Quality Assessment of DIBR-synthesized views: An Overview

Shishun Tian, Member, IEEE, Lu Zhang, Wenbin Zou, Xia Li, Member, IEEE, Ting Su, Luce Morin, and Olivier Déforges

Abstract—The Depth-Image-Based-Rendering (DIBR) is one of the main fundamental techniques to generate new views in 3D video applications, such as Multi-View Video (MVV), Free-Viewpoint Video (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 detailed survey in 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 by a summary and 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.

Index Terms—DIBR, Multi-view videos (MVV), view synthesis, distortions, quality assessment.

I. 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 receiver by using the Multiview-Video-Plus-Depth (MVD) data format and Depth-Image-Based-Rendering (DIBR) techniques [1]. 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 content. 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 named 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 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. 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 [2]–[9] 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.

Fig. 1. Procedure of DIBR

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 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. (3) the performances of these metrics are evaluated on different datasets, and the essential reasons of their performance on different type of distortions are analysed. (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. The subjective methods are first surveyed in Section II. Section III introduces the state-of-the-art objective quality metrics in detail. The experimental results are presented and discussed in Section IV. Finally, the conclusions are given in Section V.

II. 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 also different.

The Absolute Category Rating (ACR) method used in IVC image / video datasets [10], [11] randomly present the test sequences to the observers and ask them to rate on five-scales 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 [12] dataset allows the observer to rate on a continuous scale instead of a discrete five-scales evaluation. The IVY [13] 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 too much time since all the image pairs need to be tested. Compared to PC method, 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 less time than PC since it allows the observer to freely view several image multiple times and a continuous rating scale. Besides, the IVY [13], IETR [14] and SIAT datasets normalize the obtained scores to z-score to make the results more intuitive. The IVC and MCL-3D [15] datasets directly use the average scores. Except for the subjective test methodology, as shown in Table I, they use different sequences, DIBR algorithms etc. In the following part, we will introduce them respectively in detail.

A. IVC DIBR datasets

The IVC DIBR-image dataset [10] was proposed by Bosc et al. in 2011. It contains 84 DIBR-synthesized view images synthesized by 7 DIBR algorithms [1], [16]–[21]. 3 Multi-view plus Depth (MVD) sequences, BookArrival, Lovebird1 and Newspaper, are extracted as the source contents. MVD sequences contain several view-points acquired with real cameras in rectified configuration capturing the same scene, together with a depth image for each acquired view. 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 [11] 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 in the state-of-the-art view synthesis algorithms.

B. IETR DIBR image dataset

Similar to the IVC datasets, the IETR dataset [14] 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 [17], [22]–[27], including both interview synthesis and single view based synthesis, to exclude some “old fashioned” distortions, such as “black holes”. The interview DIBR algorithms use two neighbouring views to synthesize the virtual viewpoint. 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 the images captured by real cameras at the virtual viewpoints.

C. IVY stereoscopic image dataset

Jung et al. proposed the IVY stereoscopic 3D image dataset for the quality assessment of DIBR-synthesized stereoscopic images [13]. Different from the above two datasets, besides the DIBR view synthesis distortion, the IVY dataset also explores binocular perception 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 [22], [24], [28], [29] in this dataset. All the virtual view images in the IVY dataset are generated by single-view-based synthesis.
### Table I

