Quality assessment of DIBR-synthesized views: An overview

S. Tian, Lu Zhang, W. Zou, Xiaojian Li, T. Su, Luce Morin, O. Déforges

Quality Assessment of DIBR-synthesized views: An Overview

Shishun Tian[a,b], Lu Zhang[c,d], Wenbin Zou[a,b,*], Xia Li[a,b], Ting, Su[e], Luce, Morin[c,d] and Olivier, Déforges[c,d]

[a]College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China.
[b]Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China.
[c]National Institute of Applied Sciences of Rennes (INSA Rennes), Rennes, France.
[d]IETR (Institut d'Electronique et des Technologies du numŔique), UMR CNRS 6164, Rennes, France.
[e]Research Center for Medical Articial Intelligence, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China.

ARTICLE INFO

Keywords:
DIBR
Multi-view videos (MVV)
view synthesis
distortions
quality assessment.

ABSTRACT

The Depth-Image-Based-Rendering (DIBR) is one of the main fundamental technique to generate new views in 3D video applications, such as Multi-View Videos (MVV), Free-Viewpoint Videos (FVV) and Virtual Reality (VR). However, the quality assessment of DIBR-synthesized views is quite different from the traditional 2D images/videos. In recent years, several efforts have been made towards this topic, but there ~~lacks a~~ is a lack of detailed survey in the literature. In this paper, we provide a comprehensive survey on various current approaches for DIBR-synthesized views. The current accessible datasets of DIBR-synthesized views are firstly reviewed~~. Followed~~ , followed by a summary analysis of the representative state-of-the-art objective metrics. Then, the performances of different objective metrics are evaluated and discussed on all available datasets. Finally, we discuss the potential challenges and suggest possible directions for future research.

1. Introduction

Providing more immersive experiences with depth perception to the observers, the 3D applications, such as the Multi-View Video (MVV) and Free-Viewpoint Video (FVV), have drawn great public attention in recent years. These 3D applications allow the users to view the same scene at various angles which may result in a huge information redundancy and cost tremendous bandwidth or storage space. To reduce these limitations, researchers attempt to transmit and store only a subset of these views and synthesize the others at the receiver by using the Multiview-Video-Plus-Depth (MVD) data format and Depth-Image-Based-Rendering (DIBR) techniques [1, 2]. Only limited viewpoints (both texture images and depth maps) are included in the MVD data format, the other view images are synthesized through DIBR.

This MVD plus DIBR scenario greatly reduces the burden on the storage and transmission of 3D video contents. However, the DIBR view synthesis technique also raises new challenges in the quality assessment of virtual synthesized views. During the DIBR process, the pixels in the texture image at the original viewpoint are back-projected to the real 3D space, and then re-projected to the target virtual viewpoint using the depth map, which is called 3D image warping in the literature. As shown in Fig. 1, DIBR view synthesis can be divided into two parts: 3D image warping and hole filling. During the 3D image warping procedure, the pixels in the original view are warped to the corresponding position in the target view. ~~Owing to the changing of~~ Because of the change in the viewpoint, some objects which are invisible in the original view may become visible in the target one, which is called dis-occlusion and causes black holes in the synthesized view. Then, the second step is to fill the black holes. The holes can be filled by typical image in-painting algorithms [3]. Most of the image in-painting algorithms use the pixels around the "black holes" to search the similar regions in the same image, and then use this similar region to fill the "black holes". Due to the imprecise depth map and imperfect image in-painting method, various distortions, which are quite different from the traditional ones in 2D images/videos, may be caused. Most of the 2D objective quality metrics [4, 5, 6, 7, 8] which focus on the traditional distortions will fail to evaluate the quality of DIBR-synthesized views. Subjective test is the most accurate and reliable way to assess the quality of media content since the human observers are the ultimate users in most applications. The subjective tests offer the datasets along with subjective quality scores. The objective metrics are designed to mathematically model and predict the subjective quality scores. In other words, an ideal objective model is expected to be consistent with the subjective results. Since the subjective test is time consuming and practically not suitable for real-time applications, effective objective metrics are highly desired.

Although several efforts have been made targeting at the objective quality assessment of DIBR-synthesized views in recent years, to the best of our knowledge, there is not a detailed survey on these works in the current literature. In this paper, we provide a comprehensive survey on the quality assessment approaches for DIBR-synthesized views ranging from the subjective to objective methods. The main contributions can be summarized as follows: (1) the state-of-the-art metrics are introduced and classified based on their ~~used~~ approaches; (2) the metrics ~~in terms of the contributions, advantages and disadvantages are analysed in depth~~

*Corresponding author
wzou@szu.edu.cn (W. Zou)
ORCID(s): 0000-0002-7616-8382 (S. Tian); 0000-0003-1389-9089 (W. Zou)

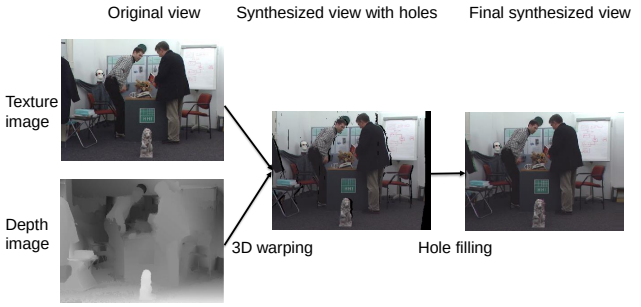


Figure 1: Procedure of DIBR.

(a) Reference view (b) Synthesized view

Figure 2: Object shifting caused by depth low-passing filter, the right borders of the character's faces are slightly modified. These images are from IVC DIBR image dataset [12].

are analyzed deeply in terms of the contributions, advantages and disadvantages; (3) the performances of these metrics are evaluated on different datasets, and ~~the essential reasons of~~ their performances on different type of distortions are ~~analysed~~ reasoned; (4) furthermore, the limitations of current works are discussed and the possible directions for future research are given.

The rest of this paper is organized as follows. Firstly, Section 2 introduces the DIBR view synthesis technique and analyses the view synthesis distortions. Secondly, the subjective methods are surveyed in Section 3. Section 4 introduces the state-of-the-art objective quality metrics in detail. The experimental results are presented and discussed in Section 5. Finally, the conclusions are given in Section 6.

2. Depth-Image-Based-Rendering (DIBR) and distortion analysis

As introduced in the previous section, the DIBR view synthesis procedure consists of two parts: 3D warping and hole filling *cf.* Fig. 1. Due to the lack of original texture information, various distortions may be induced in the DIBR-synthesized views which significantly degrade the image quality. In this section, we give a review of the algorithms that are designed to improve the visual quality of DIBR-synthesized views, and then analyze the distortions that may occur in the DIBR-synthesized views.

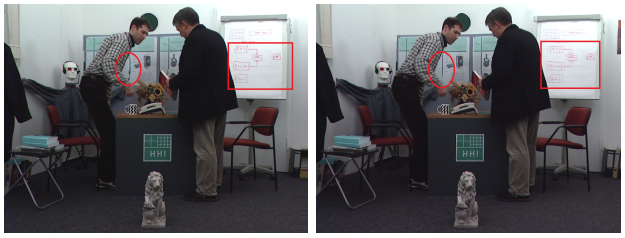
2.1. Review of state-of-the-art DIBR algorithms

During the 3D warping process, a large number of small cracks may be induced by the numerical rounding operations of pixel positions since the corresponding pixel position in the target viewpoint may be not an integer. These distortions mainly happen in the regions where the depth values are significantly different from their neighbours. Normally, these small cracks are handled by filtering the warped depth map with a low-pass filter [9, 10, 11]. However, this may also cause slightly object shift in the synthesized views *cf.* Fig. 2.

Dis-occlusion hole filling also plays an important role in generating a high quality synthesized view. Many image in-painting algorithms have been used to fill the dis-occlusion holes, such as the Criminisi's Examplar based algorithm [13] and the Telea's algorithm [14]. However, these in-painting algorithms do not consider the view synthesis characteristics. For example, the dis-occlusion regions are non-visible background objects in the original viewpoint but become visible in the target viewpoint. In other words, the dis-occlusion regions should be filled with background content. Face this issue, many studies [15, 16, 9] tried to extend the main idea of these image in-painting methods to DIBR view synthesis. Oliveira [15] extends the Criminisi's image in-painting method by changing the hole filling order with depth information. The texture propagation is enforced from the background to the foreground. Muddala [16] constrains the confidence and data terms to the background areas and local information. Ahn [9] improves the Criminisi's image in-painting method by optimizing the filling priority and the patch-matching measure. The optimized matched patch is selected only through the data term on the background areas which are extracted using warped depth map. It greatly reduces the ghost effect in the DIBR-synthesized views.

Instead of optimizing the priorities and searching regions of in-painting method, [10, 17] try to reconstruct the background content and then use the reconstructed background to eliminate the dis-occlusion holes in the virtual viewpoint. Jantet *et al.* proposed an object-based Layered Depth Image (LDI) representation to improve the quality of virtual synthesized views [10]. They firstly segment the foreground and background based on a region growing algorithm, which allows organising the LDI pixels into two object-based layers. Once the extracted foreground is obtained, an in-painting method is used to reconstruct the complete background image on both depth and texture images. Luo *et al.* proposed a hole filling approach for DIBR systems based on background reconstruction [17]. The foreground is firstly removed by using morphological operations and random walker segmentation. Then, the background is reconstructed based on motion compensation and a modified Gaussian Mixture model.

All the DIBR view synthesis algorithms introduced above are single view based synthesis method. They use only one neighbouring view to extrapolate the synthesized views. Differently, the interview algorithms use two neighbouring views to synthesize the virtual viewpoint images. The most popular interview synthesis method would be the View Synthesis



(a) single view based mode (b) interview mode

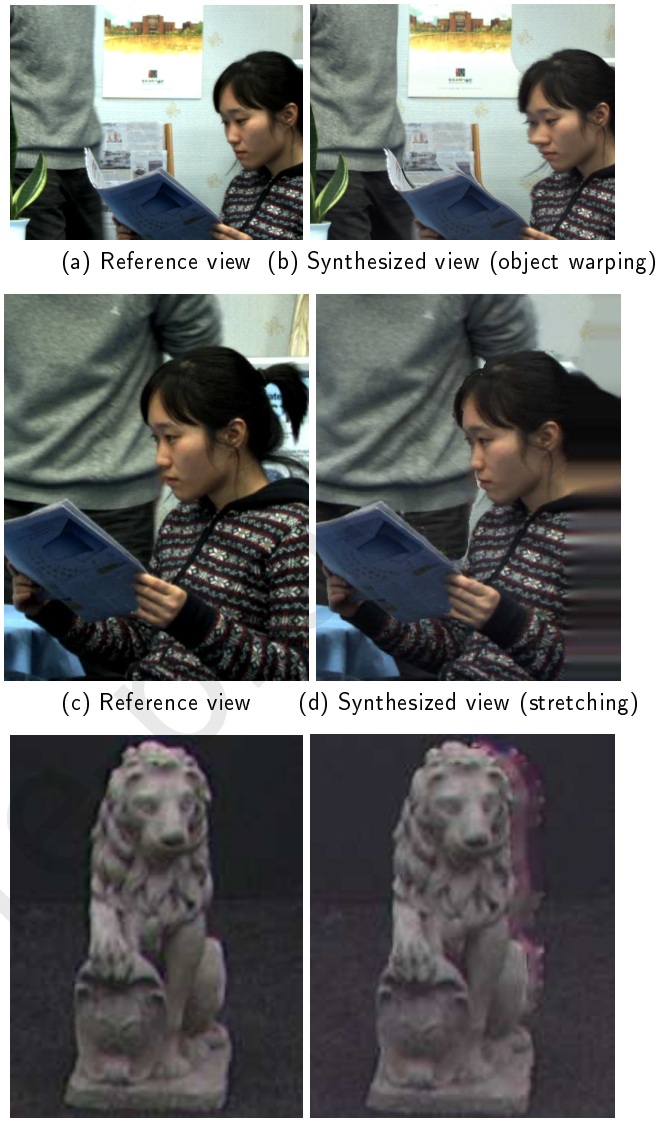
Figure 3: Examples of images synthesized by VSRS using single view based synthesis mode and interview synthesis mode. These images are from IETR DIBR image dataset [19].

Reference Software (VSRS) [11] which has been adopted by the MPEG 3D video Group. The depth discontinuity artefacts are firstly solved by performing a post-filter on the projected depth map. Then, the in-painting method proposed in [14] is used to fill the holes in the dis-occluded regions. Note that this approach is primarily used in the inter-view synthesis applications which only have small holes to be filled, but it can also be used in single view based rendering cases. Instead of in-painting the warped images directly, [18] focuses on the use of the occluded information to identify the relevant background pixels around the holes. Firstly, the occluded background information is registered in both texture and depth during 3D warping. Then, the un-occluded background information around the holes is extracted based on the depth map. After that, a virtual image is generated by integrating the occluded background and un-occluded background information. The dis-occluded holes are filled based on this generated image with the help of a depth-enhanced Criminisi's in-painting method and a simplified block-averaged filling method.

With more information, the interview synthesis cases only have smaller dis-occlusion regions to be filled, they thus outperform the single view based view synthesis methods in most circumstances. However, due to the inaccuracy of depth map, the same object in the two base views could be rendered to different positions which results in a "ghost" effect in the synthesized view. This phenomenon does not happen in the single view base synthesis method. As shown in Fig. 3, there exists a "ghost" effect of the "chat flow" on the board marked by red block in (b); but according to the synthesized content marked by red circle, the interview synthesis method (b) works better than the single view based one (a) in generating the object texture.

2.2. Distortion analysis

Imperfect hole filling methods may induce various distortions in the DIBR-synthesized views, such as object warping, stretching and blurry regions, *cf.* Fig. 4. Fig. 4 (a), (b) give an example of object warping distortion caused by Telea's image in-painting algorithm [14]. It could be observed that the "newspaper" and the "girl's nose" are extreme warped. The stretching distortion (the "girl's hair and clothes") mainly happen in the out-of-field areas *cf.* Fig. 4



(a) Reference view (b) Synthesized view (object warping)

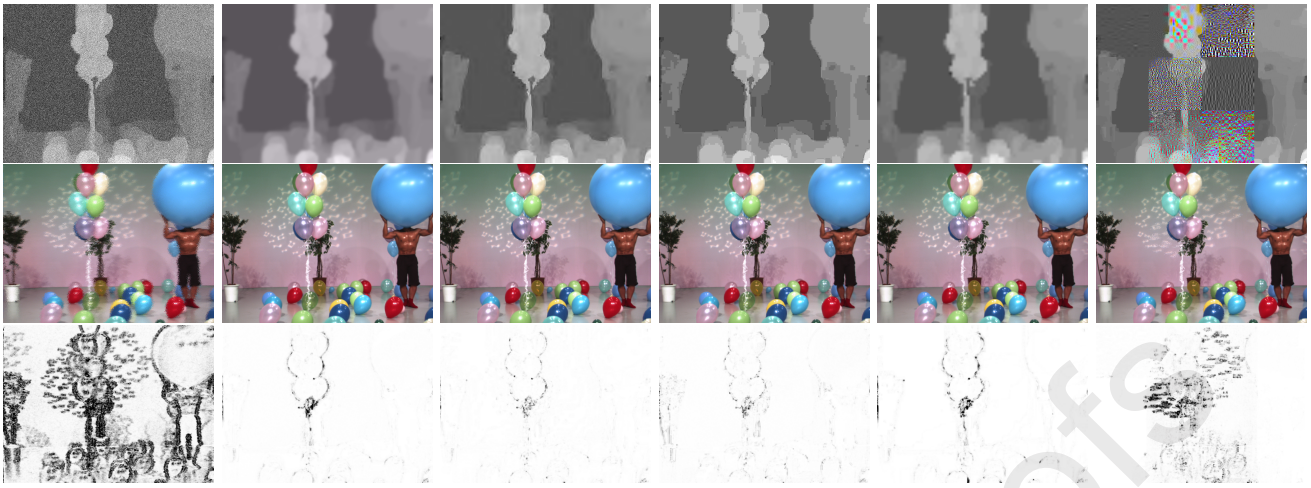
(c) Reference view (d) Synthesized view (stretching)

(e) Reference view (f) Synthesized view (blurry region)

Figure 4: Example of distortions caused by imperfect image in-painting method. These images are from IVC DIBR-image dataset [12].

