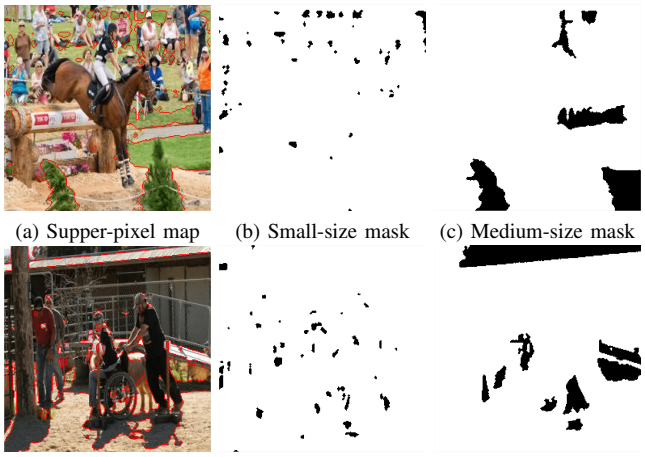
(a) Image (b) Labeled map (c) Mask I (d) Mask II

Fig. 5. Example of simulation procedure for dis-occluded regions by mask I and II in our training set.

B. Bag-of-Distortion-Word codebook learning with pre-trained discriminator

As discussed before, the discriminator serves as an indicator telling whether an input is well inpainted or not. Thus the output of the discriminator is related to the quality of the patch. Therefore, it is reasonable to hypothesize that the intermediate output of D is strongly related to rendering-related distortions, which affect the perceived quality significantly. Based on this hypothesis, we propose to use the discriminator to get a latent

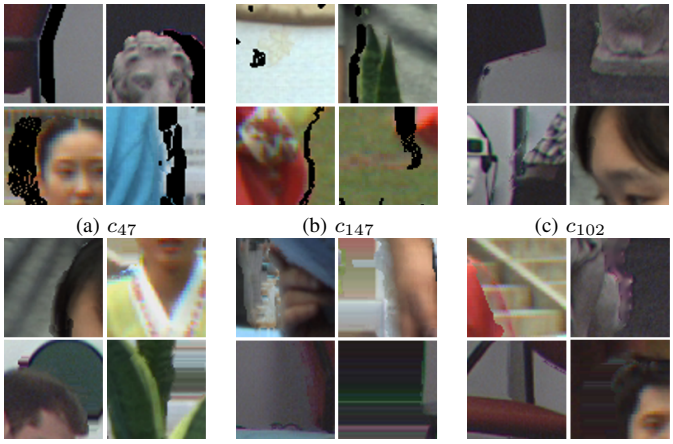
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3038305, IEEE Transactions on Multimedia



(a) Supper-pixel map (b) Small-size mask (c) Medium-size mask

(d) Supper-pixel map (e) Small-size mask (f) Medium-size mask

Fig. 6. Example of images in the training set and the corresponding designed mask III. Two mask sizes are considered.



(a) c_{47} (b) c_{147} (c) c_{102}

(d) c_{121} (e) c_{05} (f) c_{98}

Fig. 7. Selected 'Words' in the learned BDW Codebook.

codebook with 'codewords' that represent different types of distortions. With this codebook, a higher-level representation could be obtained for each image. Details are illustrated below.

To predict the image level quality by considering local distortions, the image needed to be processed locally. Therefore, a set of multiple overlapping patches $P_i = \{p_{ij} | j = 1, \cdots, n_p\}$, where n_p is the total number of patches, is used to represent the image x_i as done in [30]. In this study, the overlap size is selected as half of the patch size, and the patches are sampled over the whole image (along both the horizontal and vertical direction) to maintain as much structural information as possible. Afterwards, with the pre-trained GANs model, these patches are fed into the adversarial discriminator to extract higher-level features for later patches categorizations. For each patch p_{ij} in the entire dataset I for codebook training, its corresponding feature vector v_{ij} is extracted from the l_{th} layer in the discriminator as:

$$v_{ij} = D(p_{ij}, l). \tag{4}$$

In this study, the feature vector is extracted from the last second convolutional layer of the discriminator (*i.e.,* the feature maps of $conv4$ for D_1, $conv5$ for D_2 and D_3 as depicted in Table II. Details are shown in Section III-C). Finally, $m \times n$ vectors can be obtained for the m images in the codebook training set I.

With the set of extracted features in correspondence to their patches, now we want to look for a new representation of the entire image to link the local information with the entire image quality.

Intuitively, the idea is to categorize image patches into different clusters that can be representatives of perceived quality, so the quality of the tested image can be quantified by checking how many 'good' or 'poor' patches it contains. Formally, the $m \times n$ patches v_{ij}, $i = \{1, \cdots, m\}$, $j = \{1, \cdots, n\}$ are reshaped to $v_o, o = \{1, \cdots, n \times m\}$. Then the v are clustered into K clusters $\{c_1, \cdots, c_K\}$ using an advanced clustering algorithm [59] (a fast nearest neighbor algorithm robust in high dimensional vectors matching) with the dimensionality reduction algorithm proposed in [60]. Selected cluster results are shown in Fig. 7. It can be observed that patches with similar type of distortions are gathered in the same cluster as a 'distortion word'. For example, both of the cluster c_{47} and c_{147} consist of patches with *dark holes*, and the ones in cluster c_{47} are obviously larger than that of c_{147}, which indicates worse quality. The distortions of c_{102} is imperceivable (guarantee good quality), while c_{121}, c_{05} and c_{98} are with more obvious rendering-related artifacts. Naturally, different 'codeword' in the clustered 'codebook' actually represents a certain level of quality with respect to the types and magnitudes of distortions, which is in consistent with our hypothesis. Based on this observation, in this study, the learned codebook is named after 'Bag-of-Distortion-Words (BDW)'. With the BDW codebook, each image x_i can then be encoded as a histogram $h_{adv}(i) = \{\mu_{i1}, \cdots, \mu_{iK}\}$, where each μ_{ik} is defined as

$$\mu_{ik} = \frac{\sum_{j=1}^{n_p} \mathbf{1}(p_{ij} \subset c_k)}{n_p}, \tag{5}$$

where $\mathbf{1}(c)$ is an indicator function that equals to 1 if the specified binary clause c is true, and n_p is the number of patches within the image. An intuitive interpretation of this BDW based representation of the image is that the histogram statistically quantifies how many 'good quality' and 'poor quality' patches that a synthesized image has. As local significant rendering distortion is more annoying than the global uniform one, this new representation is a higher-level quality descriptor, which can indirectly predict the overall quality of one image. During clustering, K is an important parameter that will have an impact on the final performance. Therefore, further discussion about the selection of K is given in Section IV-C1.