**Summary of existing DIBR related datasets**

| Name          | Sequence          | Resolution       | Methodology   | DIBR algos                      | Other distortions     | No. PVS | Reference | Display |
|---------------|-------------------|------------------|---------------|---------------------------------|-----------------------|---------|-----------|---------|
| **IVC DIBR-image** |                   |                  |               |                                 |                       |         |           |         |
| BookArrival   | 1024 x 768        |                  | ACR<sup>2</sup> | Fehn's 2004                     | None                  | 84      | Original  | 2D      |
| Lovebird1     | 1024 x 768        |                  | ACR<sup>2</sup> | Telea's 2003                    | VSRS 2009             |         |           |         |
| Newspaper     | 1024 x 768        |                  | PC<sup>3</sup>  | Mller 2008                      | Nduki-Nya 2010        |         |           |         |
|               |                   |                  |               |                                 | Koppel 2010           |         |           |         |
| **IVC DIBR-video** |                 |                  |               |                                 |                       |         |           |         |
| BookArrival   | idem              |                  | ACR<sup>2</sup> | Criminisi 2004                  | VSRS 2009             | 93      | Original  | 2D      |
| Lovebird1     | idem              |                  | SAMVIQ<sup>4</sup> | None                            | Ahn's 2013            | 140     |           |         |
| Newspaper     | idem              |                  |                |                                 | VSRS 2009             |         |           |         |
| Balloons      | 1024 x 768        |                  |                |                                 | Ahn's 2013            |         |           |         |
| Kendo         | 1024 x 768        |                  |                |                                 | LDI 2011              |         |           |         |
| Dancer        | 1920 x 1088       |                  |                |                                 | HFF 2012              |         |           |         |
| Shark         | 1920 x 1088       |                  |                |                                 | Zhu's 2016            |         |           |         |
| Poznan Street | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| PoznanHall2   | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| GT_Fly        | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| **IETR-image** |                   |                  | SAMVIQ<sup>4</sup> | None                            | Ahn's 2013            | 140     |           |         |
| Kendo         | 1024 x 768        |                  |                |                                 | VSRS 2009             |         |           |         |
| Lovebird1     | 1024 x 768        |                  |                |                                 | Ahn's 2013            |         |           |         |
| Balloons      | 1024 x 768        |                  |                |                                 | LDI 2011              |         |           |         |
| Kendo         | 1024 x 768        |                  |                |                                 | HFF 2012              |         |           |         |
| Dancer        | 1920 x 1088       |                  |                |                                 | Zhu's 2016            |         |           |         |
| Shark         | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| Poznan Street | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| PoznanHall2   | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| GT_Fly        | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| **IVY-image** |                   |                  | DSCQS<sup>5</sup> | None                            | Ahn's 2013            | 84      | Original  | Stereo. |
| Aloe          | 1280 x 1100       |                  |                |                                 | VSRS 2009             |         |           |         |
| Dolls         | 1300 x 1100       |                  |                |                                 | Ahn's 2013            |         |           |         |
| Reindeer      | 1300 x 1100       |                  |                |                                 | Yoon 2014             |         |           |         |
| Laundry       | 1300 x 1100       |                  |                |                                 |                       |         |           |         |
| Lovebird1     | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| Newspaper     | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| BookArrival   | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| **MCL-3D image** |                 |                  | PC<sup>3</sup>  | Fehn's 2004 Additive White Noise |                       | 684     | Synthesized | Stereo. |
| Kendo         | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| Lovebird1     | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| Balloons      | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| Dancer        | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| Shark         | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| Poznan Street | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| PoznanHall2   | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| GT_Fly        | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| Microworld    | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| BookArrival   | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| Balloons      | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| Kendo         | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| Lovebird1     | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| Newspaper     | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| Dancer        | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| Shark         | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| Poznan Hall2  | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| Poznan Street | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| GT_Fly        | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| **SIAT video** |                   |                  | SSCQE<sup>6</sup> | VSRS 2009                       | 3DV-ATM (QP: 24-50)   | 140     | Original  | 2D      |
| Balloons      | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| Kendo         | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| Lovebird1     | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| Newspaper     | 1024 x 768        |                  |                |                                 |                       |         |           |         |
| Dancer        | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| Shark         | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| Poznan Hall2  | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| Poznan Street | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| GT_Fly        | 1920 x 1088       |                  |                |                                 |                       |         |           |         |
| Shark         | 1920 x 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.

### D. MCL-3D image dataset

Song et al. proposed the MCL-3D stereoscopic image dataset<sup>15</sup> 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)<sup>17</sup>. 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.

### E. SIAT synthesized video dataset

The SIAT synthesized video dataset<sup>12</sup> 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. 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 [10], [11], [31], IVY [13] and IETR [14] datasets focus on the distortions caused by different DIBR view synthesis algorithms; while the MCL-3D [15] and SIAT [12] 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.

III. 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. 2. 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 II classify 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 synthesized image, and evaluates the quality of the synthesized view based on the reconstruction error. These metrics are introduced as follows.

A. FR and RR metrics

In this subsection, we review 19 well-known FR metrics and 3 RR metrics.

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 [32] 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.