(c), (d). The blurry regions can be noticed around the sculpture in Fig. 4 (e), (f).

Depth map represents the distance of objects to the camera. It is composed of a series of flat homogeneous regions and sharp edges. The flat areas indicate the objects at a certain distance while the edges relate to the transition of foreground and background objects. This is quite different from the natural scene images. In DIBR view synthesis, depth maps are used to guide the 3D warping. The distortions in the depth map will certainly induce degradations in the DIBR-synthesized views. In order to analyze the effect of depth distortions on the quality of DIBR-synthesized views, we compare the images that are synthesized with undistorted depth map and depth maps with various distortions. As shown in Fig. 5, we can easily observe that most of the distortions



(a) AWN (b) Gaussian blur (c) JP2K (d) JPEG (e) Sampling blur (f) Transmission loss

Figure 5: Example of synthesized images using depth map with different distortions. The first row shows the distorted depth maps while the second row gives the DIBR-synthesized images using the corresponding distorted depth maps. The third row presents the SSIM maps between the synthesized and reference images. Note that, the reference images are the images synthesized with undistorted depth maps. These images are from MCL-3D image dataset [20].

distribute around the edge regions of the depth map. It is logical that the edge of depth map represents the transition of foreground and background objects, the noise in these edge regions will certainly cause aliasing of foreground and background texture. Besides, we also notice that the synthesized view quality is more sensitive to high-frequency distortions (*e.g.* additive white noise (AWN), transmission loss) in the depth map compared to the low-frequency distortions (*e.g.* Gaussian blur). The main reason would be that high-frequency distortions in depth map will cause great local shift in the synthesized view, which is much more annoying to human vision system and can be easily penalized by pixel-based IQA metrics.

3. Subjective image/video quality assessment of DIBR-synthesized views

Subjective test is the most direct method for image/video quality assessment. During the test, a group of human observers are asked to rate the quality of each tested image or video. The subjective test results obtained from the subjective ratings are recognized as the quality of the tested images/videos. In different subjective test methodologies, the acquisition of subjective scores ~~are~~ is also different.

The Absolute Category Rating (ACR) method used in IVC image / video datasets [12, 21] randomly present the test sequences to the observers and ask them to rate on five-scale~~s~~ quality judgement (excellent, good, fair, poor, bad). The subjective quality scores are calculated by simply averaging the ratings. The Single Stimulus Continuous Quality Evaluation (SSCQE) in SIAT [22] dataset allows the observer to rate on a continuous scale instead of a discrete five-scale~~s~~ evaluation. The IVY [23] image dataset uses the Double Stimulus Continuous Quality Scale (DSCQS). The test image along with its associated reference image are presented in succession. It is usually used when the test and reference images are similar. Pairwise Comparison (PC) method directly performs a one-to-one comparison of every image pair in the dataset. It is the most accurate and reliable way to get the subjective quality scores, but it ~~costs~~ takes too much time since all the image pairs need to be tested. The Subjective Assessment Methodology for VIdeo Quality (SAMVIQ) method used in IETR dataset can achieve much higher accuracy than ACR method for the same number of observers ~~and cost~~ . It takes less time than PC since it allows the observer to freely view several image multiple times and adopts a continuous rating scale. Besides, the IVY [23], IETR [19] and SIAT datasets normalize the obtained scores to *z-score* to make the results more intuitive. The IVC and MCL-3D [20] datasets directly use the average scores. Except for the subjective test methodology, as shown in Table 1, they use different sequences, DIBR algorithms, etc. In the following part, we will introduce them respectively in detail.

3.1. IVC DIBR datasets

The IVC DIBR-image dataset [12] was proposed by Bosc *et al.* in 2011. It contains 84 DIBR-synthesized view images synthesized by 7 DIBR algorithms [1, 14, 24, 25, 26, 27, 28]. 3 Multi-view plus Depth (MVD) sequences, *BookArrival*, *Lovebird1* and *Newspaper*, are extracted as the source contents. For each sequence, 4 virtual views are synthesized from the adjacent viewpoint by using the above algorithms. Note that in this dataset, virtual views were only generated by single-view-based synthesis, which means that the virtual view is synthesized with only one image and its associated depth map. The IVC DIBR-video dataset [21] uses almost the same contents and methodologies except that it adds the H.264 compression (with 3 quantization levels) distortion for each test sequence. In other words, there are 93 distorted

videos in this dataset, among which 84 ones only contain the DIBR view synthesis distortions. As one of the first DIBR related image datasets, the IVC datasets play an important role in the first research phase of this topic. However, because of the fast development of DIBR view synthesis algorithms, some of the distortions in these datasets do not exist any more in the state-of-the-art view synthesis algorithms.

3.2. IETR DIBR image dataset

Similar to the IVC datasets, the IETR dataset [19] is dedicated to investigate the DIBR view synthesis distortions as well. Compared to the IVC datasets, it uses more and newer DIBR view synthesis algorithms [13, 10, 9, 17, 29, 24, 18], ~~including~~ includes both interview synthesis and single view based synthesis, ~~to exclude~~ and excludes some "old fashioned" distortions, *e.g.* "black holes". In addition, the IETR dataset also uses more MVD sequences, of which 7 sequences (*Balloons*, *BookArrival*, *Kendo*, *Lovebird1*, *Newspaper*, *Poznan Street* and *PoznanHall*) are natural images and 3 sequences (*Undo Dancer*, *Shark* and *Gt_Fly*) are computer animation images. It contains 140 synthesized view images and their associated 10 reference images which are also captured by real cameras at the virtual viewpoints.

3.3. IVY stereoscopic image dataset

Jung *et al.* proposed the IVY stereoscopic 3D image dataset for the quality assessment of DIBR-synthesized stereoscopic images [23]. Different from the above two datasets, in addition to the DIBR view synthesis distortion, the IVY dataset explores binocular perception [30, 31] by showing the synthesized image pairs on a stereoscopic display. A total of 7 sequences and three MVD sequences are selected. 84 stereo images are synthesized by four DIBR algorithms [13], [9], [11], [32] in this dataset. All the virtual view images in the IVY dataset are generated by single-view-based synthesis methods.

3.4. MCL-3D image dataset

Song *et al.* proposed the MCL-3D stereoscopic image dataset [20] to evaluate the quality of DIBR-synthesized stereoscopic images. Although 4 DIBR algorithms are included, the number of images synthesized by these algorithms is quite limited (36 pairs). The major part of this dataset focuses on the traditional distortions in the synthesized views. 6 types of traditional distortions are considered in this dataset: additive white noise, Gaussian blur, down sampling blur, JPEG, JPEG2000 and transmission loss. Nine MVD sequences are collected, among which *Kendo*, *Lovebird1*, *Balloons*, *PoznanStreet* and *PoznanHall2* are natural images; *Shark*, *Microworld*, *GT_Fly* and *Undodancer* are Computer Graphics images. For each sequence, these traditional distortions are first applied on the base views. Then, the left and right view images are synthesized from these distorted base view images by using the view synthesis reference software (VSRS) [24]. Different from the above IVC, IETR and IVY datasets, the reference images in the MCL-3D dataset are the images synthesized from undistorted base view images instead of the ones captured by real cameras.

3.5. SIAT synthesized video dataset

The SIAT synthesized video dataset [22] focuses on the distortions caused by compressed texture and depth images in the synthesized views. It uses the same 10 MVD sequences as the IETR image dataset. For each sequence, 4 different texture and depth image quantization levels and their combinations are applied on the base views. Then, the videos at the virtual viewpoints are synthesized using the VSRS-1D-Fast software [33]. This dataset uses the real images (captured by real cameras at the virtual viewpoint) as references. Only interview synthesis is used in this dataset.

In the above datasets, the distortions in the DIBR-synthesized views come from not only the DIBR view synthesis algorithms, but also from the distorted texture and depth images. The IVC [34, 12, 21], IVY [23] and IETR [19] datasets focus on the distortions caused by different DIBR view synthesis algorithms; while the MCL-3D [20] and SIAT [22] datasets explore the influence of traditional 2D distortions of original texture and depth map on the DIBR-synthesized views. These datasets were usually used to evaluate and validate several quality metrics. In the next section, we will introduce the objective approaches for the quality assessment of DIBR-synthesized views.

4. Objective image/video quality assessment of DIBR-synthesized views

Several methods have been proposed to evaluate the quality of DIBR-synthesized views in the past decade. Based on the amount of reference information, these methods can be divided into 4 categories: Full-reference (FR), Reduced-reference (RR), Side View based Full-reference (SV-FR) and No-reference (NR), as shown in Fig. 6. The FR methods use the original undistorted image/video at the virtual viewpoint as reference to assess the quality of synthesized views, while the RR methods only use some features extracted from the original reference. Especially, the SV-FR methods use the undistorted image/video at the original viewpoint, from which the virtual view is synthesized, as the reference. The NR methods need no access to the original image/video.

Table 2 ~~classify~~ classifies the metrics based on their ~~used~~ approaches. Most of them (VSQA, MP-PSNR, MW-PSNR, EM-IQA and CT-IQA) evaluate the quality of synthesized views by considering the contour or gradient degradation between the synthesized and the reference images which is one of the most annoying characteristics of geometric distortions. Meanwhile some metrics (DSQM, 3DSwIM) calculate the quality score by comparing the extracted perceptual features between the synthesized and the reference images. Especially, the APT metric uses a local image description model to reconstruct the synthesize image, and evaluates the quality of the synthesized view based on the reconstruction error. These metrics are introduced as follows.

4.1. FR and RR metrics

In this subsection, we review 20 well-known FR metrics and 4 RR metrics.

Table 1
Summary of existing DIBR related datasets.

Name	Sequence	Resolution	Method.	DIBR algos		Other distortions	No. PVS [1]	Ref.	Disp.
				Name	Year				
IVC DIBR-image	BookArrival	1024 × 768	ACR[2] PC[3]	Fehn's	2004	None	84	Ori.	2D
	Lovebird1	1024 × 768		Telea's	2003				
	Newspaper	1024 × 768		VSRS	2009				
				Müller	2008				
				Ndjiki-Nya	2010				
				Köppel	2010				
				Black hole	—				
IVC DIBR-video	idem		ACR[2]	idem		H.264	93	Ori.	2D
IETR-image	BookArrival	1024 × 768	SAMVIQ[4]	Criminisi	2004	None	140	Ori.	2D
	Lovebird1	1024 × 768		VSRS	2009				
	Newspaper	1024 × 768		LDI	2011				
	Balloons	1024 × 768		HHF	2012				
	Kendo	1024 × 768		Ahn's	2013				
	Dancer	1920 × 1088		Luo's	2016				
	Shark	1920 × 1088		Zhu's	2016				
	Poznan_Street	1920 × 1088							
	PoznanHall2	1920 × 1088							
	GT_fly	1920 × 1088							
IVY image	Aloe	1280 × 1100	DSCQS[5]	Criminisi	2004	None	84	Ori.	Stereo.
	Dolls	1300 × 1100		Ahn's	2013				
	Reindeer	1300 × 1100		VSRS	2009				
	Laundry	1300 × 1100		Yoon	2014				
	Lovebird1	1024 × 768							
	Newspaper	1024 × 768							
	BookArrival	1024 × 768							
MCL-3D image	Kendo	1024 × 768	PC[3]	Fehn's	2004	Additive White Noise	684	Syn.	Stereo.
	Lovebird1	1024 × 768		Telea's	2003	Blur			
	Balloons	1024 × 768		HHF	2012	Down Sampling			
	Dancer	1920 × 1088		Black hole	—	JPEG			
	Shark	1920 × 1088				JPEG2k			
	Poznan_Street	1920 × 1088				Trans. Loss [9]			
	PoznanHall2	1920 × 1088							
	GT_fly	1920 × 1088							
	Microworld	1920 × 1088							
SIAT video	BookArrival	1024 × 768	SSCQE[6]	VSRS	2009	3DV-ATM	140	Ori.	2D
	Balloons	1024 × 768							
	Kendo	1024 × 768							
	Lovebird1	1024 × 768							
	Newspaper	1024 × 768							
	Dancer	1920 × 1088							
	PoznanHall2	1920 × 1088							
	Poznan_Street	1920 × 1088							
	GT_fly	1920 × 1088							
	Shark	1920 × 1088							

[1] PVS: Processed Video Sequences.
[2] ACR: Absolute Categorical Rating.
[3] PC: Pairwise Comparison.
[4] SAMVIQ: Subjective Assessment Methodology for VIdeo Quality.
[5] DSCQS: Double Stimulus Continuous Quality Scale.
[6] SSCQ: Single Stimulus Continuous Quality Scale.

4.1.1. Edge/Contour based FR metrics

The distortions in DIBR-synthesized views are mostly geometrical and structural distortions, which may degrade the object shape in the synthesized image. It can be measured by the change of object edges. In addition, the sharp edges in the depth map may also induce large dis-occlusions in the synthesized views which may result in dramatic distortions. Thus, a few edge-based methods have been proposed to evaluate the quality of DIBR-synthesized views.

The FR metric proposed by Bosc *et al.* in [35] indicates the structural degradations by calculating the contour displacement between the synthesized and the reference images. Firstly, a Canny edge detector is used to extract the image contours; then, the contour displacements between the synthesized and reference images are estimated. Based on the contour displacement map, three parameters are computed: the mean ratio of inconsistent displacement vectors per contour pixel, the ratio of inconsistent vectors, the ratio of new contours. The final quality score is obtained as a weighted sum of these three parameters.

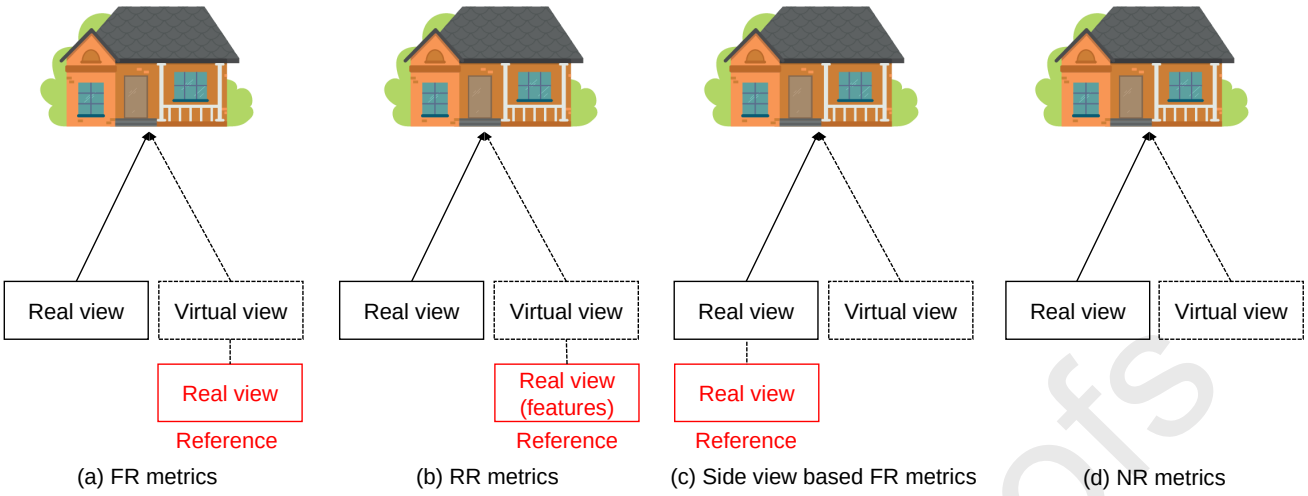


Figure 6: Categories of quality assessment metrics for DIBR-synthesized views.