C. Local distortion regions selection

Generally, artifacts located in *regions of interest* are much more annoying than those located at an inconspicuous area [27]. In our case, 'poor' quality regions (*e.g.*, *dark holes* or regions with rendering artifacts) are generally along the *regions of interest* (such as the foreground object), thus, they

are more likely to attract observers than the 'good' ones. Therefore, images with even a small number of 'poor' regions are penalized more gravely by the observers. Accordingly, it is reasonable to do the same penalization in the objective model as well.

Moreover, as discussed before, the discriminator is trained to distinguish artificially generated picture (rendered images in this case) from the real one. Therefore, a well-trained discriminator is supposed to be able to indicate poor rendered regions. The output of the discriminator D is a boolean value indicating whether the input patch p_{ij} is rendered or not, where '1' for real patches and '0' for generated patches. It is intuitive to hypothesize that patches assigned with '0' by the discriminator are those with poorer quality. Hence, the discriminator is further utilized as a 'poor' quality patches selector. As thus, Equation (5) could be modified to:

$$\mu_{ik} = \frac{\sum_{j=1}^{n_p} \mathbf{1}(p_{ij} \subset c_k) \cdot XOR(D(p_{ij}), 1)}{n_p}, \qquad (6)$$

where $D(\cdot)$ is the direct boolean output of the discriminator trained to render simulated problematic regions when taking a patch p_{ij} as the input. $XOR(\cdot)$ is the exclusive OR operation, $XOR(D(p_{ij}), 1)$ equals to 1 if $D(p_{ij}) = 0$.

Apart from using the final boolean output of the discriminator for selecting the possible poorly rendered regions, another possibility is to use the one dimension output just before the final sigmoid layer (*i.e.*, the feature map of layer $conv5$ in D_1 or $conv6$ in D_2, D_3 as depicted in Table II) with normalization. The rationale behind this is that the intermediate feature layer before the sigmoid activation layer provides more informative inter-class knowledge compared to the final boolean prediction, which is helpful to learn a finer-grained quality predictor. To do this, the output of the last convolutional layer of the discriminator for all the training patches $p_{ij}, i = \{1, \cdots, m\}, j = \{1, \cdots, n\}$ are collected and normalized into a range of $[0, 1]$. After the normalization, the output of the last convolutional layer serves as a probability value indicating whether the test patch is rendered or not. A smaller value represents a higher probability that this patch is rendered with a greater magnitude of distortions. Afterwards, patches that are with a certain magnitude of rendering-related distortions can be selected according to a certain threshold ε, meaning that only poorly rendered regions with certain low-quality level are selected for the final quality decision. By doing so, Equation (6) could be further rewritten as:

$$\mu_{ik} = \frac{\sum_{j=1}^{n_p} \mathbf{1}(p_{ij} \subset c_k) \cdot \mathbf{1}(D_{BS}(p_{ij}) < \varepsilon)}{n_p}, \qquad (7)$$

where $D_{BS}(\cdot)$ means that we only consider the output of the last convolutional layer in the discriminator with a patch p_{ij} as input. ε is a threshold for poor-quality patches selection. The setting of threshold ε is discussed in Section IV-C4.

D. Final quality prediction

After extracting the histogram h_{adv}, Support Vector Regression (SVR) is then applied on it with a linear kernel to predict the final quality score. In the experiment, the entire database is divided into different sets:

- **Validation set**. 20% of the whole dataset is used as the validation set. It is used for model parameters selection (*e.g.*, codebook training).
- **Training-Testing set**. The remaining 80% of the whole dataset is used as Training-Testing set. During the performance evaluation procedure, 1000-fold cross-validation is applied. For each fold, the Training-Testing dataset is further randomly split into 80% for SVR training and 20% for testing, with no overlap between them [61] (no same viewpoint of the same content).

The median Pearsons Correlation Coefficient (PCC), Spearman rank order Correlation Coefficient (SCC), and Root Mean Square Error (RMSE) between subjective and objective scores are reported across the 1000 runs for performance evaluation. Higher PCC, SCC, and lower RMSE values indicate better performance.

IV. Experiment and Result

A. Training of GANs

1) Training data: To generate a new dataset with the three masks mentioned in Section III-A, we collected images from the *PASCAL VOC 2012* [53] and the *Places* database [54]. There are in total 10K training images are used to train the GANs in this study.

- **PASCAL VOC 2012 database**: The original objective of this database is for a challenge to recognize objects from a number of visual object classes in realistic scenes. It contains thousands of images with different categories, which diverse from people, animals to vehicles, and indoor scenes. One of the merits of this database is that it provides us with images of pixel-wise segmentation labels (around 3K images), which gives the boundary of 'objects' against the 'background' label. An example is given in Fig. 5a and Fig. 5b. In our study, we utilize this segmentation label to generate mask I and mask II mentioned above, which leads to 6K training data.
- **Places database**: To have a balanced dataset with the mask I and II, the validation set from the 'Places Challenge 2017', which contains around 2K images, are selected as a part of the training set in this study with mask III mentioned above. This dataset contains images with diverse contents, which vary from outdoor landscapes, cities views to indoor people portrait images. As there are two mask sizes in Mask III, this leads to 4K training images.

2) Training process: The framework of the *context renderer* is implemented based on the pipeline developed by Pathak *et al.* [18] with Caffe and Torch packages. The commonly used stochastic gradient descent method Adam [62] is used for optimization. We start with a learning rate of 2×10^{-4}, as done in DCGAN [63]. In our experiment, the impact of the trade-off between the content loss and the adversarial loss, *i.e.*, different λ in Equation (2), on the performance of the proposed metric has been tested (please refer to Section IV-C3 for more details).

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/TMM.2020.3038305, IEEE Transactions on Multimedia

Furthermore, since the main focus of this section is to explore the discriminator for quality assessment of synthesized views with local non-uniform distortions, different architectures of the discriminator have been proposed and tested. Details are summarized in Table II. The main difference between D_1 and the other two architectures is the size of images that can be fed into. D_3 is of fewer parameters than D_2 and D_1, where the number of convolutional kernels is halved in each layer. With such a design, we could check how the input size and complexity of the discriminator influence the performance of the proposed scheme.

