


Article

Blind Quality Prediction for View Synthesis Based on Heterogeneous Distortion Perception

Haozhi Shi ¹, Lanmei Wang ^{1,*}  and Guibao Wang ²¹ School of Physics, Xidian University, Xi'an 710071, China² School of Physics and Telecommunication Engineering, Shaanxi University of Technology, Hanzhong 723001, China

* Correspondence: lmwang@mail.xidian.edu.cn

Abstract: The quality of synthesized images directly affects the practical application of virtual view synthesis technology, which typically uses a depth-image-based rendering (DIBR) algorithm to generate a new viewpoint based on texture and depth images. Current view synthesis quality metrics commonly evaluate the quality of DIBR-synthesized images, where the DIBR process is computationally expensive and time-consuming. In addition, the existing view synthesis quality metrics cannot achieve robustness due to the shallow hand-crafted features. To avoid the complicated DIBR process and learn more efficient features, this paper presents a blind quality prediction model for view synthesis based on Heterogeneous Distortion Perception, dubbed HEDIP, which predicts the image quality of view synthesis from texture and depth images. Specifically, the texture and depth images are first fused based on discrete cosine transform to simulate the distortion of view synthesis images, and then the spatial and gradient domain features are extracted in a Two-Channel Convolutional Neural Network (TCCNN). Finally, a fully connected layer maps the extracted features to a quality score. Notably, the ground-truth score of the source image cannot effectively represent the labels of each image patch during training due to the presence of local distortions in view synthesis image. So, we design a Heterogeneous Distortion Perception (HDP) module to provide effective training labels for each image patch. Experiments show that with the help of the HDP module, the proposed model can effectively predict the quality of view synthesis. Experimental results demonstrate the effectiveness of the proposed model.

Keywords: view synthesis; quality prediction; two-channel convolutional neural network; heterogeneous distortion perception



Citation: Shi, H.; Wang, L.; Wang, G. Blind Quality Prediction for View Synthesis Based on Heterogeneous Distortion Perception. *Sensors* **2022**, *22*, 7081. <https://doi.org/10.3390/s22187081>

Academic Editor: Christophoros Nikou

Received: 14 August 2022

Accepted: 8 September 2022

Published: 19 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the wide application of multi-view video and free-view television, virtual view synthesis technology has developed rapidly [1,2]. The virtual multi-view synthesis technology interacts with texture and depth images from different viewpoints to generate new viewpoints, of which the depth-image-based rendering (DIBR) algorithm is the most commonly used and recognized [3,4]. In practice, distortions may occur in the acquisition, compression, and transmission of texture and depth images, which affect the image quality of view synthesis [5]. As a result, it is necessary to give a corresponding quality evaluation to measure and optimize the effect of view synthesis [6].

Image quality assessment (IQA) is divided into full-reference (FR), reduced-reference (RR), and no-reference (NR) [7]. FR-IQA methods require reference to the original distortion-free image for scoring. Typical FR-IQA methods include Structural Similarity (SSIM) [8], Information Fidelity Criterion (IFC) [9], and Visual Information Fidelity (VIF) [10]. RR-IQA methods use only a small amount of edge information extracted from the original distortion-free image as a reference for scoring [11,12]. However, in practical applications, the original image of the distorted image rarely exists. Hence, it is more practical to use

the NR-IQA method, which does not require any information from the original distortion-free image to be referenced for scoring [13]. Traditional NR-IQA methods include the Blind Image Quality Index (BIQI) [14], the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [15], and the Natural Image Quality Evaluator (NIQE) [16]. With the development of deep learning in recent years, Kang et al. [17] first proposed an NR-IQA model using a Convolutional Neural Network (CNN) to extract image features. After that, some deep-learning-based NR-IQA models were also proposed [18,19]. Although the above IQA models have outstanding performance in the quality assessment of natural scene images, their application in view synthesis is very limited. This is because there are also local geometric distortions generated by the depth images and DIBR process in the view synthesis, which cannot be handled by the general IQA model [6].

In this case, view synthesis quality metrics have been proposed [20–22] to evaluate the image quality after DIBR-based view synthesis. However, the DIBR process is computationally expensive and time-consuming. For this reason, it is very meaningful and valuable to predict the image quality after view synthesis from pre-synthesis texture and depth images [23–25]. Nevertheless, current quality evaluation methods for view synthesis basically use hand-designed features. The performance improvement of traditional methods is relatively slow because of the shallow feature extraction of hand-designed methods [26]. In contrast, CNN has a strong expressive ability and is widely used in the field of quality evaluation of natural scene images [17]. Therefore, we consider applying deep learning in quality prediction for view synthesis.