In [42], Ling *et al.* proposed a contour-based FR metric ST-SIAQ for the quality assessment of DIBR-synthesized views. Instead of directly using the contour information in [35], ST-SIAQ uses mid-level contour descriptor called "Sketch Token" [80]. The "Sketch Token" stands as a codebook of image contour representation, of which each dimension can be recognized as the possibility which indicates how likely the current patch belongs to one certain category of contour from the codebook. To reduce the shifting effect in the feature comparison stage, the patches in the reference image are firstly matched to the synthesized image. The "Sketch Token" is clustered into 151 categories, which means the "Sketch Token" descriptor has 151 dimensions. A Random Forests decision model associated with a set of low-level features (including oriented gradient channels [81], color channels, and self-similarity channels [82]) ~~are~~ is used to obtain the "Sketch Token" descriptor. The geometric distortion strength in the synthesized view is calculated as the Kullback Leibler divergence of "Sketch Token" descriptors between the synthesized and reference images. In [83], this metric is improved to evaluate the quality of DIBR-synthesized videos by considering the temporal dissimilarity.

Ling *et al.* also proposed another contour-based FR metric EM-IQA in [43]. Different from ST-SIAQ metric, EM-IQA uses an interest points matching and an elastic metric [84], instead of block matching and "Sketch Token" descriptor, to compensate the shifting and evaluate the contour degradation respectively. After the interest points matching, a Simple Linear Iterative Clustering (SLIC) is used to extract the contours in the image. SLIC is originally proposed for image segmentation~~, in~~. In the EM-IQA metric, the boundaries of the segmented objects are considered as contours. Then, the elastic metric proposed in [84, 85] is used to finally measure the degradation between the contours of synthesized and reference images, which provides the quality score of DIBR-synthesized view.

In [41], Ling *et al.* proposed a variable-length context tree based image quality assessment metric CT-IQA, dedicated to quantify the overall structure dissimilarity and dissimilarities in various contour characteristics. Firstly, the contours of the reference and synthesized images are converted to differential chain code (DCC) [86] which represents the direction of object contours. Then, an optimal context tree [87] is learned from the DCC in the reference image. The overall structural dissimilarity is calculated by subtracting the encoding cost of DCC in the synthesized image and reference images. In addition, the overall dissimilarity in contour characteristics is also obtained by measuring the difference of total contour number, total contour start information and total number of symbols between the reference and synthesized image. The final quality score is calculated by combining the overall structure dissimilarity and contour characteristics dissimilarity.

Liu *et al.* proposed a gradient-based FR video quality assessment metric VQA-SIAT [22] by considering the "Activity" and "Flickering" which is the most annoying temporal distortion in the DIBR-synthesized views. The main contribution of this metric is the two following proposed structures: Quality Assessment Group of Pictures (QA-GoP) and Spatio-Temporal (S-T) tube. The QA-GoP acts as a process unit on a whole video sequence, it contains a group of 2N+1 frames (N frames before and N frames after the central frame). Besides, a block matching method is used to search the corresponding blocks of the central frame blocks in the forward and backward frames. The 2N + 1 blocks along the motion trajectory construct a S-T tube. The distortion of "Activity" is calculated from the difference of the spatial gradient in the (S-T) tube and (QA-GoP) between the synthesized and reference videos. The "Flickering" distortion is measured from the difference of temporal gradient, which is defined below:

$$\vec{\nabla} I^{temporal}_{x,y,i} = I(x, y, i) - I(x', y', i - 1), \quad (1)$$

where (x', y') is the coordinate in frame $i - 1$ corresponding to (x, y) along the motion trajectory in the previous frame i. The final quality score of DIBR-synthesized view video

Table 2
Overview of the existing metrics. The features in the first column indicate hand-craft feature (HF), deep feature (DF), contour/gradient (C/G), JND, Multi-scale decomposition (MSD), local image description (LID), depth estimation (DE), dis-occlusion Region (DR), Sharpness Evaluation (SE), Shift compensation (SC), Image complexity (IC), ML (Machine Learning).

Metric \ Approach		HF	DF	C/G	JND	MSD	LID	DE	DR	SE	SC	IC	ML
FR	Bosc *et al.* 2012 [35]	-	-	✓	-	-	-	-	-	-	-	-	-
	VSQA [36]	-	-	✓	-	-	-	-	-	-	-	-	-
	3DSwIM [37]	✓	-	-	-	-	-	-	-	-	✓	-	-
	MW-PSNR [38, 39]	✓	-	-	-	✓	-	-	-	-	-	-	-
	MP-PSNR [40]	✓	-	-	-	✓	-	-	-	-	-	-	-
	CT-IQA [41]	✓	-	-	-	-	-	-	-	-	-	-	-
	ST-SIAQ [42]	✓	-	✓	-	-	-	-	-	-	✓	-	-
	EM-IQA [43]	✓	-	✓	-	-	-	-	-	-	✓	-	-
	PSPTNR [44]	-	-	-	✓	-	-	-	-	-	-	-	-
	VQA-SIAT [22]	-	-	✓	-	-	-	-	-	-	✓	-	-
	SR-3DVQA [45]	-	-	✓	-	-	-	-	-	-	✓	-	✓
	SDRD [46]	-	-	-	-	-	-	-	✓	-	✓	-	-
	SCDM [47]	-	-	-	-	-	-	-	✓	-	✓	-	-
	SC-IQA [48]	-	-	-	-	-	-	-	-	-	✓	-	-
	CBA [23]	-	-	-	-	-	-	-	✓	-	-	-	-
	Zhou [49]	✓	-	✓	-	✓	-	-	-	-	-	-	✓
	Ling [50]	✓	-	✓	-	-	-	-	-	-	-	-	✓
	Wang [51]	✓	-	✓	-	-	-	-	✓	✓	-	-	-
RR	MP-PSNRr [52]	✓	-	-	-	✓	-	-	-	-	-	-	-
	MW-PSNRr [52]	✓	-	-	-	✓	-	-	-	-	-	-	-
	RRLP [53]	-	-	✓	-	-	✓	-	-	✓	-	-	-
Depth IQA FR/RR/NR	(FR) Li [54]	-	-	✓	-		-	-	-	-	-	-	-
	(RR) RR-DQM [55]	-	-	✓	-	-	✓	-	-	-	-	-	-
	(NR) BDQM [56]	-	-	✓	-	-	-	-	-	-	-	-	-
	(FR) Xiang [57]	-	-	✓	-	-	-	-	-	-	-	-	-
	(NR) SEP [58]	✓	-	✓	-	-	-	-	-	-	-	-	✓
SV-FR	3VQM [59]	-	-	-	-	-	-	✓	-	-	-	-	-
	LOGS [60]	-	-	-	-	-	-	-	✓	✓	✓	-	-
	DSQM [61]	✓	-	-	-	-	-	-	-	-	✓	-	-
	SIQE [62]	✓	-	-	-	-	-	-	-	-	✓	-	-
	SIQM [63]	✓	-	-	-	-	-	-	-	-	✓	-	-
NR	APT [64]	-	-	-	-	-	✓	-	-	-	-	-	-
	OUT [65]	-	-	-	-	-	✓	-	-	-	-	-	-
	MNSS [66]	✓	-	-	-	✓	-	-	-	-	-	-	✓
	NR_MWT [67]	✓	-	✓	-	✓	-	-	-	✓	-	-	-
	NIQSV [68]	-	-	✓	-	-	✓	-	-	-	-	-	-
	NIQSV+ [69]	-	-	✓	-	-	✓	-	-	-	-	-	-
	HEVSQP[70]	-	-	✓	-	-	-	-	-	-	-	-	✓
	CLGM [71]	-	-	✓	-	-	-	-	✓	✓	-	-	-
	GDIC [72]	✓	-	✓	-	-	-	-	-	-	-	✓	-
	Wang [73]	✓	-	✓	-	-	-	-	-	✓	-	✓	-
	SET [74]	✓	-	✓	-	✓	-	-	-	-	-	-	✓
	CTI [75]	-	-	-	-	-	-	-	-	-	✓	-	-
	FDI [76]	-	-	✓	-	-	-	-	-	-	✓	-	-
	CSC-NRM [77]	-	-	-	-	-	-	-	-	-	-	-	✓
	SIQA-CFP [78]	-	✓	-	-	-	-	-	-	-	-	-	✓
	GANs-NRM [79]	-	✓	-	-	-	-	-	-	-	-	-	✓

is obtained by integrating both "Activity" and "Flickering" distortions.

Furthermore, in [45], Zhang *et al.* proposed a FR metric SR-3DVQA combining the "Activity" measurement module in VQA-SIAT with a sparse representation-based flicker estimation method. In the SR-3DVQA metric, a DIBR-synthesized video is treated as a 3D volume data by stacking the frames sequentially. Then, the volume data is decomposed as a num-

ber of spatially neighboring temporal layers i.e. X-T or Y-T planes, where X, Y are the spatial coordinate and T is the temporal coordinate. In order to effectively evaluate the flicker distortion in the synthesized video, the gradient in the temporal layers and sharp edges in the associate depth map are extracted as key features for the dictionary learning and sparse representation. The rank-based method in [60] is used to pool the flicker score from the temporal layers. The final quality score is calculated by combining the flicker score and "Activity" score in the previous VQA-SIAT [22].

Jakhetiya *et al.* proposed a free-energy-principle-based IQA metric RRLP for Screen Content and DIBR-synthesized view images based on prediction model and distortion categorization [53]. The image quality is measured by calculating the disorder and sharpness similarity between the distorted and reference images. The disorder is obtained from a prediction model. As shown in Eq. 2, an observation-model-based bilateral filter (OBF) [88] is firstly used to divide the predicted and disorder parts.

$$\hat{X_{d_i}} = \frac{X_{d_i}\lambda + \sum_{k\in N_i}\omega_{k_i} I_{k_i}}{\lambda + \sum_{k\in N_i}\omega_{k_i}} \tag{2}$$

where $\hat{X_{d_i}}$ represents the predicted part, I_{k_i} and ω_{k_i} are respectively the pixels and their associated weights in the surrounding 3×3 window N_i of the ith pixel, λ is a parameter. The disorder part is computed as the difference between the predicted part and the original image:

$$R_{d_i} = |\hat{X_{d_i}} - X_{d_i}| \tag{3}$$

Then, the sharpness (edge structures) is calculated by four filters in[89]. Finally, the disorder and sharpness similarity between the distorted and reference images are estimated by using the similarity function in SSIM [4].

4.1.2. Wavelet transform based FR metrics

In the previous part, we introduced the metrics that use the edge/contour in luminance domain to evaluate the geometric distortions in DIBR-synthesized views. According to previous research, the wavelet transform representation can not only capture the image edges, but also some other texture unnaturalness. In this part, the wavelet transform based FR metrics will be reviewed.

Battisti *et al.* proposed an FR metric (3DSwIM) for DIBR-synthesized views based on the comparison of statistical features of wavelet sub-bands [37, 90]. The same as EM-IQA [43] and VQA-SIAT [22], 3DSwIM uses a block matching to ensure the "shifting-resilience". The distortions in each block of the synthesized view is measured by the Kolmogorov-Smirnov [91] distance between ~~of the two matched blocks.~~ the histograms of the matched blocks in the synthesized and reference images. In addition, since the Human Vision System (HVS) pays more attention on the human body, a skin detector is used to weight the skin regions in the matched blocks.

Sandić-Stanković *et al.* proposed another multi-scaled decomposition based FR metric MW-PSNR [39, 38]. The MW-PSNR uses morphological wavelet filters for decomposition. Then a multi-scale wavelet mean square error (MW-MSE) is calculated as the average MSE of all sub-bands and finally the MW-PSNR is calculated from it.

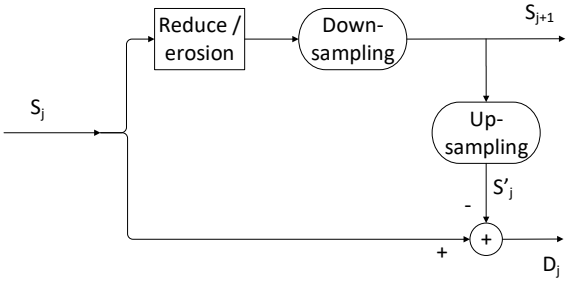


Figure 7: Decomposition scheme of MP-PSNR. S_j represents the image at scale j ($j \in [1,5]$), D_j represent the detail image at scale j [40].

The wavelet transform based FR metrics can be recognized as a kind of edge/contour based metrics. For example, the higher sub-bands of the wavelet transformed image represent the edge information of the original image. Compared to the pixel level edge/contour used in the previous subsection, the metrics in this subsection use the features in wavelet transformed domain to represent both the image edges and other characteristics.

4.1.3. Morphological operation based FR metrics

Morphological operations are widely used in image processing, especially a couple of erosion and dilation operations can be used to detect the image edges [92]. In [40], Sandić-Stanković *et al.* proposed the MP-PSNR based on multi-scaled pyramid decomposition using morphological filters. The basic erosion and dilation operations used in MP-PSNR are calculated as maximum and minimum in the neighbourhood defined by the structure element, as shown in the following equation:

$$D : dilation_{SE}(f)(x) = max_{y\in SE} f(x-y) \tag{4}$$

$$E : erosion_{SE}(f)(x) = min_{y\in SE} f(x+y) \tag{5}$$

where f is a gray-scale image and SE is binary structure element. Then, they use the Mean Square Error (MSE) between the reference and synthesized images in all pyramids' sub-bands to quantify the distortion. As shown in Fig 7, during the decomposition, the dilation is used as expanding operation and the erosion is used as ~~reducer~~reducing operation. The detail image of each scale is calculated as the difference between the original and processed (erosion and dilation) images. Finally, the overall quality is calculated by averaging the MSE of detail images in all the sub-bands and expressing it ~~as~~ in terms of PSNR.

In [52], Sandić-Stanković *et al.* also proposed the reduced version of MP-PSNR, and MW-PSNR. Only detail

images from higher decomposition scales are taken into account to measure the difference between the synthesized image and the reference image. The reduced version achieved significant improvement over the original FR metrics with lower computational complexity.

4.1.4. Dis-occlusion region based FR metrics

Since the DIBR view synthesis distortions mainly occur in the dis-occlusion regions, some of the FR metrics improve the performance of 2D FR metrics by using dis-occlusion maps[46, 47] instead of using weighting maps.

The SDRD metric proposed by Zhou in [46] detects the dis-occlusion regions by simply comparing the absolute difference between the synthesized and reference images. Before that, a self-adaptive scale transform model is used to eliminate the effect of view distance, and a SIFT flow-based warping is adopted to compensate the global shift in the synthesized view image. The final quality score is obtained by weighting the dis-occlusion regions with their size since the distortions with bigger size are more annoying to human vision system.

Tian *et al.* proposed a full-reference quality assessment model (SCDM) for 3D synthesized views by considering global shift compensation and dis-occlusion regions [47]. This model can be used on any pixel-based FR metrics. SCDM firstly compensates the shift by using a SURF + RANSAC approach instead of the SIFT flow used in SDRD. Then, the dis-occlusion regions are directly extracted from the depth map. It is more precise and uses more resources compared to SDRD. The final quality score is obtained as a weighted PSNR or weighted SSIM. It is reported to improve the performance of PSNR and SSIM by 36.85% and 13.33% in terms of Pearson Linear Correlation Coefficients (PLCC).

Since the distortions in the DIBR-synthesized views are not restricted in the dis-occlusion regions only, they may occur around these regions as well. In [51], Wang *et al.* proposed a critical region based metric by dilating the dis-occlusion region with a morphological operator. Similar to SDRD, the dis-occlusion region map is extracted by a SIFT-flow based approach. Then a Discrete Cosine Transform (DCT) decomposition method is used to partition and classify the critical regions into edge blocks, texture blocks and smooth blocks. Based on the perceptual properties of these three types of blocks, their distortions are measured differently. The edge and texture blocks contain more complex edges or texture information, the blur distortions in these regions would be much more annoying than that in the smooth regions. On the other hand, the smooth regions are sensitive to color degradations. Thus, the texture similarity and color contrast similarity between the synthesized and reference images are calculated to measure the local distortions in the edge, texture and smooth blocks respectively. Finally, a global sharpness detection is combined with the local distortion measurement to obtain the overall quality score.