B. Free-viewpoint image quality databases

The performance of the proposed GANs-NRM is evaluated on the IRCCyN/IVC DIBR [12] and the IETR [46] databases. Images from the IRCCyN/IVC DIBR database are obtained from three *free-viewpoint* sequences. Seven DIBR synthesis algorithms labeled with A1-A7 [6]–[10], [64] are used to process the three sequences to generate four new virtual views for each of them. The database is composed of 84 synthesized views and 12 original frames extracted from the corresponding sequences. Similarly, in the IETR database, images are generated from ten *free-viewpoint* sequences using eight rendering algorithms including the *Criminisi* [65], *Luo* [66], *HHF-v2* [67], *LDI* [68], *VSRS* with single view based mode [69], *Ahn* [70], *VSRS* with both interview mode [69], and the *Zhu* [71] algorithms. They are denoted as $R_1 - R_8$ in this paper. This database consists of 140 synthesized and 10 original images. Both of the databases are equipped with subjective quality scores in terms of Mean Opinion Score (MOS) or Differential MOS (DMOS).

Data augmentation is conducted to provide a more robust performance evaluation by rotating each image $90°$, $180°$, and $270°$ counterclockwise successively. Unlike other data augmentation methodology (such as scaling), rotation operation does not introduce new distortion. We thus assume that the qualities of the augmented images remain unchanged. The performance evaluation procedure is conducted according to [61] as described in section III-D. To train the BDW codebook, 20 % of the augmented data are utilized as the validation set, which contains around 1.5×10^4 patches of size 64×64 for the IRCCyN/IVC DIBR database, and 12×10^4 patches for IETR database.

C. Performance dependency on hyper parameters

In this work, there are in total four hyper-parameters, including K in BQW, different discriminator architectures D, solver hyper-parameter λ, and threshold ϵ in Equation (7). To find the optimal solution, the grid search method is utilized. In the following section, how these parameters influence the performances of the proposed model is evaluated. When evaluating the influence of one parameter, all the other parameters are fixed with the optimal solutions.

1) Number of 'Distortion Word' K in BDW: To check if the performance of the proposed GANs-NRM is sensitive to the cluster number K, different K are tested on the validation set. The results on the two databases are shown in Fig. 8a. It can be observed that the performances of GANs-NRM on the two databases (in terms of PCC) have similar trends, which rise gradually along with the increase of K at the beginning and then drop gradually after peaking (when K is in the range of $[150, 160]$). Thus, in this study, we set $K = 160$ for IVC database and $K = 150$ for IETR.

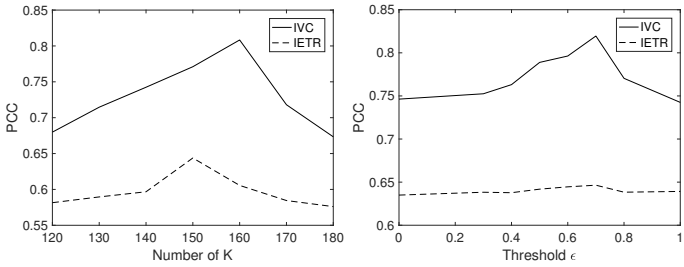


Fig. 8. Performance dependency of GANs-NRM on K and ε.

2) Different Discriminator architecture: The performances of the proposed model equipped with different discriminator architectures, which are described in Table II, are evaluated on the IRCCyN/IVC DIBR database and reported in Table III. It is found that, with any chosen λ, the proposed method always attains better PCC value with architecture D_1 than with D_2 or D_3. In the proposed model, we finally choose architecture D_1 for the discriminator.

3) Solver hyper-parameter λ: As introduced in [18], [63], the solver hyper-parameter λ in equation (2) is suggested to be set as 0.999. It is a tunable parameter balancing the reconstruction loss and the adversarial loss during training. Since the discriminator is utilized for both distortion regions selection and higher level feature extraction in this study, higher weights for the adversarial loss is tested, *i.e.* lower λ in Equation (2). The performances of the proposed model with different λ on the IRCCyN/IVC DIBR database are reported in Table III. By comparing the performances, interestingly, it is found that the PCC increases when λ increases from 0.5 to 0.9, and drops when $\lambda = 0.999$. In this study, we set $\lambda = 0.9$.

4) Threshold ε : The influence of the threshold ε on the performance of GANs-NRM is illustrated in Fig. 8b. The performance of using a threshold in Equation (7) for the selection of distortion regions is better than using the direct output of the discriminator. The performances climb with an increasing ε until they reach to 0.7 then decline on both IRCCyN/IVC and IETR databases. Therefore, in our model, we set $\varepsilon = 0.7$.

D. Overall quality prediction performance

The performance of the proposed metric is compared with the state-of-the-art metrics designed for rendered views summarized in Section II-B, including FR metrics: *MP-PSNR*$_{red}$ [72], *MW-PSNR*$_{red}$ [72], *ST-IQM* [40], *EM-IQM* [26], *LoGs* [43]; and the NR metrics: *NIQSV* [44], *NIQSV+* [45], *APT* [11]. For fair comparisons, the median performances of the compared metrics are also reported under a 1000-fold cross-validation.

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/TMM.2020.3038305, IEEE Transactions on Multimedia

TABLE II
DIFFERENT DISCRIMINATOR ARCHITECTURES TESTED IN THIS STUDY, In IS THE INPUT OF EACH LAYER, $InSize$ IS THE INPUT SIZE OF EACH LAYER, k IS THE KERNEL SIZE, **S** IS THE STRIDE, $OutL$ IS THE OUTPUT CHANNELS FOR EACH LAYER AND Act IS THE ACTIVATION FUNCTION OF EACH LAYER.