In [39], 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 [32], ST-SIAQ uses level contour descriptor called “Sketch Token” [71]. 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, color channels, and self-similarity channels) are 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 [74], 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 [40]. Different from ST-SIAQ metric, EM-IQA uses interest points matching and elastic metric instead of block matching and “Sketch Token” descriptor, to compensate the shifting and evaluate the contour degradation respectively. After 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 the EM-IQA metric, the boundaries of the segmented objects are considered as contours. Then, the elastic metric proposed in [75], [76] 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 [38], 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) [77] which represents the direction of object contours. Then, an optimal context tree is learned from the DCC in the reference image. The overall structural
Fig. 2. Categories of quality assessment metrics for DIBR-synthesized views.

TABLE II
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), SHARNESS EVALUATION (SE), SHIFT COMPENSATION (SC), IMAGE COMPLEXITY (IC), ML (MACHINE LEARNING).
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 [12] 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:

\[
\nabla_x^{temporal} = I(x, y, i) - I(x', y', i - 1),
\]

where \((x', y')\) is the coordinate in frame \(i - 1\) corresponding to \((x, y)\) along the motion trajectory in previous frame \(i\). The final quality score of DIBR-synthesized view video is obtained by integrating both “Activity” and “Flickering” distortions.

Furthermore, in [42], 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 number 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 [52] 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 [12].

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 [51]. 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) [79] is firstly used to divide the predicted and disorder parts.

\[
\hat{X_{d_i}} = \frac{X_{d_i} \lambda + \sum_{k \in N_i} \omega_k I_k}{\lambda + \sum_{k \in N_i} \omega_k}, \quad (2)
\]

where \(\hat{X_{d_i}}\) represents the predicted part, \(I_k\) and \(\omega_k\) are respectively the pixels and their associated weights in the surrounding \(3 \times 3\) window \(N_i\) of the \(i\)th pixel, \(\lambda\) is a parameter. The disorder part is computed as the difference between the predicted part and the original image:

\[
R_{d_i} = |\hat{X_{d_i}} - X_{d_i}| \quad (3)
\]

Then, the sharpness (edge structures) is calculated by four filters in [80]. Finally, the disorder and sharpness similarity between the distorted and reference images are estimated by using the similarity function in SSIM [2].

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 [34], [61]. The same as EM-IQA [40] and VQA-SIAT [12], 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 [82] distance between the histograms of the two matched blocks. 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 [35], [36]. 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.

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.

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 [83]. In [37], 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:

\[
D : dilation_{SE}(f)(x) = \max_{y \in SE} f(x - y) \quad (4)
\]

\[
E : erosion_{SE}(f)(x) = \min_{y \in SE} f(x + y) \quad (5)
\]

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 3 during the decomposition, the dilation is used as expand operation and the erosion is used as reduce 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 PSNR.

5) 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 [44], [45] instead of using weighting maps.

The SDRD metric proposed by Zhou in [44] 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 [45]. 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 [49], 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.

6) 2D related FR metrics: the main reason of the ineffectiveness of 2D quality assessment metrics on DIBR-synthesized views can be analysed as follows. Firstly, there exists large object shift in the synthesized views and this kind shift 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 in the dis-occluded regions. The 2D related metrics are based on the traditional 2D FR metrics, such as 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 [33] tries to improve the performance of SSIM [2] 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 textures, 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 [41]. The temporal noise is defined as the the difference between inter-frame change in the processed sequence and that in the reference sequence:

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

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 [84] and a motion mask [85], 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, but according to the recent research [86], Human Visual System (HVS) is more sensitive to local artefacts compared to the global object shift. In [46], 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 [87] 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 [12]. 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 [13], 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.

B. 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. 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 reference to evaluate the quality of DIBR-synthesized virtual views. These metrics are named as side view based FR metrics in this paper.

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) [52]. 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 $I_{wref}$ 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 $\sigma^2$. 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 [53]. 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:

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

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 logarithmic Gabor wavelet method proposed in [88]. 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:

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

(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 \cite{89} and divisive normalization (DN) transform \cite{90} based Synthesized Image Quality Evaluator (SIQE) in \cite{54}. 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 \cite{91}) 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 \cite{55}, 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.

C. 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.

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 \cite{51}, 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 ($3 \times 3$) is first used to construct a relationship between this pixel and its neighbouring pixels.

\[
x_i = \Omega(x_i)s + d_i
\]

where $\Omega(x_i)$ denotes a vector which is composed of the neighbouring pixels of $x_i$ in the ($3 \times 3$) 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 $7 \times 7$ 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 \cite{92} 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 \cite{57} 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.