4.1.5. 2D related FR metrics

The main reason of the ineffectiveness of 2D quality assessment metrics on DIBR-synthesized views can be ~~analysed~~ analyzed as follows. Firstly, there ~~exists~~ are large object shifts in the synthesized views and this kind of shifts ~~effect~~ can be easily penalized by 2D metrics even though the HVS is not sensitive to the global shift in the image. The second reason is the distribution of distortions. The distortions in traditional 2D images often scatter over the whole image while the DIBR view synthesis distortions are mostly local, ~~especially~~ mainly in the dis-occluded regions. The 2D related metrics are based on the traditional 2D FR metrics, ~~such as~~ e.g. PSNR, SSIM, etc. They try to improve the performance of 2D metrics by considering HVS and the characteristics of DIBR view synthesis distortions.

The VSQA metric proposed by Conze *et al.* in [36] tries to improve the performance of SSIM [4] by taking advantage of known characteristics of the ~~human visual system (~~HVS~~)~~. It aims to handle areas where disparity estimation may fail, such as thin objects, object borders, transparency, etc., by applying three weighting maps on the SSIM distortion map. The main purpose of these three weighting maps is to characterize the image complexity in terms of texture~~s~~, diversity of gradient orientations and presence of high contrast since the HVS is more sensitive to the distortions in such areas. For example, the distortions in an untextured area are much more annoying than the ones located in a high texture complexity area. It is reported that this method approaches a gain of 17.8% over SSIM in correlation with subjective measurements.

Zhao *et al.* proposed the PSPTNR metric to measure the perceptual temporal noise of the synthesized sequence [44]. The temporal noise is defined as the the difference between inter-frame change in the processed sequence and that in the reference sequence:

$$TN_{i,n} = ((P_{i,n} - P_{i,n-1}) - (R_{i,n} - R_{i,n-1})^2, \tag{6}$$

where TN indicates the temporal noise, P and R represent the distorted and reference sequence respectively. In order to better predict the perceptual quality of synthesized videos, temporal noise is filtered by a Just Noticeable Distortion (JND) model and a motion mask [93], since the human can observe noise only beyond certain level and motion may decrease the texture sharpness in the video.

The shift compensation methods included in SDRD and SCDM only consider the global shift~~, b~~ But according to the recent research[94], ~~Human Visual System (HVS)~~ the HVS is more sensitive to local artefacts compared to the global object shift. In [48], Tian *et al.* proposed a shift-compensation based image quality assessment metric (SC-IQA) for DIBR-synthesized views. The same as SCDM, a SURF + RANSAC approach is used to roughly compensate the global shift. In addition, a multi-resolution block matching method is proposed to precisely compensate the global shift and penalize the local shift at the same time. A saliency map [95] is also considered to weight the distortion map of the synthesized view. Furthermore, only the blocks

with the worst quality are used to calculate the final quality score since HVS tends to perceive poor regions in an image with more severity than the good ones [94, 22]. SC-IQA achieves the performance of SCDM without access to the depth map.

The metrics introduced above consider only the view synthesis and compression artefacts which occur on applications that show the synthesized views on a 2D display, the binocular effect in the synthesized stereoscopic images is not taken into consideration. In [23], Jung *et al.* proposed a SSIM-based FR metric to measure the critical binocular asymmetry (CBA) in the synthesized stereo images. Firstly, the disparity inconsistency between the two different views is generated to detect the critical areas in terms of Left-Right image mismatches. Then, only the SSIM value on the critical areas of each view are computed to measure the asymmetry in the corresponding view image. The final binocular asymmetry score is obtained by averaging the asymmetry score in the left and right views.

4.2. Side view based FR metrics

The major limitation of the FR metrics is that they always need the reference view which may be unavailable in some circumstances (~~eg.~~ *e.g.* FVV). In other words, there is no ground truth for a full comparison with the distorted synthesized view. In this part, four side view based FR metrics will be reviewed. This kind of metrics use the real image/video at the original viewpoint, from which the virtual view is synthesized, as the reference to evaluate the quality of DIBR-synthesized virtual views. These metrics are named ~~as~~ "side view based FR metrics" in this paper.

Solh *et al.* proposed a side view based FR metric 3VQM [59] to evaluate synthesized view distortions by deriving an "ideal" depth map from the virtual synthesized view and the reference view at a different viewpoint. The "ideal" depth is the depth map that would generate the distortion-free image given the same reference image and DIBR parameters. Three distortion measurements, spatial outliers, temporal outliers and temporal inconsistency are calculated from the difference between the "ideal" depth map and the distorted depth map:

$$SO = STD(\triangle Z) \tag{7}$$

$$TO = STD(\triangle Z_{t+1} + \triangle Z_t) \tag{8}$$

$$TI = STD(Z_{t+1} + Z_t) \tag{9}$$

where SO, TO and TI denote the spatial outliers, temporal outliers and temporal inconsistencies respectively, STD represents the standard deviation. $\triangle Z$ is the difference between the "ideal" and the distorted depth maps and t is the frame number. These three measurements are then integrated into a final quality score. Since the calculation of the "ideal" depth map is based on the assumption that the horizontal shift of the synthesized view and the original view is small, this metric would not work well when the baseline distance increases.

Li *et al.* proposed a side view based FR metric for DIBR-synthesized views by measuring local geometric distortions in dis-occluded regions and global sharpness (LOGS) [60]. This metric consists of three parts. Firstly, the dis-occlusion regions are detected by using SIFT-flow based warping. These dis-occluded regions are extracted from the absolute difference map between the synthesized view I_{syn} and the warped reference view I^{w}_{ref} followed by an additional threshold. Then, the distortion size and strength in the local dis-occlusion regions are combined to obtain the overall local geometric distortion. The distortion size is simply measured by the number of pixels in the dis-occluded regions and the distortion strength is defined as the mean value of the dis-occluded regions in the whole difference map M. The next part is to measure the global sharpness by using a reblurring-based method. The synthesized image is firstly blurred by a Gaussian smoothing filter. Both the synthesized image and its reblurred version are divided into blocks. The sharpness of each block is calculated by its textural complexity, which is represented by its variance σ^2. Then, the overall sharpness score is computed by averaging the textural distance of all blocks. Finally, the local geometric distortion and the global sharpness are pooled to generate the final quality score.

Farid *et al.* proposed a side view based FR metric (DSQM) for the DIBR-synthesized view in [61]. A block matching is firstly used to estimate the shift between the reference and synthesized image. Then the difference of Phase congruency (PC) in these two matched blocks is used to measure the quality of the block in the synthesized image, which is defined as follows:

$$PC(x) = \max_{\bar{\phi(x)}\in[0,2\pi]} \frac{\sum_n A_n cos(\phi_n(x) - \bar{\phi(x)})}{\sum_n A_n} \tag{10}$$

where A_n and $\phi_n(x)$ represent the amplitude and the local phase of the n-th Fourier component at position x respectively. The implementation of phase congruency is based on ~~an~~ a logarithmic Gabor wavelet method proposed in [96]. The quality score of each block is calculated as the absolute difference between the mean values of the phase congruency maps of the matched blocks in the synthesized and reference image:

$$Q_i = |\mu(PC_{si} - PC_{ri})| \tag{11}$$

where $\mu()$ represents the mean value of the corresponding phase congruency map, the PC_{si} and PC_{ri} indicate the PC map of the matched blocks in the synthesized and reference image. The final image quality is obtained by averaging the quality score of all the blocks.

Farid *et al.* proposed a cyclopean eye theory [97] and divisive normalization (DN) transform [98] based Synthesized Image Quality Evaluator (SIQE) in [62]. The DIBR-synthesized view image associated with the left and right side views are firstly transformed by DN. Then, the statistical characteristics of the cyclopean image are estimated from

the DN representations of the left and right side views while the statistical characteristics of the synthesized image are obtained directly from itself. The similarity (Bhattacharyya coefficient [99]) between the distribution of the cyclopean and the synthesized image's DN representations is computed to measure the quality score of the synthesized image.

The SIQE metric only considers the texture information, in [63], Farid *et al.* proposed an extended version of SIQM by considering both the texture and depth information. The depth distortion estimation is based on the fact that the edge regions in a depth image are more sensitive to noise than the flat homogeneous regions since the distorted edge in the depth map may cause very annoying structural distortions in the synthesized image. Firstly, the pixels in the depth map with a high gradient value are extracted as noise sensitive pixels (NSP). Then, for each NSP, a local histogram from the distorted depth map is constructed and analysed to estimate the distortion in the depth image. The overall depth distortions are calculated by averaging the distortions in the left and right depth image. The final quality of the synthesized view is pooled from the texture and depth distortions.

4.3. Depth image quality metrics

The quality of depth images is crucial for generating high-quality synthesized views. A few metrics have been proposed to predict the depth image quality in DIBR view synthesis.

Le *et al.* proposed a RR depth image quality metric (RR-DQM) [55] which requires a pair of color and depth images. The depth image quality is measured depending on the edge directions based on the fact that the local depth distortion and the local image characteristic are strongly correlated. A Gabor filter is applied to generate a weighting map which are then used to adaptively weight the local depth distortion.

Li *et al.* proposed a FR depth image quality metric based on weighted edge similarity [54]. Based on their observation that the distortions in the DIBR-synthesized views are mainly concentrated in the edge regions of depth maps, the proposed metric is designed with emphasis on the distortions in depth edge regions. The similarity between the distorted and reference depth map is calculated in both intensity and gradient domains. Then, a weighting map is generated by combining a location prior and a depth distance measure. Finally, the edge indication is used as a guidance to pool the overall quality of depth map.

Farid *et al.* proposed a blind depth quality metric (BDQM) [56] to evaluate the compression distortions in depth images. They noticed that the compression flattens the sharp transitions of the depth image. Therefore, the shape of the histogram around the depth boundaries are used to predict the depth quality.

In [57], Xiang *et al.* proposed a NR depth image quality metric by calculating the misalignment errors between the edges of texture and depth images. The misalignments are evaluated from three similarities: the edge orientation similarity, the spatial similarity and the segmentation length similarity. Finally, the misalignments are used to calculate the final quality scores.

Li *et al.* proposed a NR depth image quality index based on the statistics of edge profiles (SEP) [58]. The first-order and second-order statistical features are firstly extracted based on edge profiles which are the neighbouring regions around the depth edges. Then, the random forest (RF) is applied to build a quality assessment model for depth maps.

The depth image quality metrics can evaluate the quality of synthesized view before performing actual rendering and is thus more computational friendly. It can also be used in the rate distortion optimization of depth map compression. The same as the texture IQA metrics, the NR depth image quality metrics are more practical than the FR ones since the depth maps are usually acquired by depth cameras or depth estimation approaches and are not always available.

4.4. NR metrics

In this part, we will review the NR metrics which do not need ground truth images/videos to evaluate the quality of DIBR-synthesized views.

4.4.1. Local image description based NR metrics

Due to the distorted depth map and imperfect rendering method, there exists a large number of structural and geometric distortions in the DIBR-synthesized views. As introduced in the RRLP metric [53], the structural distortions may result in local disorder in the image. Similarly, several local image description based NR metrics have been proposed to evaluate the structural distortions by measuring the local inconsistency via different models.

Gu *et al.* proposed an auto-regression (AR) based model (APT) to capture the geometric distortions in the DIBR-synthesized views. For each pixel, a local AR model (3×3) is first used to construct a relationship between this pixel and its neighbouring pixels.

$$x_i = \Omega(x_i)s + d_i \tag{12}$$

where $\Omega(x_i)$ denotes a vector which is composed of the neighbouring pixels of x_i in the (3×3) patch, s is a vector of AR parameters and d_i represents the error difference between the current pixel value and its corresponding AR prediction. The AR parameters are solved on the assumption that the 7×7 local patch, which consists of the current pixel and its 48 adjacent pixels, shares the same AR model. The error difference map between the synthesized and the reconstructed images is obtained as the distortion map. Then, a Gaussian filter and a saliency map [100] associated with a maximum pooling are used to obtain the final image quality score. Due to its computational complexity, this method owns a high computing cost.

Different from the APT metric, the OUT (outliers) metric [65] proposed by Jakhetiya *et al.* uses a median filter to calculate the difference map. Then, two thresholds are used to extract the structural and geometric distortion regions. The quality score is finally obtained from the standard deviation of the structural and geometric distortion regions.

These local image description based metrics can only detect thin distortions or local noise, they do not work well on the large size distortions.

4.4.2. Morphological operation based NR metrics

The morphological operations show their effectiveness in the FR metric MP-PSNR [40]. In [68, 69], Tian *et al.* proposed two metrics NIQSV and NIQSV+ to detect the local thin structural distortions through morphological operations. These two metrics assume that the "perfect" image consists of flat areas and sharp edges, so such images are insensitive to the morphological operations while the local thin structural distortions can be easily detected by these morphological operations. The NIQSV metric firstly uses an opening operation to detect the thin distortions and followed by a closing operation with larger Structural Element (SE) to file the black holes. The NIQSV+ extend the NIQSV by proposing two additional measurements: black hole detection and stretching detection. The black hole distortion is estimated by counting the black hole pixels proportion in the image while the stretching distortion is evaluated by calculating the gradient decrease of the stretching region and its adjacent non-stretching region.

Due to the limitation of the assumption and the SE size, these two metrics do not work well on the distortions in complex texture and the distortions with large size.

4.4.3. Sharpness detection based NR metric

Sharpness detection has been widely used in 2D image quality assessment [101, 102, 103] and also in the side view based FR metric LOGS [60]. In this part, we will introduce its usage in NR metrics. Sharpness is one of the most important measurements in NR image quality assessment [104, 105, 106]. The DIBR view synthesis may introduce multiple distortions such as blur, geometric distortions around the object edges, which may significantly result in the degradation of sharpness.

Nonlinear morphological wavelet decomposition can extract high-pass image content while preserving the unblurred geometric structures [40, 39]. In the transform domain, geometry distorted areas introduced by DIBR-synthesis are characterized by coefficients of higher value compared to the coefficients of smooth, edge and textural areas. In [67], Sandić-Stanković *et al.* proposed a wavelet-based NR metric (NR_MWT) for the DIBR-synthesized view videos. The sharpness is measured by quantifying the high frequency components in the image, which are represented by the high-high wavelet sub-band. The final quality is obtained from the sub-band coefficients whose value are higher than the threshold. Similar to MW-PSNR and MP-PSNR [40, 39], the NR_MWT also ~~achieved~~ has a very low computational complexity.

Differently, in CLGM [71], the sharpness is measured as the distance of standard deviations between the synthesized image and its down-sampled version. Besides, two additional distortions, dis-occluded regions and stretching, are also taken into consideration in CLGM. The dis-occluded regions are detected through an analysis of local image similarity. Similar to NIQSV+ [69], the stretching distortion is estimated by computing the similarity between the stretching region and its adjacent non-stretching region.

In [72], Wang *et al.* also proposed a NR metric (GDIC) to measure the geometric distortions and image complexity. Firstly, different from the wavelet transform based metrics introduced above, this GDIC metric uses the edge map of wavelet sub-bands to obtain the shape of geometric distortions. Then, the geometric distortion is measured ~~by~~ in terms of edge similarity between the wavelet low-level and high-level sub-bands [107]. Besides, the image complexity is also an important factor in human visual perception. In order to evaluate the image complexity of the DIBR-synthesized images, hybrid filter [108, 109], which combines the Autoregressive (AR) and bilateral (BL), is used. The final image quality score is computed by normalizing the geometric distortion with image complexity. Furthermore, in [73], this metric is extended to achieve higher performance by adding a log-energy based sharpness detection module.

4.4.4. Flicker region based video NR metrics

In DIBR-synthesized videos, temporal flicker is one of the most annoying distortions. Extracting the flicker regions may help to evaluate the quality of DIBR-synthesized videos.