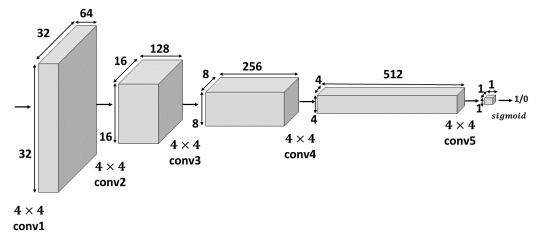
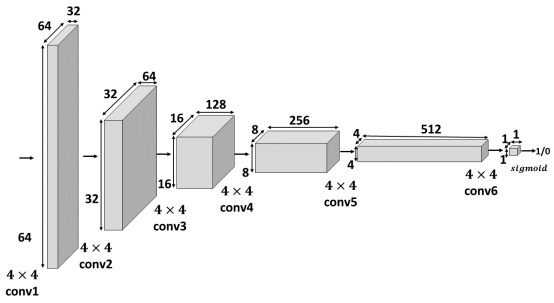
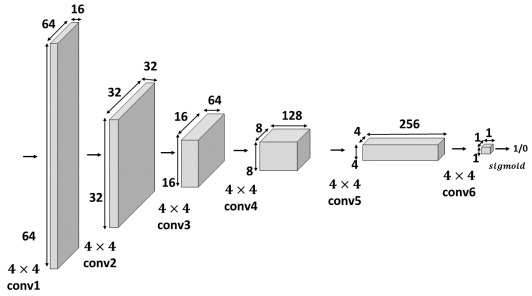
Layer	In	InSize	k	s	OutL	Act	Visualization
Discriminator architecture D_1							
conv_1	image	64×64	4	1	64	Leaky ReLU	
conv_2	conv_1	32×32	4	1	128	Leaky ReLU	
conv_3	conv_2	16×16	4	1	256	Leaky ReLU	
conv_4	conv_3	8×8	4	1	512	Leaky ReLU	
conv_5	conv_4	4×4	4	1	1	Sig moid	
Discriminator architecture D_2							
conv_1	image	128×128	4	1	32	Leaky ReLU	
conv_2	conv_1	64×64	4	1	64	Leaky ReLU	
conv_3	conv_2	32×32	4	1	128	Leaky ReLU	
conv_4	conv_3	16×16	4	1	256	Leaky ReLU	
conv_5	conv_4	8×8	4	1	512	Leaky ReLU	
conv_6	conv_5	4×4	4	1	1	Sig moid	
Discriminator architecture D_3							
conv_1	image	128×128	4	1	16	Leaky ReLU	
conv_2	conv_1	64×64	4	1	32	Leaky ReLU	
conv_3	conv_2	32×32	4	1	64	Leaky ReLU	
conv_4	conv_3	16×16	4	1	128	Leaky ReLU	
conv_5	conv_4	8×8	4	1	256	Leaky ReLU	
conv_6	conv_5	4×4	4	1	1	Sig moid	

TABLE III
PERFORMANCE DEPENDENCY OF PROPOSED METRIC WITH DIFFERENT SOLVER HYPER-PARAMETERS λ.

PCC	$\lambda = 0.5$	$\lambda = 0.9$	$\lambda = 0.999$
D_1	0.7802	**0.8083**	0.7821
D_2	0.7377	0.7536	0.7339
D_3	0.7266	0.7280	0.7273

The performance results on the two aforementioned databases are summarized in Table IV. Among the FR metrics, ST-IQM performs the best on the IVC-DIBR database, when LoGs performs the best on the IETR database in terms of PCC, SCC, and RMSE. Among the NR metrics, our proposed method attains the best performance, and it is comparable to the best performing FR metrics. The gain of GANs-NRM compared to the second best NR metric *APT* is 17% in terms of PCC on IVC-DIBR, and 35% on IETR database. As discussed in Section I that reference views are barely available in practical cases, and since the proposed metric obtains comparable performances compared to the best performing FR metrics, in the following analyses, we focus mainly on NR metrics.[1]

The scatter plots of all the tested NR quality metrics are provided in Fig. 9. We can notice that most of the objective scores predicted by the proposed metric are better distributed

[1]Experimental results/analyses compared to FR metrics are shown in the supplemental materials.

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/TMM.2020.3038305, IEEE Transactions on Multimedia

TABLE IV
PERFORMANCE OF QUALITY METRICS DESIGNED FOR RENDERED VIEWS.

	IVC-DIBR Image dataset			IETR Image dataset		
	PCC	**SCC**	**RMSE**	**PCC**	**SCC**	**RMSE**
Full Reference Metric						
MP-PSNR$_{red}$	0.748	0.701	0.414	0.556	0.494	0.219
MW-PSNR$_{red}$	0.740	0.684	0.424	0.549	0.473	0.223
ST-IQM	**0.846**	**0.768**	**0.341**	0.506	0.401	0.237
EM-IQM	0.759	0.710	0.403	0.541	0.462	0.231
LoGs	0.832	0.727	0.358	**0.752**	**0.686**	**0.178**
No Reference Metric						
NIQSV	0.687	0.508	0.460	0.323	0.307	0.253
NIQSV+	0.701	0.515	0.455	0.338	0.323	0.252
APT	0.704	0.729	0.431	0.476	0.496	0.237
GANs-NRM	**0.826**	**0.807**	**0.386**	**0.646**	**0.571**	**0.198**

along the diagonal of the plot than others. In the scatter plot of APT, NIQSV, and NIQSV+, images that synthesized using the same rendering algorithm are predicted with similar objective scores leading to a 'vertical line' as shown in the corresponding figures.

To verify the generalization capability of the proposed metric, cross-database tests are conducted. GANs-NRM is trained on one database and then tested on the another one. As no cross-validation evaluation is conducted, performances of metrics on the entire dataset are reported. Results are presented in Table V. Although the performances of GANs-NRMs drops slightly, it still outperforms the compared state-of-the-art NR metrics designed for synthesized views.

TABLE V
CROSS-DATASET EVALUATION.

	train IETR/test IVC			train IVC/test IETR		
	PCC	SCC	RMSE	PCC	SCC	RMSE
NIQSV	0.634	0.616	0.514	0.175	0.245	0.147
NIQSV+	0.711	0.666	0.467	0.209	0.219	0.242
APT	0.730	0.715	0.576	0.422	0.418	0.225
GANs-NRM	**0.794**	**0.772**	**0.410**	**0.601**	**0.552**	**0.205**

In order to examine the significance of the performances between each two tested quality metrics, Student's t-test is conducted. More specifically, the 1000 PCC values obtained during the cross performance evaluation described in Section IV-D of each tested metric are used as input for the t-test. The significant results on the two databases are concluded in Table VI with a significance level of 0.05, where '1' represents the performance of the under-test metric in row outperforms the one in column significantly, '-1' represents the inverse situation, and '0' represents there is no significant difference. According to the table, the performance of the proposed GANs-NRM is significantly better than all the other NR metrics on the two databases.

Another important application of an objective metric in *free-viewpoint* system is to benchmark different synthesized algorithms, with respect to the ground truth ranking of the rendering algorithms (*i.e.,* A1-A7 and R1-R8 utilized in the two databases). In our study, the ground truths are obtained by averaging the DMOS of each rendering algorithm. The rankings predicted by objective metrics are reported in Table VIII. According to Table VIII, the ranking of the proposed GANs-NRM is the most consistent. For GANs-NRM, only the rankings of A4, A5 on IRCCyN/IVC database and R1, R2 on IETR database are switched, which generate similar quality synthesized images. Therefore, a trustable rank order could be provided by the proposed model to select proper synthesis algorithms.

TABLE VI
SIGNIFICANT TEST RESULTS ON **IRCCYN/IETR** DATABASES.

Metric	NIQSV	NIQSV+	APT	GANs-NRM
NIQSV	-	0/0	0/-1	-1/-1
NIQSV+	0/0	-	0/-1	-1/-1
APT	0/1	0/1	-	-1/-1
GANs-NRM	1/1	1/1	1/1	-

TABLE VII
NORMALIZED EXECUTION TIME OF EACH NR METRIC.