2) Morphological operation based NR metrics: the morphological operations show their effectiveness in the FR metric MP-PSNR \cite{37}. In \cite{57}, 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.

3) Sharpness detection based NR metric: sharpness detection has been widely used in 2D image quality assessment \cite{93}–\cite{95} and also in the side view based FR metric LOGS \cite{52}. In this part, we will introduce its usage in NR metrics. Sharpness is one of the most important measurements in NR image quality assessment \cite{96}–\cite{98}. 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 \cite{56}, \cite{57}. 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 \cite{59}, 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 [36], [37], the NR_MWT also achieved a very low computational complexity.

Differently, in CLGM [63], 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+ [61], the stretching distortion is estimated by computing the similarity between the stretching region and its adjacent non-stretching region.

In [64], 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 edge similarity between the wavelet low-level and high-level sub-bands [99]. 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 [100]–[102], 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 [65], this metric is extended to achieve higher performance by adding a log-energy based sharpness detection module.

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 [103], 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 [67], 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.

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 [104]–[107]. 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, 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 [66]. The orientation selective statistics (similar to the metric in [106]) are extracted from different scale DoG images while the texture naturalness features are obtained based on the Gray level Gradient Co-occurrence Matrix (GGCM) [103] 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 [58]. The multiscale analysis on the DIBR-synthesized image and its associated reference image indicates that the distance (SSIM value [2]) 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 images can be approximately used as reference. 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 [109] 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 [62]. 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) [105], 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 [68], 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 still exists a large gap between objective quality measurement and human perception [110].

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. They make it possible to directly learn the representative features from image [111]. Unfortunately, owing to the limitation of size of DIBR-synthesized view datasets, there is not enough data to train the deep model straightforwardly. However, it is shown in the recent published literature that the deep neural network models trained on large-scale datasets, eg. ImageNet [112], can be used to extract effective representative features of human perception.

In [69], Wang et al. proposed a NR metric SIQA-CFP which uses the ResNet-50 [113] 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 I, various distortions may be introduced during the dis-occlusion region filling stage. Meanwhile, in current literature, several Generative Adversarial Networks (GAN) [114] based models have been proposed for image in-painting. As the generator is trained to inpaint 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) for DIBR-synthesized images. In GANs-NRM, a generative adversarial network for image in-painting is firstly trained on two large-scale datasets (PASCAL [115] and Places [116]). 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, eg. object detection, this metric is more targeted, and it also proposes a new way to obtain the semantic features for image quality assessment.

D. 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 [11], most of the metrics consist of multiple parts, it is thus difficult to classify them into a single specific category thoroughly, 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 used in these metrics, they can be classified into low-level [12], mid-level [39], [40] and high-level [68]–[70] metrics. As introduced in [117], 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 eg. objects, unnatural structures. Besides, there are also some hierarchical metrics which combine the above features, such as the LMS metric proposed in [47] which uses both low-level and mid-level features [39] and the metric in [48] which integrates the features on each level.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

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

A. 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) [118] recommended correlation based methods and the recently proposed Krasula’ model [119] in detail.

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:

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

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

\[
SROCC(X, Y) = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}
\]

where \(d_i\) indicates the difference of ranking of \(X\) and \(Y\). Higher PLCC and SROCC value 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 \(D/MOS_{SP}\) to remove the nonlinearities due to the subjective rating processing and to facilitate comparison of
the metrics in a common analysis space \cite{113}. The nonlinear function for regression mapping is shown as follows:

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

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 \(DMOS_S\) and \(DMOS\). As shown in Fig. 5, the nonlinearity has been removed after the regression.

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 \cite{119} 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. 5, firstly, the tested image pairs in the dataset are divided into two groups: different and similar according to their classification. The performances of objective metrics are evaluated by calculating the probability of image pairs. Then, we consider the pairs with higher than the selected significance level 0.95 to be significantly different. The others will be recognized as similar.

There are two performance analysis. The first performance

![Fig. 4. Example relationship between DMOS and objective quality scores.](image1)

![Fig. 5. Krasula’s model for performance evaluation of objective quality metrics.](image2)
analysis is conducted by evaluating how well the objective metric succeeds to distinguish significantly different image pairs from unsignificantly 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.