In [75], Kim *et al.* also proposed a NR metric (CTI) to measure the temporal inconsistency and flicker regions in the DIBR-synthesized video. First, the flicker regions are detected from the difference between motion-compensated consecutive frames. Then, the structural similarity between consecutive frames are calculated on the flicker regions to measure the structural distortions in each frame. At the same time, the number of pixels in the flicker regions is used to weight the distortion of each frame. The final quality score is obtained as the weighted sum of the quality scores of all the frames in the DIBR-synthesized video.

In [76], Zhou *et al.* proposed a NR metric FDI to measure the temporal flickering distortion in the DIBR-synthesized videos. Firstly, the gradient variations between each frame are used to extract the potential flickering regions. Followed by a refinement to precisely obtain the flickering regions through calculating the correlation between the candidate flickering regions and their neighbours. Then, the flickering distortion is estimated in SVD domain from the difference between the singular vectors of the flickering block and their associated block in the previous frame. The final video quality is computed as the average quality of all the frames.

4.4.5. Natural Scene Statistics based NR metrics

Natural Scene Statistics (NSS) based approaches, which assume that the natural images contain certain statistics and these statistics may be changed by different distortions, have achieved great success in the quality assessment of traditional 2D images [110, 111, 112, 113]. Due to the big difference between the DIBR view synthesis distortions and the traditional 2D ones, these NSS based metrics do not work well on the quality assessment of DIBR-synthesized views. Recently, several efforts have been made to fix this gap.

As introduced in the previous Edge/Contour based FR metrics part, the edge image is significantly degraded by structural and geometric distortions in DIBR-synthesized images, and the edge based FR metrics have shown their superiority. With this ~~view~~ consideration, Zhou *et al.* proposed a NR metric (SET) for DIBR-synthesized images via edge statistics and texture naturalness based on Difference-of-Gaussian (DoG) in [74]. The orientation selective statistics (similar to the metric in [112]) are extracted from ~~different scale~~ DoG images at different scales while the texture naturalness features are obtained based on the Gray level Gradient Co-occurrence Matrix (GGCM) [114] which represents the joint distribution relation of pixel gray level and edge gradient. A Random Forest (RF) regression model is finally trained based on these two groups of features to predict the quality of DIBR-synthesized images.

Gu *et al.* proposed a self-similarity and main structure consistency based Multiscale Natural Scene Statistics (MNSS) in [66]. The multiscale analysis on the DIBR-synthesized image and its associated reference image indicates that the distance (SSIM value [4]) between the synthesized and the reference image decreases significantly when the scale reduces. It is assumed that the synthesized image at a higher scale holds a better quality, which means the ~~higher scale~~ higher-scale images can be approximately used as references. Thus, the similarity between the lower scale image (first scale is used in this metric) and the higher scale images (self similarity) are used to measure the quality of DIBR-synthesized image. Besides, in the main structure ~~NSS model~~, the authors use 300 natural images from the Berkeley segmentation dataset [115] to obtain the general statistical regularity of main structure in natural images. The similarity between the main structure map of the synthesized image and the obtained prior NSS vector is calculated to evaluate the structure degradation of the DIBR-synthesized image. Finally, the statistical regularity of main structure and the structure degradation are combined to get the overall quality score.

Shao *et al.* propose a NR metric (HEVSQP) for DIBR-synthesized videos based on color-depth interactions in [70]. Firstly, the video sequence is divided into Group of Frames (GoF). Through an analysis of color-depth interactions, more than 90 features from both texture and depth videos, including gradient magnitude, asymmetric generalized Gaussian distribution (AGGD) [111], local binary pattern (LBP), are extracted. Then, a principal component analysis (PCA) is applied to reduce the feature dimension. Then, two dictionaries, color dictionary and depth dictionary, are learned to establish the relationship between the features and video quality. The final quality score is pooled from the color and depth quality.

In [77], Ling *et al.* proposed a NR learning based metric for DIBR-synthesized views, which focuses on the non-uniform distortions. Firstly, a set of convolutional kernels are learned by using the improved fast convolutional sparse coding (CSC) algorithms. Then, the convolutional sparse coding (CSC) based features of the DIBR-synthesized images are extracted, from which the final quality score is obtained via support vector regression (SVR).

Although the NSS models have made great progress for the NR IQA, the hand-craft features may not be sufficient to represent complex image textures and artefacts, there is still ~~exists~~ a large gap between objective quality measurement and human perception [116].

4.4.6. Deep feature based NR metrics

The deep learning techniques, especially the Convolutional Neural Networks (CNN), have shown their great advantages in various computer vision tasks [117, 118]. They make it possible to directly learn the representative features from image [119, 120]. Unfortunately, ~~owing~~ due to the limitation of ~~size of~~the number of images in the DIBR-synthesized view datasets, there is not enough data to train the deep models straightforwardly. However, it is shown in the recent published literature that the deep neural network models trained on large-scale datasets, *e.g.* ImageNet [121], can be used to extract effective representative features of human perception.

In [78], Wang *et al.* proposed a NR metric SIQA-CFP which uses the ResNet-50 [122] model pre-trained on ImageNet to extract multi-level features of DIBR-synthesized images. Then, a contextual multi-level feature pooling strategy is designed to encode the high-level and low-level features, and finally to get the quality scores.

As introduced in Section 1, various distortions may be introduced during the dis-occlusion region filling stage. Meanwhile, in current literature, several Generative Adversarial Networks (GAN) [123] based models have been proposed for image in-painting. As the generator is trained to in-paint the missing part, the discriminator is supposed to have the capability to capture the perceptual information which reflects the in-painted image quality. Based on this assumption, Ling *et al.* proposed a GAN based NR metric (GANs-NRM) [79] for DIBR-synthesized images. In GANs-NRM, a generative adversial network for image in-painting is firstly trained on two large-scale datasets (PASCAL [124] and Places [125]). Then, the features extracted from the pre-trained discriminator are used to learn a Bag-of-Distortion-Word (BDW) codebook. A Support Vector Regression (SVR) is trained on the encoded information of each image to predict the final quality of DIBR-synthesized images. Instead of simply using the general models trained for other tasks, *e.g.* object detection, this metric is more targeted, and it also proposes a new way to obtain the semantic features for image quality assessment.

4.5. Summary

In this section, 19 FR, 3 RR, 4 SV-FR and 15 NR DIBR quality metrics have been reviewed and categorized based on their used approaches and on the amount of reference information used. As shown in Table 2, most of the metrics consist of multiple parts~~, it~~ It is thus difficult to classify them into a single specific category thoroughly, that is why we just classify them into the most related one instead. Besides, there are also some other ways to do the classification. For example, if we focus on the image structural representation

Table 3
Performance of the DIBR dedicated metrics on DIBR-synthesized image dataset.

Metric		IVC image dataset			IETR image dataset			MCL 3D image dataset			IVY dataset		
		PLCC	RMSE	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE	SROCC
FR 2D	PSNR	0.4557	0.5927	0.4417	0.6012	0.1985	0.5356	0.7852	1.6112	0.7915	**0.6311**	**19.1227**	**0.6668**
	SSIM [4]	0.4348	0.5996	0.4004	0.4016	0.2275	0.2395	0.7331	1.7693	0.7470	0.3786	22.8172	0.3742
NR 2D	BIQI [127]	0.5150	0.5708	0.3248	0.4427	0.2223	0.4321	0.3347	2.4516	0.3696	0.5686	20.2791	5754
	BLIINDS2 [110]	0.5709	0.5467	0.4702	0.2020	0.2428	0.1458	0.6338	2.0124	0.5893	0.3508	23.0855	0.2569
FR DIBR	Bosc [35]	0.5841	0.5408	0.4903	—	—	—	0.4536	2.2980	0.4330	—	—	—
	3DSwIM [37]	0.6864	0.4842	0.6125	—	—	—	0.6519	1.9729	0.5683	—	—	—
	VSQA [36]	0.6122	0.5265	0.6032	0.5576	0.2062	0.4719	0.5078	2.9175	0.5120	—	—	—
	ST-SIAQ [42]	0.6914	0.4812	0.6746	0.3345	0.2336	0.4232	0.7133	1.8233	0.7034	—	—	—
	EM-IQA [43]	0.7430	0.4455	0.6282	0.5627	0.2020	0.5670	—	—	—	—	—	—
	MP-PSNR [40]	0.6729	0.4925	0.6272	0.5753	0.2032	0.5507	0.7831	1.6179	0.7899	0.5947	19.8182	0.5707
	MW-PSNR [39]	0.6200	0.5224	0.5739	0.5301	0.2106	0.4845	0.7654	1.6743	0.7721	0.5373	20.7910	0.5051
	SCDM [47]	0.8242	0.3771	0.7889	**0.6685**	**0.1844**	**0.5903**	0.7166	1.8141	0.7197	—	—	—
	SC-IQA [48]	**0.8496**	**0.3511**	0.7640	**0.6856**	**0.1805**	**0.6423**	**0.8194**	**1.4913**	**0.8247**	0.4326	22.2256	0.3135
	Wang [51]	**0.8512**	**0.3146**	**0.8346**	0.6118	0.1961	0.6136	**0.7910**	**1.5917**	**0.7929**	—	—	—
	CBA [23]	—	—	—	—	—	—	—	—	—	**0.826**	**8.181**	**0.829**
RR DIBR	MP-PSNRr [52]	0.6954	0.4784	0.6606	0.6061	0.1976	0.5873	0.7740	1.6474	0.7802	0.5384	20.7733	0.5454
	MW-PSNRr [52]	0.6625	0.4987	0.6232	0.5403	0.2090	0.4946	0.7579	1.7012	0.7665	0.5304	20.8993	0.5138
SV-FR DIBR	SIQE [62]	0.7650	0.5382	0.4492	0.3144	0.2353	0.3418	0.6734	1.9233	0.6976	—	—	—
	LOGS [60]	**0.8256**	**0.3601**	0.7812	**0.6687**	**0.1845**	**0.6683**	0.7614	1.6873	0.7579	**0.6442**	**18.8553**	**0.6385**
	DSQM [61]	0.7430	0.4455	0.7067	0.2977	0.2367	0.2369	0.6995	1.8593	0.6980	—	—	—
NR DIBR	APT [64]	0.7307	0.4546	0.7157	0.4225	0.2252	0.4187	0.6433	1.9870	0.6200	0.5156	21.1239	0.4754
	OUT [65]	0.7243	0.4591	0.7010	0.2007	0.2429	0.1924	0.4208	2.3601	0.3171	0.2525	23.8530	0.2409
	MNSS [66]	0.7700	0.4120	0.7850	0.3387	0.2333	0.2281	0.3766	2.4101	0.3531	0.3834	22.7681	0.2282
	NR_MWT [67]	0.7343	0.4520	0.5169	0.4769	0.2179	0.4567	0.1373	2.5771	0.0110	0.4848	21.5614	0.4558
	NIQSV [68]	0.6346	0.5146	0.6167	0.1759	0.2446	0.1473	0.6460	1.9820	0.5792	0.4113	22.4706	0.2717
	NIQSV+ [69]	0.7114	0.4679	0.6668	0.2095	0.2429	0.2190	0.6138	2.0375	0.6213	0.2823	23.6491	0.3823
	SET [74]	**0.8586**	**0.3015**	**0.8109**	—	—	—	**0.9117**	**1.0631**	**0.9108**	—	—	—
	GANs-NRM [79]	0.826	0.386	**0.807**	0.646	0.198	0.571	—	—	—	—	—	—

"—" : Due to the unavailability of source code or reference resources *e.g.* depth map and side view reference image, we just use the reported results in their corresponding publications instead, their associated results on other datasets are marked by the symbol "—" in the table.

used in these metrics, they can be classified into low-level [22]), mid-level [42, 43] and high-level [77, 78, 79] metrics. As introduced in [126], the low-level representations indicate the pixel level edges or contours; the mid-level representations mean the shapes and texture information; the high-level representations refer to the complex features *e.g.* objects, unnatural structures. Besides, there are also some hierarchical metrics which combine the above features, such as the LMS metric proposed in [49] which uses both low-level and mid-level features [42] and the metric in [50] which integrates the features on each level.

5. Experimental results and discussions

In this section, the performances of different objective quality assessment metrics are presented and analysed. Besides, some potential challenges and possible directions for future work will be discussed.

5.1. Performance evaluation methodologies

The subjective test results can be recognized as the ground truth visual quality since the human observer is the ultimate receiver of image/video content. The accuracy of an objective quality metric can be evaluated based on its consistencies with the subjective quality scores. In this part, we will introduce the Video Quality Expert Group (VQEG) [128] recommended correlation based methods and the recently proposed Krasula' model [129] in detail.

5.1.1. Correlation coefficients based methods

The reliability of objective metrics can be evaluated through their correlation with subjective test scores. Three widely used criteria, Pearson Linear Correlation Coefficients (PLCC) and Root-Mean-Square-Error (RMSE) and Spearman Rank-Order Correlation Coefficients (SROCC), are recommended by VQEG to evaluate the prediction accuracy, prediction monotonicity and prediction consistency of the objective metrics respectively, which are defined as follows:

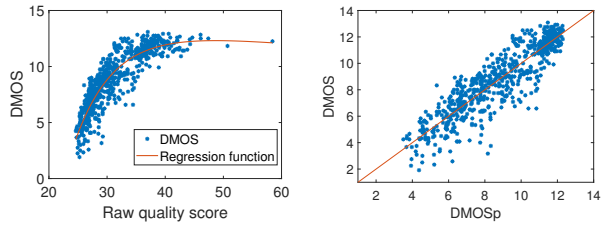
$$PLCC(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (13)$$

$$RMSE(X,Y) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2} \quad (14)$$

$$SROCC(X,Y) = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (15)$$

where d_i indicates the difference of ranking of X and Y. Higher PLCC and SROCC valuest indicate higher accuracy and better monotonicity respectively. On the contrary, a higher RMSE value refers to a lower prediction accuracy.

Before computing these three criteria, the objective scores are recommended by VQEG to be mapped to the predicted subjective score $DMOS_p$ to remove the nonlinearties due to the subjective rating processing and to facilitate comparison



(a) Before regression (b) After regression

Figure 8: Example relationship between DMOS and objective quality scores. This figure is from [130].

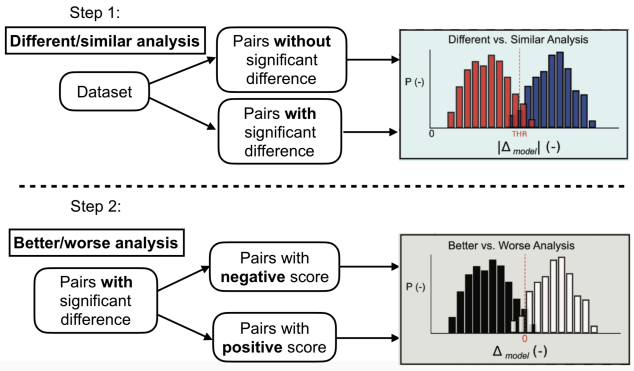


Figure 9: Krasula's model for performance evaluation of objective quality metrics [129].

of the metrics in a common analysis space [128]. The nonlinear function for regression mapping is shown as follows:

$$DMOS_p = \beta_1(0.5 - \frac{1}{(1+e^{(\beta_2(s-\beta_3))})}) + \beta_4 s + \beta_5 \quad (16)$$

where s is the score obtained by the objective metric and $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the parameters of these regression functions. They are obtained through regression to minimize the difference between $DMOS_p$ and $DMOS$. As shown in Fig. 8, the nonlinearity has been removed after the regression.

5.1.2. Analysis of Krasula's model

The above methods compare the performance of each metric by calculating their correlations with the subjective results. However they only consider the mean value of subjective scores, the uncertainty of the subjective scores are ignored. In addition, the quality scores need to be regressed by a regression function *cf.* Eq. 16, that is not the way they are exactly used in real scenarios. Thus, we further conduct a statistical test proposed by Krasula *et al.* in [129] which does not suffer from the drawbacks of the above methods. The performances of objective metrics are evaluated by their classification abilities.