Metric	NIQSV	NIQSV+	GANs-NRM	APT
Time	18	21	157	13k+

TABLE VIII
RANKING OF RENDERING ALGORITHMS (DESCEND ORDER).

	IRCCyN/IVC DIBR database							
DMOS	**A1**	**A5**	**A4**	**A6**	**A2**	**A3**	**A7**	-
NIQSV	A1	A4	A5	A2	A6	A3	A7	-
NIQSV+	A1	A6	A5	A4	A2	A3	A7	-
APT	A1	A2	A4	A3	A5	A6	A7	-
GANs-NRM	**A1**	**A4**	**A5**	**A6**	**A2**	**A3**	**A7**	-
	IETR database							
DMOS	**R7**	**R8**	**R5**	**R6**	**R3**	**R1**	**R2**	**R4**
NIASV	R7	R8	R3	R5	R1	R2	R6	R4
NIASV+	R7	R8	R1	R2	R6	R3	R5	R4
APT	R8	R7	R3	R5	R1	R2	R6	R4
GANs-NRM	**R7**	**R8**	**R5**	**R6**	**R3**	**R2**	**R1**	**R4**

Last but not least, to evaluate of the quality of synthesized view feasible in real applications, the time cost of the quality assessment metric should be reasonable, if possible, the lower, the better. To verify the efficiency of the proposed metric, as well as make comparisons with others, the execution time normalized by the run time of PSNR is computed [45]. By calculating the normalized execution time, it is then possible to compare the time complexities of different metrics on different machines and datasets. For a given image x from a database, the normalized execution time t_{norm} is defined as

$$t_{norm} = \frac{t_{metric}}{t_{PSNR}}, \tag{8}$$

where t_{metric} is the execution time of the objective quality metric for image x, and t_{PSNR} is the corresponding runtime of PSNR. For completeness, in our study, the experiments are conducted on a desktop equipped with i7 CPU (4GHz), 8 GB RAM, and an Nvidia Xeon E3-1200 v3/4th. The runtime of PSNR for one synthesized image in IRCCyN/IVC DIBR images database is 0.05 seconds. The normalized execution time for each metric is summarized in Table VII. Although GANs-NRM is slower than NIQSV, it is still within a reasonable time

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/TMM.2020.3038305, IEEE Transactions on Multimedia

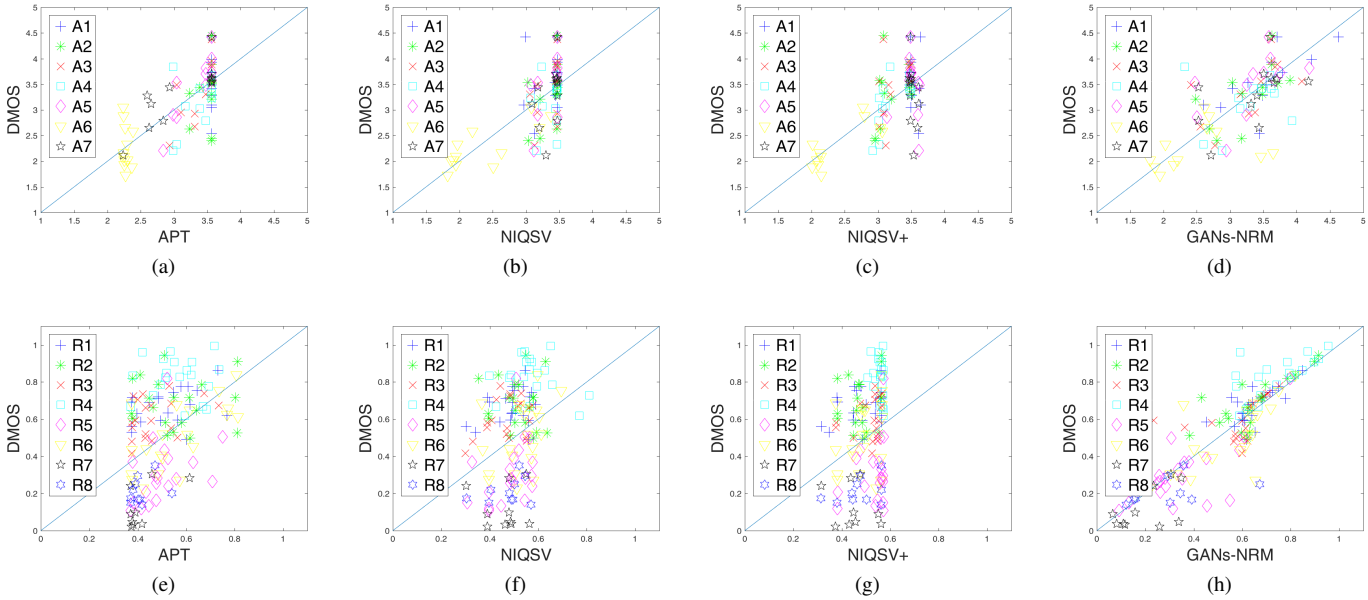


Fig. 9. Scatter plots of objective quality scores versus DMOS on IRCCyN/IVC DIBR (the first row) and IETR (the second row) databases. A1-A7 represent different DIBR algorithms tested in [12], while R1-R8 represent different DIBR algorithms tested in [46]. It has to be mentioned that DMOS provided by IRCCyN/IVC DIBR is computed according to [73], while the one in IETR is calculated as described in [46].