B. Performance on DIBR image datasets

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 III in which four 2D metrics [2],[104],[121] 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. 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, 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.) perform 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 FR metric has access to the ground truth. While on the IETR dataset, the NR metric perform worse than the FR metrics. The main reason is probably also be 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 the A1 view synthesis algorithm [16] 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 IV.

We can observe that the performance of FR and RR metrics increase significantly when large global shift artefacts are excluded.

| Metric                        | PLCC  | RMSE  | SROCC |
|-------------------------------|-------|-------|--------|
| FR 2D image metrics          | 0.7519| 0.4525| 0.6766 |
| FR DIBR image metrics        | 0.5956| 0.5513| 0.4424 |
| RR DIBR image metrics        | 0.8545| 0.3565| 0.7150 |

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 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 III, we can easily find that the performance of the DIBR-synthesized view dedicated metrics decrease greatly in IETR dataset compared to their performance in 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 the distortions in the IVC dataset, but 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 performance 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. For instance, VSQA and 3DSwIM metrics can not achieve the performance of SSIM; the SCDM, MP-PSNR and MW-PSNR metrics performs 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.

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
In the first different / similar analysis of the IVC dataset cf. (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 of the IVC dataset cf. (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. (c) (d) and the combination of the two datasets cf. (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.

C. 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...
TABLE V
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                        | 0.4081 | 0.6041 | 0.3751 | 0.4528 | 0.1144 | 0.4550 |
| MOVIE                       | 0.4971 | 0.4903 | 0.3877 | 0.646 | 0.097 | 0.693 |
| ST-RRED                     | 0.2025 | 0.6480 | 0.5777 | 0.7164 | 0.0895 | 0.6971 |
| **NR 2D video metrics**     |      |      |       |      |      |       |
| 5PeED                       | 0.3771 | 0.6128 | 0.5952 | 0.7236 | 0.0885 | 0.6987 |
| VIIDEO                      | 0.5971 | 0.5308 | 0.5877 | 0.2586 | 0.1239 | 0.2535 |
| **FR DIBR image metrics**   |      |      |       |      |      |       |
| Bosc                        | 0.5856 | 0.4602 | 0.2654 | 0.453 | 0.1144 | 0.431 |
| MF-PSNR                     | 0.5026 | 0.5720 | 0.5478 | 0.5681 | 0.1056 | 0.5044 |
| MW-PSNR                     | 0.4911 | 0.4638 | 0.4538 | 0.5745 | 0.1050 | 0.5024 |
| 3DSwIM                      | 0.4822 | 0.4974 | 0.3320 | 0.5677 | 0.1057 | 0.2626 |
| **NR DIBR image metrics**   |      |      |       |      |      |       |
| MP-PSNRr                    | 0.4617 | 0.5869 | 0.5307 | 0.5640 | 0.1059 | 0.5040 |
| MW-PSNRr                    | 0.4802 | 0.5804 | 0.5038 | 0.5757 | 0.1049 | 0.5853 |
| **SV-FR DIBR image metrics**|      |      |       |      |      |       |
| SIQE                        | 0.4084 | 0.5138 | 0.0991 | 0.3627 | 0.1195 | 0.2586 |
| DSSM                        | 0.5241 | 0.4857 | 0.3517 | 0.4001 | 0.1071 | 0.5994 |
| OUT                         | 0.6762 | 0.4874 | 0.6151 | 0.0945 | 0.1277 | 0.0926 |
| **NR DIBR image metrics**   |      |      |       |      |      |       |
| NRMWT                       | 0.7530 | 0.4354 | 0.7145 | 0.5051 | 0.1107 | 0.3092 |
| NJQSV                       | 0.6805 | 0.5025 | 0.5963 | 0.5544 | 0.1100 | 0.4562 |
| MNSS                        | 0.5180 | 0.5660 | 0.5371 | 0.1591 | 0.1266 | 0.2463 |
| **FR DIBR video metrics**   |      |      |       |      |      |       |
| CQM                         | 0.4102 | 0.5101 | 0.3265 | 0.4021 | 0.1070 | 0.4064 |
| PSPTNR                      | 0.4321 | 0.5002 | 0.4152 | 0.4461 | 0.1069 | 0.4305 |
| VQA-SIAT                    | 0.3947 | 0.5212 | 0.567 | 0.5527 | 0.0879 | 0.3852 |
| **NR DIBR video metrics**   |      |      |       |      |      |       |
| CTI                         | 0.6821 | 0.4372 | 0.6896 | 0.5736 | 0.1053 | 0.5425 |
| FDI                         | 0.7576 | 0.4319 | 0.7162 | 0.5952 | 0.1033 | 0.5425 |

(a) Different / Similar analysis on IVC Video dataset
(b) Better / Worse analysis on IVC Video dataset

Fig. 7. Performance on IVC video dataset using Krasula’s model. The metrics 1-13 represent: PSNR, SSIM, 5PeED, ST-RRED, VIIDEO, MF-PSNR, MW-PSNRr, NIQSV, OUT, MNSS, NRMWT, 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.