As shown in Fig. 9, firstly, the tested image pairs in the dataset are divided into two groups: different and similar according to their subjective scores. The cumulative distribution function (~~edf~~ CDF) of the normal distribution is used to calculate the probability of image pairs. Then, we consider the pairs with a probabilty higher than the selected significance level 0.95 to be significantly different. The others will be recognized as similar.

There are two performance analyses. The first performance analysis is conducted by ~~by~~ evaluating how well the objective metric succeeds to distinguish significantly different image pairs from ~~unsignificantly~~ non-significantly different video pairs, in a consistent way with subjective evaluation of significant difference. ~~In the case of the two videos in the pair are significantly different according to the subjective results.~~ The second analysis determines whether the objective metric can correctly identify the image of higher quality in the pair.

Compared to simply calculating the correlation coefficients, this model considers not only the mean value of subjective scores, but also their uncertainties. Besides, since no regression is used, this model less depends on the quality ranges of different datasets. Another advantage of Krasula's model is that it can easily combine the data from multiple datasets and evaluate a comprehensive performance on multiple datasets instead of simply averaging the results on different datasets.

5.2. Performance on DIBR image datasets

5.2.1. Results of PLCC, RMSE and SROCC

The obtained PLCC, RMSE and SROCC values of the objective image quality assessment metrics on the DIBR-synthesized image datasets are given in Table 3, in which four 2D metrics [127, 110, 4] and 24 DIBR metrics are tested. The best three performances among the blind IQA methods are shown in bold. We can easily observe that the DIBR-synthesized view dedicated metrics significantly outperform the traditional 2D metrics on the IVC and IETR image datasets which focus on the DIBR view synthesis distortions. In other words, the metrics initially designed for traditional 2D image distortions can not well evaluate the DIBR view synthesis distortions.

The shift compensation based FR and SV-FR metrics obtain great improvement compared to the original 2D FR metrics, ~~eg.~~ *e.g.* the SC-IQA compared to PSNR. One main reason is that the global object shift existing in the DIBR-synthesized images may not be perceived by human observers but can be easily detected by the original 2D pixel-based FR metrics. ~~So,~~ Thus this shift distortions are often overestimated by the 2D pixel-based FR metrics.

If we focus on the wavelet transform-based metrics (NR_MWT and MW-PSNR), the NR metric (NR_MWT ~~etc.~~) performs better than the FR metric (MW-PSNR) on the IVC dataset. It is surprising that the FR metric performs even worse than the NR metric since these metrics use similar features and the FR metric has access to the ground truth. While on the IETR dataset, the NR metric performs worse than the FR metric~~s~~. The main reason ~~is~~ probably ~~also be~~ lies in the global shift distortion in the IVC image dataset.

To further explore the object shift effect, we have made an additional experiment on the IVC dataset while excluding

Table 4
Performance on the IVC DIBR image dataset excluding A1 algorithm.

Metric		PLCC	RMSE	SROCC
FR 2D	PSNR	0.7519	0.4525	0.6766
	SSIM	0.5956	0.5513	0.4424
FR DIBR	MW-PSNR	0.8545	0.3565	0.7750
RR DIBR	MW-PSNRr	0.8855	0.3188	0.8298

the A1 view synthesis algorithm [14] which causes great object shift in the synthesized views. The A1 algorithm fills the black holes in the dis-occlusion regions by simply stretching the adjacent texture which may cause great global object shift in the synthesized views. The results are shown in Table 4. We can observe that the performance of FR and RR metrics increase significantly when large global shift artefacts are excluded.

The edge/contour based metrics also perform much better than the 2D pixel-based FR metrics since the edge/contour features can better represent the geometric degradations in the DIBR-synthesized images compared to simple pixel information.

The NR metrics do not need any reference information to evaluate the image quality, ~~so~~ thus the global shift does not have effect on the NR metrics. Besides, since the real reference images at virtual viewpoints are not always available in real applications, the NR metrics are more practical and useful. From Table 3, we can easily find that the performances of the DIBR-synthesized view dedicated metrics decrease greatly ~~in~~ on the IETR dataset compared to ~~their performance in~~ theirs on the IVC dataset. Among these metrics, the NR ones decrease the most, especially the learning based NR metrics. This is because of the fact that these NR metrics ~~focus on~~ are designed for the distortions in the IVC dataset~~, but~~. However, in the IETR dataset, many "old fashioned" distortions are excluded.

As introduced in Section II, the MCL-3D dataset does not focus on the DIBR view synthesis distortions, but on the traditional distortion effects on the synthesized views. Thus, the performances of the tested objective metrics are quite different. Some of the metrics (Bosc, VSQA and NR_MWT) that only consider the DIBR view synthesis distortions perform not as good as the traditional 2D metrics. Some 2D related FR metrics perform even worse than their ~~original version~~ backbones. For instance, VSQA and 3DSwIM ~~metrics~~ can not achieve the performance of SSIM; SCDM, MP-PSNR and MW-PSNR ~~metrics~~ perform worse than PSNR. Among these metrics, the feature-based FR metrics perform better than the simple edge/contour based metrics. It can be inferred that the frequency domain features can represent not only the edge/contour information, but also some other texture characteristics. The SET metric contains not only the DoG features for the DIBR view synthesis distortions, but also the GGCM based features for the texture naturalness. That may explain its good performance on both IVC and MCL-3D datasets.

The IVY dataset considers not only the view synthesis distortion, but also de binocular asymmetry in synthesized stereoscopic images. The baseline distance between the virtual viewpoint and the original viewpoint is much bigger than that in ~~the~~ other datasets. Thus, the metrics which do not consider the binocular asymmetry perform not well on this dataset.

5.2.2. Results of Krasula's model

Only the IVC and IETR datasets are tested in this part since the MCL-3D and IVY datasets do not provide the standard deviation which represents the subject uncertainty. The obtained Area Under the Curves (AUC) and significant test results on IVC and IETR are shown in Fig. 10 (a) (b) (c) (d). The Fig. 10 (e) and (f) demonstrate the results on the combination of IVC and IETR datasets. A higher AUC value indicates a higher performance. In the significant test results, the white block indicates that the metric in the row performs significantly better that the metric in the column and vice versa for the black block. The gray block means these two metrics are statistically equivalent.

In the first different / similar analysis on the IVC dataset *cf.* Fig. 10 (a), none of these metrics perform well since most AUC values are below 0.7 and there even exist some metrics whose AUC values are under 0.5. Generally, the DIBR FR metrics perform better than the other metrics.

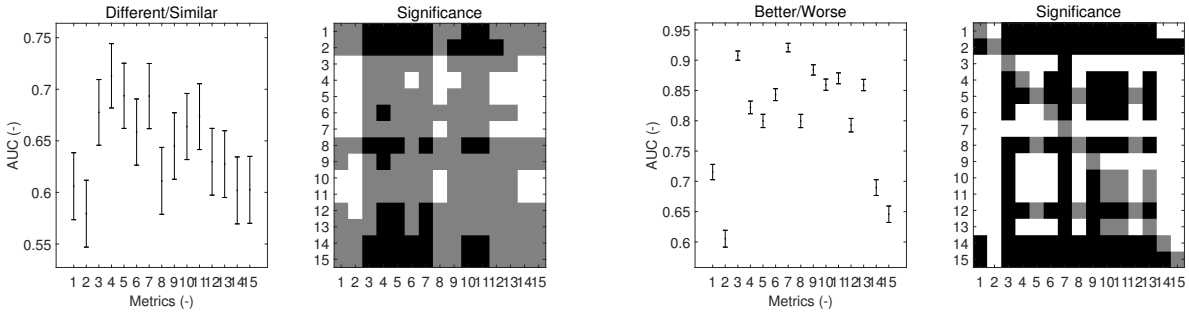
In the second different / similar analysis on the IVC dataset *cf.* Fig. 10 (b), the DIBR-synthesized view dedicated metrics perform significantly better than the 2D metrics (first and last 2 metrics) since the DIBR metrics can achieve higher AUC values. Among these metrics, the SCDM and SC-IQA metrics perform the best, they achieve AUC values higher than 0.9.

The results on the IETR dataset *cf.* Fig. 10 (c) (d) and the combination of the two datasets *cf.* Fig. 10 (e) (f) show that most of the FR metrics outperform the NR metrics except the SSIM metric. The 2D NR metrics achieve similar results compared to their performance on IVC dataset, while the performance of the DIBR NR metrics decrease greatly compared to their performance on IVC dataset. The results of Krasula's model are consistent with the correlation coefficients results in the previous part.

5.3. Performance on DIBR video datasets

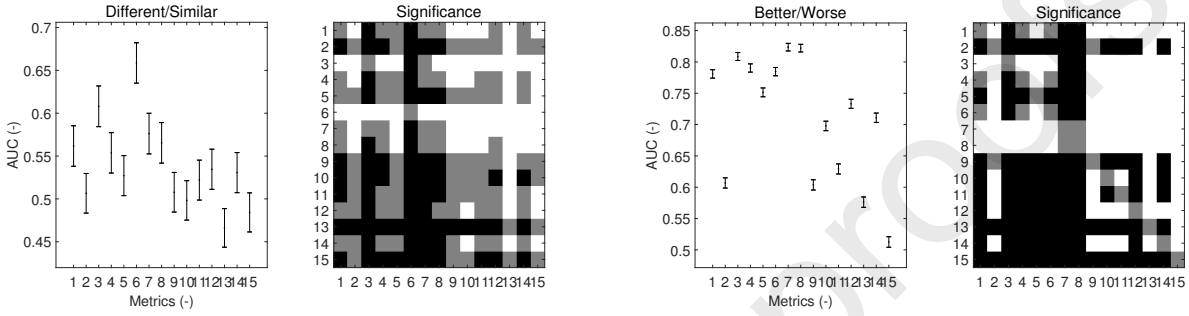
The DIBR-synthesized videos contain some temporal distortions, such as flickering, in addition to the spatial distortions in images. In this experiment, 12 state-of-the-art DIBR image metrics in addition to 5 DIBR video metrics are tested. To compare the performance of DIBR metrics and traditional 2D metrics, 5 widely used 2D video metrics and 2 2D image metrics are tested. The quality scores of image metrics are obtained by averaging the quality of all the frames. The three metrics which performance the best among the BIQA methods are marked in bold.

The obtained PLCC, RMSE and SROCC values on IVC video and SIAT video datasets are given in Table 5. Only the results of Krasula's model on IVC video dataset are shown



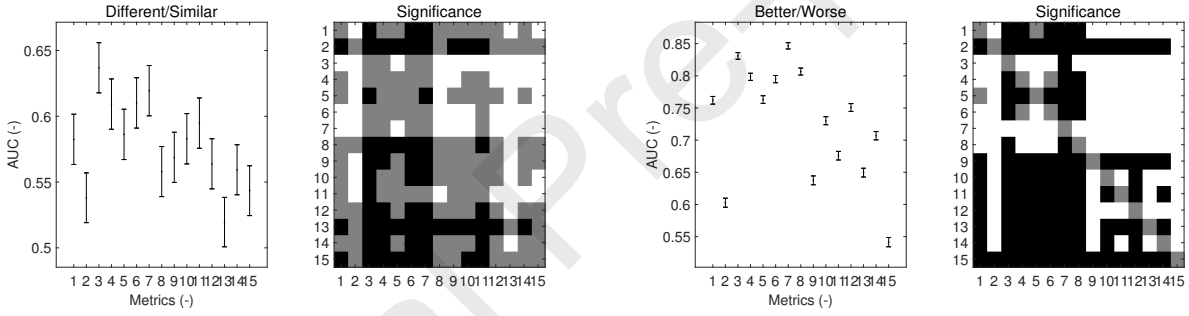
(a) Different / Similar analysis on IVC image dataset

(b) Better / Worse analysis on IVC image dataset



(c) Different / Similar analysis on IETR image dataset

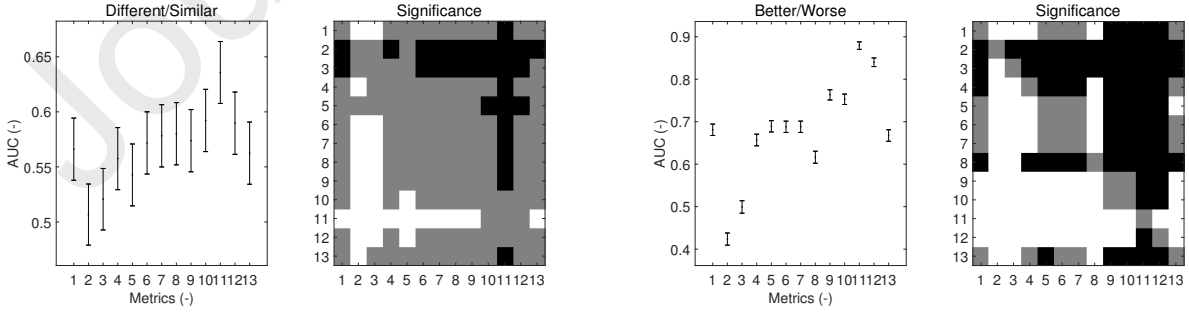
(d) Better / Worse analysis on IETR image datasets



(e) Different / Similar analysis combining two datasets

(f) Better / Worse analysis combining two datasets

Figure 10: Performance on IVC and IETR image datasets using Krasula's model. The metrics 1-15 indicate PSNR, SSIM, SCDM. MP-PSNRr, MW-PSNRr, EM-IQA, SC-IQA, LOGS, NIQSV+, APT, MNSS, NR_MWT, OUT, BIQI, BLiindS2 respectively. In the significant test results, the white block indicates that the metric in the row performs significantly better that the metric in the column and vice versa for the black block. The gray block means these two metrics are statistically equivalent.



(a) Different / Similar analysis on IVC Video dataset

(b) Better / Worse analysis on IVC Video dataset

Figure 11: Performance on IVC video dataset using Krasula's model. The metrics 1-13 represent: PSNR, SSIM, SpEED, ST-RRED, VIIDEO, MP-PSNRr, MW-PSNRr, NIQSV, OUT, MNSS, NR_MWT, FDI, SIAT-VQA respectively. In the significant test results, the white block indicates that the metric in the row performs significantly better that the metric in the column and vice versa for the black block. The gray block means these two metrics are statistically equivalent.

Table 5
Performance on the IVC and SIAT DIBR video dataset.

Metric		IVC video dataset			SIAT video dataset		
		PLCC	RMSE	SROCC	PLCC	RMSE	SROCC
FR 2D image metrics	PSNR	0.5104	0.5690	0.4647	0.6525	0.0972	0.6366
	SSIM [4]	0.4081	0.6041	0.3751	0.4528	0.1144	0.4550
FR 2D video metrics	MOVIE [131]	0.4971	0.4903	0.3877	0.646	0.097	0.693
	ST-RRED [132]	0.2025	0.6480	0.5777	**0.7164**	**0.0895**	**0.6971**
NR 2D video metrics	SpEED [133]	0.3771	0.6128	0.5952	**0.7236**	**0.0885**	**0.6987**
	VIIDEO [134]	0.5971	0.5308	0.5877	0.2586	0.1239	0.2535
FR DIBR image metrics	Bosc [35]	0.5856	0.4602	0.2654	0.453	0.114	0.431
	MP-PSNR [40]	0.5026	0.5720	0.5478	0.5681	0.1056	0.5044
	MW-PSNR [40]	0.4911	0.4638	0.4558	0.5745	0.1050	0.5024
	3DSwIM [37]	0.4822	0.4974	0.3320	0.5677	0.1057	0.2762
RR DIBR image metrics	MP-PSNRr [52]	0.4617	0.5869	0.5307	0.5640	0.1059	0.5040
	MW-PSNRr [52]	0.4802	0.5804	0.5038	0.5757	0.1049	0.5853
SV-FR DIBR image metrics	SIQE [62]	0.4084	0.5138	0.0991	0.3627	0.1195	0.2586
	DSQM [61]	0.5241	0.4857	0.3157	0.4001	0.1071	0.3994
NR DIBR image metrics	OUT [65]	0.6762	0.4874	0.6151	0.0945	0.1277	0.0926
	NR_MWT [67]	**0.7530**	**0.4354**	**0.7145**	0.5051	0.1107	0.3092
	NIQSV [68]	0.6505	0.5025	0.5963	0.5144	0.1100	0.4562
	MNSS [66]	0.5180	0.5660	0.5371	0.1591	0.1266	0.2463
FR DIBR video metrics	CQM [135]	0.4102	0.5101	0.3265	0.4021	0.1070	0.4064
	PSPTNR [44]	0.4321	0.5002	0.4152	0.4461	0.1069	0.4305
	VQA-SIAT [22]	0.5943	0.5321	0.5879	**0.8527**	**0.0670**	**0.8583**
NR DIBR video metrics	CTI [75]	**0.6821**	**0.4372**	**0.6896**	0.5736	0.1053	0.5425
	FDI [76]	**0.7576**	**0.4319**	**0.7162**	0.5952	0.1033	0.5425

in Fig. 11 since the SIAT video dataset does not provide the uncertainty of subject ratings.