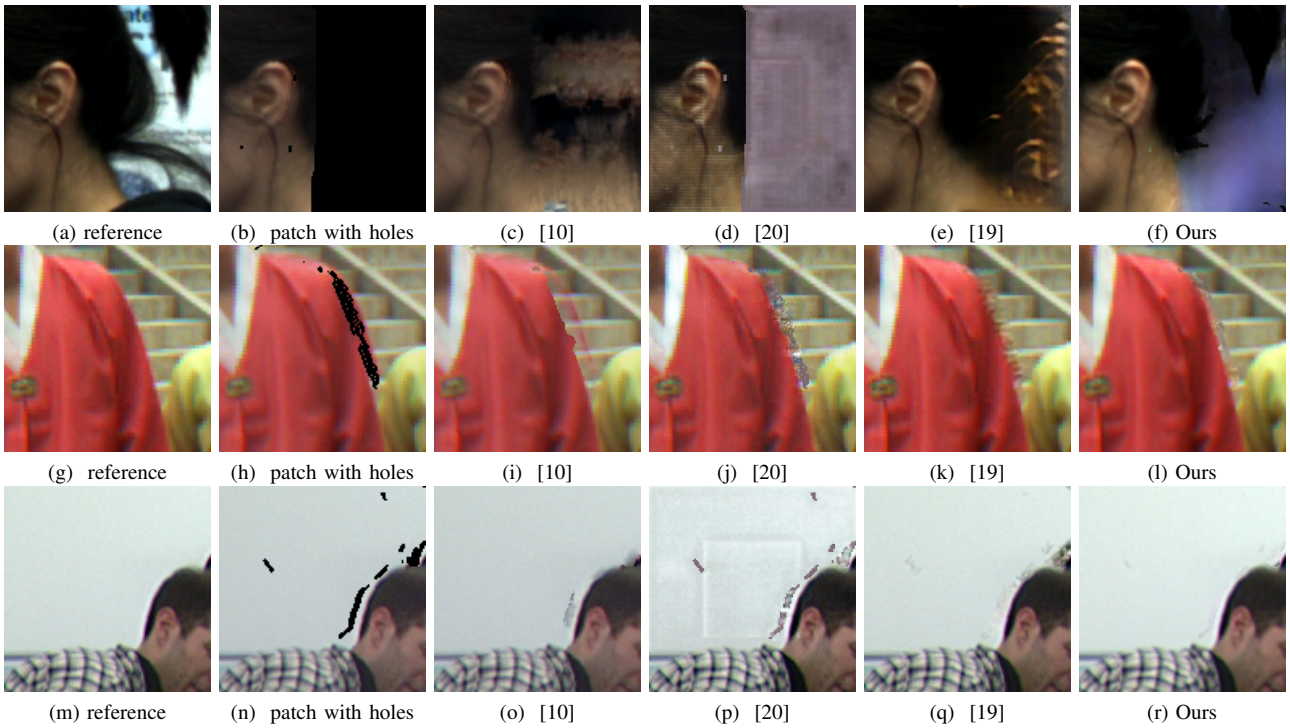


Fig. 10. Rendered results: 1st column: reference image; 2nd column: patches with dis-occluded regions (holes); From the 3rd to 5th column, rendered results using algorithms proposed in [10], [19], [20], respectively; the last column: our results.

cost and much faster than the second best NR metric APT.

E. Inpainting/rendering results

In this paper, the generator/discriminator are trained simultaneously to render/evaluate the *dis-occluded regions* within rendered viewpoints. The performance of utilizing discriminator to predict the quality of the rendered views has been demonstrated in the previous section. As a byproduct of this paper, it would be interesting to evaluate the performance of our context renderer (generator) in real rendering cases. Its performance of rendering *dis-occluded regions* (back holes) on IRCCyN/IVC DIBR images database is reported.

One typical DIBR inpainting algorithms [10] (where depth information is used) and two state-of-the-art semantic inpainting/rendering algorithms [19], [20] are used for comparison. Due to the limitation of space, selected results[2] are shown in Fig.10. Based on the results, it is observed that 1) By comparing our inpainted results in Fig. 10f to the others with respect to the reference, the shape of the braid of the girl is better remained by our model; 2) The shape of the dis-occluded regions in Fig. 10h are better inpainted by the proposed models as shown in Fig. 10l. There are obvious 'double-edge like' shapes, *i.e.* ghosting artifacts, along the objects after being inpainted by other methods; 3) In the condition that holes appear in homogeneous texture regions, which are also close to the borders of foreground objects, our inpainted result is with higher texture consistency than the others as shown in Fig. 10r.

In conclusion, our proposed *context renderer* could maintain the structures of the *dis-occluded/textureless regions*, especially when those regions are large. For the challenging *dis-occluded regions* that lie on the border of foregrounds and backgrounds, as well as in the homogeneous texture regions close to the border of foreground objects, the proposed renderer performs better than the others.

The appealing performance of our *context renderer* (generator) validates the effectiveness of the proposed training strategy, where specifically designed masks are used to mimic the typical textureless regions induced in the rendering process. The proposed strategy is more flexible in using the large-scale image databases in the computer vision domain rather than limited *free-viewpoint* contents. It should also be noted that the training data scale in our study is only 10K, which could be definitely further augmented by further employing the existing datasets. Therefore, there is still an improvement space for our current trained model, no matter for quality assessment or for textureless regions rendering in free-viewpoint applications.

V. Future work

As a first attempt to build the bridge between the task of generating synthetic contents and quality assessment of the generated contents, we apply the proposed model for quality assessment of virtual views rendering in free-viewpoint applications. With the popularity of generative models in immersive multimedia and computer graphics applications, it would be worthy of employing the proposed framework for the quality control of target services and the benchmarking of relative techniques. For instance, the GANs model proposed in [74] for light field synthesis could be adopted with the proposed framework using the light field quality dataset released from [75].

[2]More inpainting/rendering results are shown in the supplemental materials.

VI. Conclusions

In this work, we propose a GANs-based NR quality metric to evaluate the perceptual quality of *free-viewpoint* rendered views based on the assumption that if a generator of a GANs could be trained to render the *dis-occluded/textureless regions*, then the discriminator could be used to predict the quality. To resolve the challenges of the training data scales in DNNs, a novel strategy is proposed, which exploits the current existing large-scale 2D computer vision datasets. The spirit of the strategy can be easily applied to other applications in immersive multimedia, computer graphics domain, or even other community. By exploring the intermediate output of the discriminator, we learned a Bag-of-Distortions-Word codebook, proposed a local distortion region selector, and eventually mapped the non-uniform inpainting related artifacts to perceptual quality with limited available quality databases for *free-viewpoint* rendered views. According to experimental results, the proposed GANs-NRM provides the best performance compared to the state-of-the-art NR quality metrics for synthesized views. As a byproduct, the *context renderer* also shows appealing performances in rendering the challenging *dis-occluded/textureless* holes in the virtual views.