The obtained PLCC, RMSE and SROCC values on IVC video and SIAT video datasets are given in Table V. Only the results of Krasula’s model on IVC video dataset are shown in Fig. 7 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 effect on the synthesized views. We can easily observe that the best three metrics on IVC and SIAT datasets are the DIBR metrics and 2D metrics respectively except VQA-SIAT metric. 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 achieve high correlation with the subjective results. Moreover, there is no significant difference between the performance 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: the global shift effect.

D. Discussions

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

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 mis-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.

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 [48], [74]. 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 [13]. It could be interesting to try to design the metrics for these applications since they are currently rarely explored.

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 dataset. 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. Creating a very large-scale dataset may significantly help train a good deep model, but unlike the datasets for other tasks eg. 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. one-shot learning and few-shot learning [122], [123]. 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 [69], [70], more effort could be made to create a more general and effective end-to-end deep model.

V. 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.

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] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, April 2004.
[3] Z. Wang, E. P. Simoncelli, and 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. 1396–1402.
[4] Z. Wang and Q. Li, “Information content weighting for perceptual image quality assessment,” IEEE Transactions on Image Processing, vol. 20, no. 5, pp. 1185–1198, 2011.
[5] L. Ma, S. Li, F. Zhang, and K. N. Ngan, “Reduced-reference image quality assessment using reorganized dct-based image representation,” IEEE Transactions on Multimedia, vol. 13, no. 4, pp. 824–829, Aug 2011.
[6] Y. Liu, K. Gu, S. Wang, D. Zhao, and W. Gao, “Blur quality assessment of camera images based on low-level and high-level statistical features,” IEEE Transactions on Multimedia, vol. 21, no. 1, pp. 135–146, Jan 2019.
[7] Y. Liu, G. Zhai, K. Gu, X. Liu, D. Zhao, and W. Gao, “Reduced-reference image quality assessment in free-energy principle and sparse representation,” IEEE Transactions on Multimedia, vol. 20, no. 2, pp. 379–391, Feb 2018.
[8] Q. Li, W. Lin, J. Xu, and Y. Fang, “Blind image quality assessment using statistical structural and luminance features,” IEEE Transactions on Multimedia, vol. 18, no. 12, pp. 2457–2469, Dec 2016.
[9] K. Gu, G. Zhai, X. Yang, and W. Zhang, “Using free energy principle for blind image quality assessment,” IEEE Transactions on Multimedia, vol. 17, no. 1, pp. 50–63, Jan 2015.
[10] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Presigout, and L. Morin, “Towards a new quality metric for 3-d synthesized view assessment,” IEEE Journal of Selected Topics in Signal Processing, vol. 5, no. 7, pp. 1323–1343, 2011.
[11] E. Bosc, P. Le Callet, L. Morin, and 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.
[12] X. Liu, Y. Zhang, S. Hu, S. Kwong, C.-C. J. Kuo, and Q. Peng, “Subjective and objective video quality assessment of 3D synthesized views with texture/depth compression distortion,” IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 4847–4861, 2015.
[13] Y. J. Jung, H. G. Kim, and 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, vol. 26, no. 7, pp. 1201–1214, 2016.
[14] S. Tian, L. Zhang, L. Morin, and O. Duforges, “A benchmark of dibr synthesized view quality assessment metrics on a new database for immersive media applications,” IEEE Transactions on Multimedia, vol. 21, no. 5, pp. 1235–1247, May 2019.
G. J. Sullivan, J. M. Boyce, Y. Chen, J.-R. Ohm, C. A. Segall, and M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, “Ref-
I. Ahn and C. Kim, “A novel depth-based virtual view synthesis method for free viewpoint video,” IEEE Transactions on Broadcasting, vol. 59, no. 4, pp. 614–626, 2013.
G. Luo, Y. Zhu, Z. Li, and L. Zhang, “A hole filling approach based on background reconstruction for view synthesis in 3d videos,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1781–1789.
M. Soli and G. AlRegib, “Hierarchical hole-filling for depth-based virtual view synthesis in fiv and 3d video,” IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 5, pp. 495–504, 2012.
C. Zhu and S. Li, “Depth image based view synthesis: New insights and perspectives on hole generation and filling,” IEEE Transactions on Broadcasting, vol. 62, no. 1, pp. 82–93, 2016.
M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, “Reference softwares for depth estimation and view synthesis,” ISO/IEC JTC1/SC29/WG11 MPEG, vol. 20081, p. M15377, 2008.
S. S. Yoon, H. Sohn, Y. J. Jung, and 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.
G. J. Sullivan, J. M. Boyce, Y. Chen, J.-R. Ohm, C. A. Segall, and A. Vetro, “Standardized extensions of high efficiency video coding (hevc),” IEEE Journal of selected topics in Signal Processing, vol. 7, no. 6, pp. 1001–1016, 2013.
IVC-IRCCyN lab, “IRCCyN/IvC DIBR image database,” http://ivc.univ-nantes.fr/en/databases/DIBR_Images last accessed Aug. 30th, [Online].
E. Bosc, P. Le Callet, L. Morin, and 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.
P-H. Conze, P. Robert, and L. Morin, “Objective view synthesis quality assessment,” in IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2012, pp. 82881M–82881M.
F. Battisti, E. Bosc, M. Carli, P. Le Callet, and S. Perugia, “Objective image quality assessment of 3D synthesized views,” Signal Processing: Image Communication, vol. 30, pp. 78–88, 2015.
D. Sandić-Stanković, F. Battisti, D. Kuljak, P. L. Callet, and M. Carli, “Free viewpoint video quality assessment based on morphological multiscale metrics,” in 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), June 2016, pp. 1–6.
D. Sandić-Stanković, D. Kuljak, and 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.
D. Sandić-Stanković, D. Kuljak, and P. Le Callet, “Multi–scale synthesized view assessment based on morphological pyramids,” Journal of Electrical Engineering, vol. 67, no. 1, pp. 3–11, 2016.
[103] H. G. Kim and 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.