The IVC video dataset focuses on the DIBR view synthesis distortions while the SIAT dataset focuses on the compression effects on the synthesized views. We can easily observe that the best three metrics on IVC dataset are all DIBR metrics while the best three metrics on SIAT dataset are VQA-SIAT and two 2D metrics. The VQA-SIAT metric mainly focuses on the compression effect which may lead obvious flicker in the DIBR-synthesized views. The spatial view synthesis distortions considered in this metric are very limited. That may explain why it significantly outperforms the other metrics on SIAT dataset while it can not obtain a very good performance on the IVC dataset. When we focus on the IVC video dataset, none of FR metrics achieves a high correlation with the subjective results. Moreover, there is no significant difference between the performances of DIBR FR and 2D FR metrics. However, the DIBR NR metrics perform the best compared to other metrics~~. The main reason is the same as that on IVC image dataset:~~, also due to the global shift effect.

5.4. Discussions

The experimental results show that although great progress has been made towards the quality assessment of synthesized views, there is still ~~significant~~ a large room for improvement.

5.4.1. Synthesized video quality assessment

The DIBR-synthesized videos contain not only the compression distortions but also the distortions induced by DIBR. The VQA-SIAT metric works well on capturing the temporal flicker caused by video compression, but it fails to assess the DIBR view synthesis distortions in the synthesized video frames. In addition, the imperfect view synthesis algorithms may also result in great miss-match between the adjacent frames in the synthesized video, which causes very annoying temporal distortions that the 8 by 8 block matching (in VQA-SIAT) may fail to detect. Therefore, we could try to further analyse the specific spatial-temporal distortions in the synthesized videos and design a complete metric for the DIBR-synthesized videos.

5.4.2. Quality assessment of synthesized views in real applications

As introduced previously, DIBR can be used in various applications, but the quality assessment for these applications are rarely researched. For example, the free viewpoint videos (FVV) and multi-view videos (MVV) provide the images from multiple viewpoints at the same time instant. The temporal distortions in FVV or MVV are mainly introduced by the changing of viewpoints instead of timeline [83, 50]. This type of distortions are different from that in normal DIBR-synthesized views videos. Besides, in order to provide immersive perception for the observer, the AR or VR applications need to generate multiple synthesized images

and change the viewpoint with the motion of the observer. The synthesized video contains both the inter-frame and inter-viewpoint temporal distortions, as well as the binocular asymmetric distortions which may happen in stereoscopic applications [23]. It could be interesting to try to design the metrics for these applications since they are currently rarely explored.

5.4.3. Deep learning approaches

The main limitation of the usage of deep learning on the quality assessment of DIBR-synthesized views is the limited size of available datasets. Unlike the homogeneous distortions in the traditional 2D images, the distortions in the DIBR-synthesized views mostly occur in the dis-occlusion regions. In other words, the major part of the DIBR-synthesized view holds a perfect quality. ~~The synthesized image can not be split into several patches and directly use the quality of the whole image as the quality of all the patches.~~ Thus we cannot split the synthesized image into several patches and then directly use the quality of the whole image as the quality of the patches. Creating a very large-scale dataset may significantly help to train a ~~good~~ deep model. But unlike the datasets for other tasks ~~eg.~~ , *e.g.* object recognition, creating an image quality dataset necessarily requires subjective tests which are quite expensive and time-consuming. Thus, exploring how to train a comprehensive model on limited data could be more practical, ~~eg.~~ maybe via one-shot learning or few-shot learning [136, 137]. Besides, in addition to the individual predicted image quality scores (precision), the ranking of the predicted scores (monotonicity) is also an important index to evaluate the performance of an IQA metric. Therefore, learning from rankings [138, 7] may help to solve the problem of IQA dataset size limit. Firstly, the ranked image sets can be automatically generated without subjective tests [138]. We can pre-train our model on the generated ranked image sets and then fine-tune it on the target IQA datasets. Secondly, a reliable ranking loss can enhance the ability of the model to rank images in terms of quality and thus help to generate more precise quality scores [7]. The fact that quality score of the whole synthesized image can not directly be distributed to all the image patches does not mean that the image can not be processed patch by patch. The main challenge is to find a proper pooling method to get the overall quality score. Although the pre-trained deep features have been successfully used in metrics [78, 79], more efforts could be made to create a more general and effective end-to-end deep model.

6. Conclusion

In this paper, we present an up-to-date overview for the quality assessment methods of DIBR-synthesized views. We firstly described the existing DIBR-synthesized view datasets. Secondly, we analysed and discussed the recently proposed state-of-the-art objective quality metrics for DIBR-synthesized views, and classified them into different categories based on their ~~used~~ approaches. Then, we conducted a reliable experiment to compare the performance of each metric, and analysed their advantages and disadvantages ~~at the same time~~. Furthermore, we discussed the potential challenges and directions for future research. We hope this overview can help to better understand the state-of-the-art of this research topic and provide insights to design better metrics and experiments for effective DIBR-synthesized images/videos quality evaluation.

Acknowledgment

The authors would like to thank Dr. Suiyi Ling and Dr. Yu Zhou for sharing their code. We would also like to thank Prof. Patrick Le Callet and Dr. Lucas Krasula for their kind advices on the experiment. This work was supported in part by the NSFC Project under Grants 61771321 and 61872429, in part by the Guangdong Key Research Platform of Universities under Grants 2018WCXTD015, in part by the Natural Science Foundation of Guangdong Province, China, under Grants 2020A1515010959, and in part by the Interdisciplinary Innovation Team of Shenzhen University.

References

[1] 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.

[2] W. Sun, L. Xu, O. C. Au, S. H. Chui, C. W. Kwok, An overview of free view-point depth-image-based rendering (dibr), in: APSIPA Annual Summit and Conference, 2010, pp. 1023–1030.

[3] L. Jiao, H. Wu, H. Wang, R. Bie, Multi-scale semantic image inpainting with residual learning and gan, Neurocomputing 331 (2019) 199–212.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612. doi:10.1109/TIP.2003.819861.

[5] Z. Wang, E. P. Simoncelli, A. C. Bovik, Multiscale structural similarity for image quality assessment, in: Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2004., Vol. 2, IEEE, 2003, pp. 1398–1402.

[6] L. Li, Y. Zhou, W. Lin, J. Wu, X. Zhang, B. Chen, No-reference quality assessment of deblocked images, Neurocomputing 177 (2016) 572–584.

[7] X. Jiang, L. Shen, L. Yu, M. Jiang, G. Feng, No-reference screen content image quality assessment based on multi-region features, Neurocomputing (2019).

[8] Q. Li, W. Lin, K. Gu, Y. Zhang, Y. Fang, Blind image quality assessment based on joint log-contrast statistics, Neurocomputing 331 (2019) 189–198.

[9] I. Ahn, C. Kim, A novel depth-based virtual view synthesis method for free viewpoint video, IEEE Transactions on Broadcasting 59 (4) (2013) 614–626.

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

[11] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, Y. Mori, Reference softwares for depth estimation and view synthesis, ISO/IEC JTC1/SC29/WG11 MPEG 20081 (2008) M15377.

[12] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, L. Morin, Towards a new quality metric for 3-d synthesized view assessment, IEEE Journal of Selected Topics in Signal Processing 5 (7) (2011) 1332–1343.

[13] A. Criminisi, P. Pérez, K. Toyama, Region filling and object removal by exemplar-based image inpainting, IEEE Transactions on image processing 13 (9) (2004) 1200–1212.

[14] A. Telea, An image inpainting technique based on the fast marching method, Journal of graphics tools 9 (1) (2004) 23–34.

[15] A. Oliveira, G. Fickel, M. Walter, C. Jung, Selective hole-filling for depth-image based rendering, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 1186–1190.

[16] S. M. Muddala, M. Sjöström, R. Olsson, Virtual view synthesis using layered depth image generation and depth-based inpainting for filling disocclusions and translucent disocclusions, Journal of Visual Communication and Image Representation 38 (2016) 351–366.

[17] G. Luo, Y. Zhu, Z. Li, 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.

[18] C. Zhu, S. Li, Depth image based view synthesis: New insights and perspectives on hole generation and filling, IEEE Transactions on Broadcasting 62 (1) (2016) 82–93.

[19] S. Tian, L. Zhang, L. Morin, O. Déforges, A benchmark of dibr synthesized view quality assessment metrics on a new database for immersive media applications, IEEE Transactions on Multimedia 21 (5) (2019) 1235–1247. doi:10.1109/TMM.2018.2875307.

[20] R. Song, H. Ko, C. Kuo, Mcl-3d: A database for stereoscopic image quality assessment using 2d-image-plus-depth source, Journal of Informatin Science and Engineering 31 (5) (2015) 1593–1611.

[21] E. Bosc, P. Le Callet, L. Morin, M. Pressigout, Visual quality assessment of synthesized views in the context of 3d-tv, in: 3D-TV system with depth-image-based rendering, Springer, 2013, pp. 439–473.

[22] X. Liu, Y. Zhang, S. Hu, S. Kwong, C.-C. J. Kuo, Q. Peng, Subjective and objective video quality assessment of 3d synthesized views with texture/depth compression distortion, IEEE Transactions on Image Processing 24 (12) (2015) 4847–4861.

[23] Y. J. Jung, H. G. Kim, Y. M. Ro, Critical binocular asymmetry measure for the perceptual quality assessment of synthesized stereo 3d images in view synthesis, IEEE Transactions on Circuits and Systems for Video Technology 26 (7) (2016) 1201–1214.

[24] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, M. Tanimoto, View generation with 3d warping using depth information for ftv, Signal Processing: Image Communication 24 (1) (2009) 65–72.

[25] K. Mueller, A. Smolic, K. Dix, P. Merkle, P. Kauff, T. Wiegand, View synthesis for advanced 3d video systems, EURASIP Journal on Image and Video Processing 2008 (1) (2009) 1–11.

[26] P. Ndjiki-Nya, M. Köppel, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, T. Wiegand, Depth image based rendering with advanced texture synthesis, in: 2010 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2010, pp. 424–429.

[27] P. Ndjiki-Nya, M. Koppel, D. Doshkov, H. Lakshman, P. Merkle, K. Muller, T. Wiegand, Depth image-based rendering with advanced texture synthesis for 3-d video, IEEE Transactions on Multimedia 13 (3) (2011) 453–465.

[28] M. Köppel, P. Ndjiki-Nya, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, 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.

[29] M. Solh, G. AlRegib, Hierarchical hole-filling for depth-based view synthesis in ftv and 3d video, IEEE Journal of Selected Topics in Signal Processing 6 (5) (2012) 495–504.

[30] Y. Wang, Y. Shuai, Y. Zhu, J. Zhang, P. An, Jointly learning perceptually heterogeneous features for blind 3d video quality assessment, Neurocomputing 332 (2019) 298–304.

[31] J. Yang, Y. Zhu, C. Ma, W. Lu, Q. Meng, Stereoscopic video quality assessment based on 3d convolutional neural networks, Neurocomputing 309 (2018) 83–93.

[32] S. S. Yoon, H. Sohn, Y. J. Jung, Y. M. Ro, Inter-view consistent hole filling in view extrapolation for multi-view image generation, in: Image Processing (ICIP), 2014 IEEE International Conference on, IEEE, 2014, pp. 2883–2887.

[33] G. J. Sullivan, J. M. Boyce, Y. Chen, J.-R. Ohm, C. A. Segall, A. Vetro, Standardized extensions of high efficiency video coding (hevc), IEEE Journal of selected topics in Signal Processing 7 (6) (2013) 1001–1016.

[34] IVC-IRCCyN lab, IRCCyN/IVC DIBR image database, http://ivc.univ-nantes.fr/en/databases/DIBR_Images/, last accessed Aug. 30th 2017, [Online].

[35] E. Bosc, P. Le Callet, L. Morin, M. Pressigout, An edge-based structural distortion indicator for the quality assessment of 3d synthesized views, in: 2012 Picture Coding Symposium, IEEE, 2012, pp. 249–252.

[36] P.-H. Conze, P. Robert, L. Morin, Objective view synthesis quality assessment, in: IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics, 2012, pp. 82881M–82881M.

[37] F. Battisti, E. Bosc, M. Carli, P. Le Callet, S. Perugia, Objective image quality assessment of 3D synthesized views, Signal Processing: Image Communication 30 (2015) 78–88.

[38] D. Sandić-Stanković, F. Battisti, D. Kukolj, P. L. Callet, M. Carli, Free viewpoint video quality assessment based on morphological multiscale metrics, in: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), 2016, pp. 1–6. doi:10.1109/QoMEX.2016.7498949.

[39] D. Sandić-Stanković, D. Kukolj, P. Le Callet, DIBR synthesized image quality assessment based on morphological wavelets, in: 2015 Seventh International Workshop onQuality of Multimedia Experience (QoMEX), IEEE, 2015, pp. 1–6.

[40] D. Sandić-Stanković, D. Kukolj, P. Le Callet, Multi–scale synthesized view assessment based on morphological pyramids, Journal of Electrical Engineering 67 (1) (2016) 3–11.

[41] S. Ling, P. Le Callet, G. Cheung, Quality assessment for synthesized view based on variable-length context tree, in: Multimedia Signal Processing (MMSP), 2017 IEEE 19th International Workshop on, IEEE, 2017, pp. 1–6.

[42] S. Ling, P. Le Callet, Image quality assessment for free viewpoint video based on mid-level contours feature, in: 2017 IEEE International Conference on Multimedia and Expo (ICME), 2017, pp. 79–84. doi:10.1109/ICME.2017.8019431.

[43] S. Ling, 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.

[44] Y. Zhao, L. Yu, A perceptual metric for evaluating quality of synthesized sequences in 3dv system, in: Visual Communications and Image Processing 2010, International Society for Optics and Photonics, 2010, pp. 77440X–77440X.

[45] Y. Zhang, H. Zhang, M. Yu, S. Kwong, Y. Ho, Sparse representation-based video quality assessment for synthesized 3d videos, IEEE Transactions on Image Processing 29 (2020) 509–524.

[46] Y. Zhou, L. Li, K. Gu, Y. Fang, W. Lin, Quality assessment of 3d synthesized images via disoccluded region discovery, in: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 1012–1016.

[47] S. Tian, L. Zhang, L. Morin, O. Deforges, A full-reference image quality assessment metric for 3d synthesized views, in: Image Quality and System Performance Conference, at IS&T Electronic Imaging 2018, Society for Imaging Science and Technology, 2018.

[48] S. Tian, L. Zhang, L. Morin, O. Déforges, Sc-iqa: Shift compensation based image quality assessment for dibr-synthesized views, in: IEEE International Conference on Visual Communications and Image Processing, 2018.

[49] Y. Zhou, L. Li, S. Ling, P. Le Callet, Quality assessment for view synthesis using low-level and mid-level structural representation, Signal Processing: Image Communication 74 (2019) 309–321.

[50] S. Ling, J. Li, P. Le Callet, J. Wang, Perceptual representations of structural information in images: application to quality assessment of synthesized view in ftv scenario, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 1735–

1739.