References

[1] C. Fehn, "A 3d-tv approach using depth-image-based rendering (dibr)," in *Proc. of VIIP*, vol. 3, no. 3, 2003.
[2] W. Zhou, Z. Chen, and W. Li, "Dual-stream interactive networks for no-reference stereoscopic image quality assessment," *IEEE Transactions on Image Processing*, 2019.
[3] M. Tanimoto, M. P. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint tv," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 67–76, 2011.
[4] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn, "Towards foveated rendering for gaze-tracked virtual reality," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, p. 179, 2016.
[5] W. Zhou, L. Shi, Z. Chen, and J. Zhang, "Tensor oriented no-reference light field image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4070–4084, 2020.
[6] C. Fehn, "Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv," in *Electronic Imaging 2004*. International Society for Optics and Photonics, 2004, pp. 93–104.
[7] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, "View generation with 3d warping using depth information for ftv," *Signal Processing: Image Communication*, vol. 24, no. 1, pp. 65–72, 2009.
[8] K. Mueller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "View synthesis for advanced 3d video systems," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 1–11, 2009.
[9] P. Ndjiki-Nya, M. Koppel, D. Doshkov, H. Lakshman, P. Merkle, K. Muller, and T. Wiegand, "Depth image-based rendering with advanced texture synthesis for 3-d video," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 453–465, 2011.
[10] M. Köppel, P. Ndjiki-Nya, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, and T. Wiegand, "Temporally consistent handling of disocclusions with texture synthesis for depth-image-based rendering," in *2010 IEEE International Conference on Image Processing*. IEEE, 2010, pp. 1809–1812.
[11] K. Gu, V. Jakhetiya, J.-F. Qiao, X. Li, W. Lin, and D. Thalmann, "Model-based referenceless quality metric of 3d synthesized images using local image description," *IEEE Transactions on Image Processing*, 2017.

[12] 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 Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1332–1343, 2011.

[13] K. Wolski, D. Giunchi, N. Ye, P. Didyk, K. Myszkowski, R. Mantiuk, H.-P. Seidel, A. Steed, and R. K. Mantiuk, "Dataset and metrics for predicting local visible differences," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 5, p. 172, 2018.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[15] B. Yang, S. Rosa, A. Markham, N. Trigoni, and H. Wen, "Dense 3d object reconstruction from a single depth view," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[16] W. Wang, Q. Huang, S. You, C. Yang, and U. Neumann, "Shape inpainting using 3d generative adversarial network and recurrent convolutional networks," *arXiv preprint arXiv:1711.06375*, 2017.

[17] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5485–5493.

[18] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[19] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.

[20] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-net: Image inpainting via deep feature rearrangement," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 1–17.

[21] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *arXiv preprint arXiv:1710.10916*, 2017.

[22] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[23] C. Wang, C. Xu, C. Wang, and D. Tao, "Perceptual adversarial networks for image-to-image transformation," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4066–4079, 2018.

[24] D. M. Chandler, "Seven challenges in image quality assessment: past, present, and future research," *ISRN Signal Processing*, vol. 2013, 2013.

[25] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.

[26] S. Ling and P. Le Callet, "Image quality assessment for dibr synthesized views using elastic metric," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 1157–1163.

[27] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, vol. 2. IEEE, 2007, pp. II–169.

[28] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1733–1740.

[29] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3773–3777.

[30] D. Li, T. Jiang, and M. Jiang, "Exploiting high-level semantics for no-reference image quality assessment of realistic blur images," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 378–386.

[31] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image and Video Processing*, vol. 12, no. 2, pp. 355–362, 2018.

[32] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2018.

[33] K.-Y. Lin and G. Wang, "Hallucinated-iqa: No-reference image quality assessment via adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 732–741.

[34] J. Xu, W. Zhou, Z. Chen, S. Ling, and P. L. Callet, "Predictive auto-encoding network for blind stereoscopic image quality assessment," *arXiv preprint arXiv:1909.01738*, 2019.

[35] P.-H. Conze, P. Robert, and L. Morin, "Objective view synthesis quality assessment," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2012, pp. 82 881M–82 881M.

[36] F. Battisti, E. Bosc, M. Carli, P. Le Callet, and S. Perugia, "Objective image quality assessment of 3d synthesized views," *Signal Processing: Image Communication*, vol. 30, pp. 78–88, 2015.

[37] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "Dibr synthesized image quality assessment based on morphological wavelets," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*. IEEE, 2015, pp. 1–6.

[38] D. Sandic-Stankovic, D. Kukolj, and P. Le Callet, "Dibr synthesized image quality assessment based on morphological pyramids," in *2015 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE, 2015, pp. 1–4.

[39] P. L. C. Ling, Suiyi and C. Gene, "Quality assessment for synthesized view based on variable-length context tree," in *Multimedia Signal Processing (MMSP), 2017 IEEE 19th International Workshop on*. IEEE, 2017.

[40] S. Ling and P. Le Callet, "Image quality assessment for free viewpoint video based on mid-level contours feature," in *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 2017, pp. 79–84.

[41] S. Ling, J. Gutiérrez, K. Gu, and P. Le Callet, "Prediction of the influence of navigation scan-path on perceived quality of free-viewpoint videos," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 204–216, 2019.

[42] S. Ling, J. Li, Z. Che, X. Min, G. Zhai, and P. L. Callet, "Quality assessment of free-viewpoint videos by quantifying the elastic changes of multi-scale motion trajectories," *arXiv preprint arXiv:1903.12107*, 2019.

[43] 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 Transactions on Multimedia*, vol. 20, no. 4, pp. 914–926, 2018.

[44] S. Tian, L. Zhang, L. Morin, and O. Deforges, "Niqsv: A no reference image quality assessment metric for 3d synthesized views," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1248–1252.

[45] S. Tian, L. Zhang, L. Morin, and O. Déforges, "Niqsv+: A no-reference synthesized view quality assessment metric," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1652–1664, 2018.

[46] 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 Transactions on Multimedia*, 2018.

[47] S. Ling and P. Le Callet, "How to learn the effect of non-uniform distortion on perceived visual quality? case study using convolutional sparse coding for quality assessment of synthesized views," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 286–290.

[48] S. Ling, G. Cheung, and P. Le Callet, "No-reference quality assessment for stitched panoramic images using convolutional sparse coding and compound feature selection," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[49] D. D. Sandić-Stanković, D. D. Kukolj, and P. Le Callet, "Fast blind quality assessment of dibr-synthesized video based on high-high wavelet subband," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5524–5536, 2019.

[50] 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 Transactions on Image Processing*, vol. 29, pp. 1802–1814, 2019.

[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.

[52] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

[53] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[54] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[55] Y. J. Jung, H. Sohn, S.-i. Lee, Y. M. Ro, and H. W. Park, "Quantitative measurement of binocular color fusion limit for non-spectral colors," *Optics express*, vol. 19, no. 8, pp. 7325–7338, 2011.

[56] J. Li, "Methods for assessment and prediction of qoe, preference and visual discomfort in multimedia application with focus on s-3dtv," Ph.D. dissertation, Université de Nantes, 2013.

[57] P. Hedman, J. Philip, T. Price, J.-M. Frahm, G. Drettakis, and G. Brostow, "Deep blending for free-viewpoint image-based rendering," in *SIGGRAPH Asia 2018 Technical Papers*. ACM, 2018, p. 257.