[104] M. A. Saad, A. C. Bovik, and C. Charrier, “Blind image quality assessment: A natural scene statistics approach in the dct domain,” *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.

[105] ———, “DCT statistics model-based blind image quality assessment,” in *2011 18th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2011, pp. 3093–3096.

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

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

[108] C. Li, Y. Zhang, X. Wu, W. Fang, and L. Mao, “Blind multiply distorted image quality assessment using relevant perceptual features,” in *2015 IEEE International Conference on Image Processing (ICIP)*, Sep. 2015, pp. 4883–4886.

[109] D. Martin, C. Fowlkes, D. Tal, and 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, July 2001, pp. 416–423 vol.2.

[110] X. Yang, F. Li, and H. Liu, “A survey of dnn methods for blind image quality assessment,” *IEEE Access*, vol. 7, pp. 123788–123806, 2019.

[111] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, “Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment,” *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 130–141, Nov 2017.

[112] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in *Advances in Neural Information Processing Systems 29*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2016, pp. 4077–4085. [Online]. Available: http://papers.nips.cc/paper/6131-prototypical-networks.pdf

[113] 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.

[114] L. Krasula, K. Fliegel, P. Le Callet, and 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.

[115] S. Tian, “Image quality assessment of 3d synthesized views,” Ph.D. dissertation, Rennes, INSA, 2019.

[116] M. Manassi, B. Sayim, and M. H. Herzog, “When crowding of crowding leads to uncrowding,” *Journal of Vision*, vol. 13, no. 13, pp. 10–10, 2013.

[117] 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.

[118] L. Krasula, K. Fliegel, P. Le Callet, and 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.

[119] S. Tian, “Image quality assessment of 3d synthesized views,” Ph.D. dissertation, Rennes, INSA, 2019.

[120] A. Moorthy and A. Bovik, “A modular framework for constructing blind universal quality indices,” *IEEE Signal Processing Letters*, vol. 17, 2009.

[121] L. Fei-Fei, R. Fergus, and P. Perona, “One-shot learning of object categories,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

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