[51] X. Wang, F. Shao, Q. Jiang, R. Fu, Y. Ho, Quality assessment of 3d synthesized images via measuring local feature similarity and global sharpness, IEEE Access 7 (2019) 10242–10253.

[52] D. Sandic-Stankovic, D. Kukolj, P. Le Callet, DIBR-synthesized image quality assessment based on morphological multi-scale approach, EURASIP Journal on Image and Video Processing 2017 (1) (2016) 4.

[53] V. Jakhetiya, K. Gu, W. Lin, Q. Li, S. P. Jaiswal, A prediction backed model for quality assessment of screen content and 3-d synthesized images, IEEE Transactions on Industrial Informatics 14 (2) (2017) 652–660.

[54] L. Li, X. Chen, Y. Zhou, J. Wu, G. Shi, Depth image quality assessment for view synthesis based on weighted edge similarity., in: CVPR Workshops, 2019, pp. 17–25.

[55] T.-H. Le, S.-W. Jung, C. S. Won, A new depth image quality metric using a pair of color and depth images, Multimedia Tools and Applications 76 (9) (2017) 11285–11303.

[56] M. S. Farid, M. Lucenteforte, M. Grangetto, Blind depth quality assessment using histogram shape analysis, in: 2015 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), IEEE, 2015, pp. 1–5.

[57] S. Xiang, L. Yu, C. W. Chen, No-reference depth assessment based on edge misalignment errors for t+ d images, IEEE Transactions on Image Processing 25 (3) (2015) 1479–1494.

[58] L. Li, X. Chen, J. Wu, S. Wang, G. Shi, No-reference quality index of depth images based on statistics of edge profiles for view synthesis, Information Sciences 516 (2020) 205–219.

[59] M. Solh, G. AlRegib, J. M. Bauza, 3vqm: A vision-based quality measure for dibr-based 3d videos, in: 2011 IEEE International Conference on Multimedia and Expo, 2011, pp. 1–6. doi:10.1109/ICME.2011.6011992.

[60] L. Li, Y. Zhou, K. Gu, W. Lin, S. Wang, Quality assessment of dibr-synthesized images by measuring local geometric distortions and global sharpness, IEEE Transactions on Multimedia 20 (4) (2018) 914–926.

[61] M. S. Farid, M. Lucenteforte, M. Grangetto, Perceptual quality assessment of 3d synthesized images, in: Multimedia and Expo (ICME), 2017 IEEE International Conference on, IEEE, 2017, pp. 505–510.

[62] M. S. Farid, M. Lucenteforte, M. Grangetto, Objective quality metric for 3d virtual views, in: 2015 IEEE International Conference on Image Processing (ICIP), IEEE, 2015, pp. 3720–3724.

[63] M. S. Farid, M. Lucenteforte, M. Grangetto, Evaluating virtual image quality using the side-views information fusion and depth maps, Information Fusion 43 (2018) 47–56.

[64] K. Gu, V. Jakhetiya, J.-F. Qiao, X. Li, W. Lin, D. Thalmann, Model-based referenceless quality metric of 3d synthesized images using local image description, IEEE Transactions on Image Processing (2017).

[65] V. Jakhetiya, K. Gu, T. Singhal, S. C. Guntuku, Z. Xia, W. Lin, A highly efficient blind image quality assessment metric of 3d-synthesized images using outlier detection, IEEE Transactions on Industrial Informatics (2018).

[66] K. Gu, J. Qiao, S. Lee, H. Liu, W. Lin, P. Le Callet, Multiscale natural scene statistical analysis for no-reference quality evaluation of dibr-synthesized views, IEEE Transactions on Broadcasting (2019).

[67] D. D. Sandić-Stanković, D. D. Kukolj, P. Le Callet, Fast blind quality assessment of dibrsynthesized video based on high-high wavelet subband, IEEE Transactions on Image Processing (2019).

[68] S. Tian, L. Zhang, L. Morin, O. Déforges, NIQSV: A no reference image quality assessment metric for 3D synthesized views, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017.

[69] S. Tian, L. Zhang, L. Morin, O. Déforges, NIQSV+: A No-Reference Synthesized View Quality Assessment Metric, IEEE Transactions on Image Processing 27 (4) (2018) 1652–1664.

[70] F. Shao, Q. Yuan, W. Lin, G. Jiang, No-reference view synthesis quality prediction for 3-d videos based on color–depth interactions, IEEE Transactions on Multimedia 20 (3) (2017) 659–674.

[71] G. Yue, C. Hou, K. Gu, T. Zhou, G. Zhai, Combining local and global measures for dibr-synthesized image quality evaluation, IEEE Transactions on Image Processing 28 (4) (2018) 2075–2088.

[72] G. Wang, Z. Wang, K. Gu, Z. Xia, Blind quality assessment for 3d-synthesized images by measuring geometric distortions and image complexity, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 4040–4044.

[73] G. Wang, Z. Wang, K. Gu, L. Li, Z. Xia, L. Wu, Blind quality metric of dibr-synthesized images in the discrete wavelet transform domain., IEEE transactions on image processing: a publication of the IEEE Signal Processing Society (2019).

[74] Y. Zhou, L. Li, S. Wang, J. Wu, Y. Fang, X. Gao, No-reference quality assessment for view synthesis using dog-based edge statistics and texture naturalness, IEEE Transactions on Image Processing (2019).

[75] H. G. Kim, Y. M. Ro, Measurement of critical temporal inconsistency for quality assessment of synthesized video, in: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 1027–1031.

[76] Y. Zhou, L. Li, S. Wang, J. Wu, Y. Zhang, No-reference quality assessment of dibr-synthesized videos by measuring temporal flickering, Journal of Visual Communication and Image Representation 55 (2018) 30–39.

[77] S. Ling, 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.

[78] X. Wang, K. Wang, B. Yang, F. W. B. Li, X. Liang, Deep blind synthesized image quality assessment with contextual multi-level feature pooling, in: 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 435–439. doi:10.1109/ICIP.2019.8802943.

[79] S. Ling, J. Li, J. Wang, P. L. Callet, Gans-nqm: A generative adversarial networks based no reference quality assessment metric for RGB-D synthesized views, CoRR abs/1903.12088 (2019). arXiv:1903.12088.
URL http://arxiv.org/abs/1903.12088

[80] J. J. Lim, C. L. Zitnick, P. Dollár, Sketch tokens: A learned mid-level representation for contour and object detection, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3158–3165.

[81] P. Dollár, Z. Tu, P. Perona, S. Belongie, Integral channel features (2009).

[82] E. Shechtman, M. Irani, Matching local self-similarities across images and videos., in: CVPR, Vol. 2, Minneapolis, MN, 2007, p. 3.

[83] S. Ling, J. Gutiérrez, K. Gu, 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 9 (1) (2019) 204–216.

[84] W. Mio, A. Srivastava, S. Joshi, On shape of plane elastic curves, International Journal of Computer Vision 73 (3) (2007) 307–324.

[85] A. Srivastava, E. Klassen, S. H. Joshi, I. H. Jermyn, Shape analysis of elastic curves in euclidean spaces, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (7) (2010) 1415–1428.

[86] H. Freeman, Application of the generalized chain coding scheme to map data processing., Tech. rep., RENSSELAER POLYTECHNIC INST TROY NY DEPT OF ELECTRICAL AND SYSTEMS ENGINEERING (1978).

[87] A. Zheng, G. Cheung, D. Florencio, Context tree-based image contour coding using a geometric prior, IEEE Transactions on Image Processing 26 (2) (2016) 574–589.

[88] V. Jakhetiya, O. C. Au, S. Jaiswal, L. Jia, H. Zhang, Fast and efficient intra-frame deinterlacing using observation model based bilateral filter, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 5819–5823. doi:10.1109/ICASSP.2014.6854719.

[89] J. Wu, W. Lin, G. Shi, A. Liu, Reduced-reference image quality assessment with visual information fidelity, IEEE Transactions on Multimedia 15 (7) (2013) 1700–1705. doi:10.1109/TMM.2013.2266093.

[90] F. Battisti, 3DSwIM Source Code, http://www.comlab.uniroma3.it/3DSwIM.html, last accessed Aug. 30th 2017, [Online].

[91] H. W. Lilliefors, On the kolmogorov-smirnov test for normality with mean and variance unknown, Journal of the American statistical Association 62 (318) (1967) 399–402.

[92] P. Maragos, R. W. Schafer, Morphological systems for multidimensional signal processing, Proceedings of the IEEE 78 (4) (1990) 690–710.

[93] J. H. Westerink, K. Teunissen, Perceived sharpness in complex moving images, Displays 16 (2) (1995) 89–97.

[94] A. K. Moorthy, A. C. Bovik, Visual importance pooling for image quality assessment, IEEE journal of selected topics in signal processing 3 (2) (2009) 193–201.

[95] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, S. Li, Salient object detection: A discriminative regional feature integration approach, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 2083–2090.

[96] P. Kovesi, et al., Image features from phase congruency, Videre: Journal of computer vision research 1 (3) (1999) 1–26.

[97] B. Julesz, Cyclopean perception and neurophysiology, Investigative Ophthalmology & Visual Science 11 (6) (1972) 540–548.

[98] P. C. Teo, D. J. Heeger, Perceptual image distortion, in: Human Vision, Visual Processing, and Digital Display V, Vol. 2179, International Society for Optics and Photonics, 1994, pp. 127–141.

[99] B. K. Patra, R. Launonen, V. Ollikainen, S. Nandi, A new similarity measure using bhattacharyya coefficient for collaborative filtering in sparse data, Knowledge-Based Systems 82 (2015) 163–177.

[100] K. Gu, G. Zhai, W. Lin, X. Yang, W. Zhang, Visual saliency detection with free energy theory, IEEE Signal Processing Letters 22 (10) (2015) 1552–1555.

[101] L. Tang, L. Li, K. Gu, X. Sun, J. Zhang, Blind quality index for camera images with natural scene statistics and patch-based sharpness assessment, Journal of Visual Communication and Image Representation 40 (2016) 335–344.

[102] Y. Zhang, T. D. Phan, D. M. Chandler, Reduced-reference image quality assessment based on distortion families of local perceived sharpness, Signal Processing: Image Communication 55 (2017) 130–145.

[103] K. Gu, G. Zhai, W. Lin, X. Yang, W. Zhang, No-reference image sharpness assessment in autoregressive parameter space, IEEE Transactions on Image Processing 24 (10) (2015) 3218–3231.

[104] K. Gu, G. Zhai, W. Lin, X. Yang, W. Zhang, No-reference image sharpness assessment in autoregressive parameter space, IEEE Transactions on Image Processing 24 (10) (2015) 3218–3231. doi:10.1109/TIP.2015.2439035.

[105] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, C. W. Chen, Blind quality assessment based on pseudo-reference image, IEEE Transactions on Multimedia 20 (8) (2018) 2049–2062.

[106] R. Ferzli, L. J. Karam, A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb), IEEE Transactions on Image Processing 18 (4) (2009) 717–728.

[107] A. Cohen, I. Daubechies, J.-C. Feauveau, Biorthogonal bases of compactly supported wavelets, Communications on pure and applied mathematics 45 (5) (1992) 485–560.

[108] K. Gu, J. Zhou, J.-F. Qiao, G. Zhai, W. Lin, A. C. Bovik, No-reference quality assessment of screen content pictures, IEEE Transactions on Image Processing 26 (8) (2017) 4005–4018.

[109] K. Gu, W. Lin, G. Zhai, X. Yang, W. Zhang, C. W. Chen, No-reference quality metric of contrast-distorted images based on information maximization, IEEE transactions on cybernetics 47 (12) (2016) 4559–4565.

[110] M. A. Saad, A. C. Bovik, C. Charrier, Blind image quality assessment: A natural scene statistics approach in the dct domain, IEEE Transactions on Image Processing 21 (8) (2012) 3339–3352.

[111] M. A. Saad, A. C. Bovik, C. Charrier, DCT statistics model-based blind image quality assessment, in: 2011 18th IEEE International Conference on Image Processing (ICIP),, IEEE, 2011, pp. 3093–3096.

[112] A. K. Moorthy, A. C. Bovik, Blind image quality assessment: From natural scene statistics to perceptual quality, IEEE transactions on Image Processing 20 (12) (2011) 3350–3364.

[113] L. Liu, H. Dong, H. Huang, A. C. Bovik, No-reference image quality assessment in curvelet domain, Signal Processing: Image Communication 29 (4) (2014) 494 – 505. doi:https://doi.org/10.1016/j.image.2014.02.004.

[114] C. Li, Y. Zhang, X. Wu, W. Fang, L. Mao, Blind multiply distorted image quality assessment using relevant perceptual features, in: 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 4883–4886. doi:10.1109/ICIP.2015.7351735.

[115] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vol. 2, 2001, pp. 416–423 vol.2.

[116] X. Yang, F. Li, H. Liu, A survey of dnn methods for blind image quality assessment, IEEE Access 7 (2019) 123788–123806.

[117] X. Ye, X. Ji, B. Sun, S. Chen, Z. Wang, H. Li, Drm-slam: Towards dense reconstruction of monocular slam with scene depth fusion, Neurocomputing (2020).

[118] Y. Lei, W. Du, Q. Hu, Face sketch-to-photo transformation with multi-scale self-attention gan, Neurocomputing (2020).

[119] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, A. C. Bovik, Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment, IEEE Signal Processing Magazine 34 (6) (2017) 130–141.

[120] N. Zhuang, Q. Zhang, C. Pan, B. Ni, Y. Xu, X. Yang, W. Zhang, Recognition oriented facial image quality assessment via deep convolutional neural network, Neurocomputing 358 (2019) 109–118.

[121] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV) 115 (3) (2015) 211–252. doi:10.1007/s11263-015-0816-y.

[122] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

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

[124] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, International journal of computer vision 111 (1) (2015) 98–136.

[125] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (6) (2018) 1452–1464. doi:10.1109/TPAMI.2017.2723009.

[126] M. Manassi, B. Sayim, M. H. Herzog, When crowding of crowding leads to uncrowding, Journal of Vision 13 (13) (2013) 10–10.

[127] A. Moorthy, A. Bovik, A modular framework for constructing blind universal quality indices, IEEE Signal Processing Letters 17 (2009).

[128] V. Q. E. Group, Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, VQEG (March 2008).

[129] L. Krasula, K. Fliegel, P. Le Callet, M. Klíma, On the accuracy of objective image and video quality models: New methodology for performance evaluation, in: Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on, IEEE, 2016, pp. 1–6.

[130] S. Tian, Image quality assessment of 3d synthesized views, Ph.D.

thesis, Rennes, INSA (2019).

[131] K. Seshadrinathan, A. C. Bovik, Motion tuned spatio-temporal quality assessment of natural videos, IEEE transactions on image processing 19 (2) (2009) 335–350.

[132] R. Soundararajan, A. C. Bovik, Video quality assessment by reduced reference spatio-temporal entropic differencing, IEEE Transactions on Circuits and Systems for Video Technology 23 (4) (2012) 684–694.

[133] C. G. Bampis, P. Gupta, R. Soundararajan, A. C. Bovik, Speed-qa: Spatial efficient entropic differencing for image and video quality, IEEE signal processing letters 24 (9) (2017) 1333–1337.

[134] A. Mittal, M. A. Saad, A. C. Bovik, A completely blind video integrity oracle, IEEE Transactions on Image Processing 25 (1) (2015) 289–300.

[135] C. Sun, X. Liu, W. Yang, An efficient quality metric for dibr-based 3d video, in: 2012 IEEE 14th International Conference on High Performance Computing and Communication & 2012 IEEE 9th International Conference on Embedded Software and Systems, IEEE, 2012, pp. 1391–1394.

[136] L. Fei-Fei, R. Fergus, P. Perona, One-shot learning of object categories, IEEE transactions on pattern analysis and machine intelligence 28 (4) (2006) 594–611.

[137] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: Advances in Neural Information Processing Systems, 2017, pp. 4077–4087.

[138] X. Liu, J. van de Weijer, A. D. Bagdanov, Rankiqa: Learning from rankings for no-reference image quality assessment, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1040–1049.