[58] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels," Tech. Rep., 2010.

[59] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 11, pp. 2227–2240, 2014.

[60] C. Boutsidis, A. Zouzias, M. W. Mahoney, and P. Drineas, "Randomized dimensionality reduction for k-means clustering," *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 1045–1062, 2015.

[61] P. Gastaldo, R. Zunino, and J. Redi, "Supporting visual quality assessment with machine learning," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 54, 2013.

[62] D. Kinga and J. B. Adam, "A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[63] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[64] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of graphics tools*, vol. 9, no. 1, pp. 23–34, 2004.

[65] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on image processing*, vol. 13, no. 9, pp. 1200–1212, 2004.

[66] G. Luo, Y. Zhu, Z. Li, and L. Zhang, "A hole filling approach based on background reconstruction for view synthesis in 3d video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1781–1789.

[67] M. Solh and G. AlRegib, "Hierarchical hole-filling for depth-based view synthesis in ftv and 3d video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 495–504, 2012.

[68] V. Jantet, C. Guillemot, and L. Morin, "Object-based layered depth images for improved virtual view synthesis in rate-constrained context," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 125–128.

[69] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Reference softwares for depth estimation and view synthesis," 2008.

[70] I. Ahn and C. Kim, "A novel depth-based virtual view synthesis method for free viewpoint video," *IEEE Transactions on Broadcasting*, vol. 59, no. 4, pp. 614–626, 2013.

[71] C. Zhu and S. Li, "Depth image based view synthesis: New insights and perspectives on hole generation and filling," *IEEE Transactions on Broadcasting*, vol. 62, no. 1, pp. 82–93, 2016.

[72] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "Dibr-synthesized image quality assessment based on morphological multi-scale approach," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, p. 4, 2016.

[73] ITU, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," *ITU-T Recommendation P.913*, 2014.

[74] L. Ruan, B. Chen, and M.-L. Lam, "Light field synthesis from a single image using improved wasserstein generative adversarial network," in *Proceedings of the 39th Annual European Association for Computer Graphics Conference: Posters*. Eurographics Association, 2018, pp. 19–20.

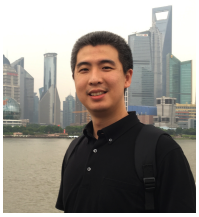
[75] V. Kiran Adhikarla, M. Vinkler, D. Sumin, R. K. Mantiuk, K. Myszkowski, H.-P. Seidel, and P. Didyk, "Towards a quality metric for dense light fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 58–67.

Suiyi Ling received the B.S. degree in computer sicence from GuangDong University of Technology in 2013. He received the M.S. degree in multimedia and big data management from University of Nantes and in computer science from GuangDong University of Technology (double degree) in 2015 and 2016 respectively. He received the Ph.D. degree in computer science from University of Nantes in 2018. He is currently working as a research scientist at CAPACITÉS SAS, France. His research interests include computer vision, machine learning, multimedia quality assessment, and perceptual image processing.

Jing Li received her M.S. degree in Pattern Recognition and Intelligence Systems from Xidian University, China in 2010, and her Ph.D. degree in computer science from the Image and Video Communications (IVC/IRCCyN) Group in Polytech Nantes, University of Nantes, France in 2013. From 2013 to 2019, she has been working in IPI/LS2N lab as a researcher. During 2014 to 2016, she was an assistant professor in University of Nantes, France. In 2019, she joined Alibaba Group as a Senior Staff Algorithm Engineer, leading the Moku Laboratory. She is a member of Video Quality Experts Group (VQEG) and IEEE Standard P3333.1. She is the contributor of International Telecommunication Union (ITU) Standards ITU-T Rec. P.914, P.915 and P.916, and IEEE Standard P3333.1.1. Her research interests include image quality assessment, QoE of immersive multimedia including both psychophysical study and objective modeling, deep learning, active learning and information retrieval.

Zhaohui Che received the B.E. degree from School of Electronic Engineering, Xidian University, Xi'an, China, in 2015. He is currently working toward the Ph.D. degree at the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include visual attention, perceptual quality assessment, deep learning, and adversarial attack and defense. From 2018 to 2019, he was a Visiting Student at the Ecole Polytechnique de l'Universite de Nantes, Nantes, France. He won the Grand Prize of the ICME 2018 Grand Challenge on "Salient360!" for visual attention modeling for panoramic content. He was a co-organizer of the Grand Challenge "Saliency4ASD" at IEEE ICME 2019.

Wei Zhou is currently a joint Ph.D. Student with the Department of Electronic Engineer and Information Science at University of Science and Technology of China and the Department of Electrical and Computer Engineering at University of Waterloo, ON, Canada.



Junle Wang is currently a senior researcher with Tencent Interactive Entertainment Group, Shenzhen, China. He received his BSc and M.S degree from South China University of Technology. In 2012, he received his PhD degree in computer science from the Image and Video Communications (IVC/IRCCyN) Lab in University of Nantes. He then became an assistant professor in the Department of Electronic and Digital Technologies of University of Nantes. In 2013, he worked as a France-Singapore Merlion research fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. In 2014, he created a startup on computer vision and deep learning in Nantes, France. In 2018, he joined the Quality Management department of Tencent Games. Currently he is a member of Video Quality Experts Group (VQEG). He has served as reviewers for several journals including IEEE TIP, TMM, TCSVT, TNNLS, J-STSP, SPIE JEI. His research interests include QoE and cloud gaming, image quality assessment, image processing for immersive media, human visual perception, image classification and object detection.

Patrick Le Callet (IEEE Fellow) is full professor at University of Nantes, in the Electrical Engineering and the Computer Science departments of Polytech Nantes. He is one of the steering director of CNRS LS2N lab (450 researchers). He is also the scientific director of the cluster “Ouest Industries Créatives”, gathering more than 10 institutions (including 3 universities). “Ouest Industries Créatives” aims to strengthen Research, Education and Innovation of the Region Pays de Loire in the field of Creative Industries. He is mostly engaged in research dealing with cognitive computing and the application of human vision modeling in image and video processing. His current centers of interest are AI boosted QoE Quality of Experience assessment, Visual Attention modeling and applications. He is co-author of more than 300 publications and communications and co-inventor of 16 international patents on these topics. He serves or has been served as associate editor or guest editor for several Journals such as IEEE TIP, IEEE STSP, IEEE TCSVT, Springer EURASIP Journal on Image and Video Processing, and SPIE JEI. He is serving in IEEE IVMSP-TC (2015- to present) and IEEE MMSP-TC (2015-to present) and one the founding member of EURASIP TAC (Technical Areas Committee) on Visual Information Processing.