This paper proposes a blind quality prediction model based on Heterogeneous Distortion Perception (HEDIP), which predicts the image quality of view synthesis from pre-synthesis texture and depth images. The distortions of texture and depth images usually lead to traditional and geometric distortions [25], i.e., heterogeneous distortions, in the DIBR-synthesized images. To obtain more edge information, the proposed model is designed as a Two-Channel Convolutional Neural Network (TCCNN) structure, which can extract features in the image spatial and gradient domain, respectively. Among them, the edge features extracted by the gradient channel can effectively reflect the geometric distortions. Furthermore, to better describe the geometric distortions, we add a Contextual Multi-Level Feature Fusion (CMLFF) module, which can fuse shallow detail features and deep semantic features. At the input of the proposed HEDIP model, the texture and depth images are fused by Discrete Cosine Transform (DCT) [27] to imitate the distortions of DIBR-synthesized images. The fused images are then fed to the TCCNN to extract features in the spatial and gradient domains. Additionally, a fully connected layer linearly regresses the extracted features into a quality score. Considering the presence of non-uniform distortions in the view synthesis image [25], the ground-truth score of the source image cannot effectively represent the labels of each image patch during training. Therefore, we design a Heterogeneous Distortion Perception (HDP) module with the help of the classic BRISQUE [15] metric and combine it with the ground-truth score of the source image to provide effective training labels for each image patch. The advantage of the proposed HEDIP model is demonstrated through extensive experiments and comparisons. The contributions of this paper are as follows.

1. We propose a deep-learning-based blind quality prediction model for view synthesis, a two-channel convolutional neural network structure based on the spatial-gradient domain, which operates end-to-end via input texture and depth images.
2. A heterogeneous distortion perception module is designed to provide effective training labels for each image patch.
3. Extensive experiments on different databases show that our proposed model achieves state of the art.

2. Related Work

Existing view synthesis quality metrics basically adopt hand-designed methods to extract features. Tian et al. [20] proposed a NIQSV metric by quantifying the distortions of

synthesized images based on morphological and edge operations. Furthermore, they also proposed NIQSV+ [21] metric on this basis to evaluate blurred regions, holes, and stretching distortions. Gu et al. [22] first generated reconstructed images using the autoregression (AR) model and then measured the geometric distortions based on the error between the AR-reconstructed image and the corresponding DIBR-synthesized image. In [28], a No-Reference Morphological Wavelet with Threshold (NR-MWT) metric first obtained high-frequency information based on morphological wavelet and then mapped the high-frequency information to the quality score. Gu et al. [29] reported a Multiscale Natural Scene Statistical analysis (MNSS) method, which inferred the image quality mainly based on the degree of self-similarity impairment and major structure degradation at different scales. Zhou et al. [30] addressed a blind view composite quality metric, which used Difference-of-Gaussian features to measure edge degradation and texture unnaturalness. Wang et al. [31] decomposed the DIBR-synthesized images by using discrete wavelet transform and then calculated the quality score of the synthesized image based on the geometric distortions and global sharpness of the low-frequency and high-frequency sub-bands. Recently, Li et al. [32] reported a view synthesis quality metric based on local Instance DEgradation and global Appearance (IDEA). This model used discrete orthogonal moments and superpixels to measure local and global distortions, respectively.

The above works are all about quality evaluation of the images after view synthesis. The DIBR-based view synthesis process includes the acquisition, compression, transmission, and decompression of texture and depth images, as well as deformation and rendering in the DIBR process. In practical applications, different types and degrees of distortions may occur in each link of view synthesis. Moreover, the DIBR process is computationally intensive and complex. To avoid unnecessary distortions and calculations, it is worth considering predicting the quality of view synthesis based on texture and depth images, which can make the view synthesis system more flexible. Currently, only a few studies have investigated quality prediction for view synthesis. Wang et al. [23] advised a novel FR quality prediction model, which utilized the classic SSIM [8] method to compute two quality indication maps between distorted images and reference images for texture and depth. The overall quality is calculated based on the two quality indication maps. Shao et al. [24] recommended a High-Efficiency View Synthesis Quality Prediction (HEVSQP) method with the help of sparse representation. They first achieved Color-Involved View Synthesis Quality Prediction (CI-VSQP) and Depth-Involved View Synthesis Quality Prediction (DI-VSQP), and then predicted the quality score of the synthesized view through the metrics of CI-VSQP and DI-VSQP models. Li et al. [25] put forward a prediction model based on color-depth image fusion, which fused the input texture and depth images through wavelet transform to imitate the synthesized images. The statistical features of the fused images are then mapped to quality scores.

3. Materials and Methods

The proposed HEDIP is a deep learning model that can predict the image quality of view synthesis without reference. The texture and depth images before synthesis are fused through DCT, and then the spatial and gradient domain features of the fused image are extracted to predict the quality score. Notably, for the problem that local distortion causes image patches to have no valid training labels, the designed HDP module can provide effective training labels for each image patch with the help of the classic BRISQUE metric and the ground-truth score of the source image.

3.1. Image Preprocessing

In DIBR-based view synthesis, the distortions of texture and depth images generally lead to traditional and geometric distortions in the synthesized images [31]. Therefore, we fuse texture and depth images to imitate the distortions of DIBR-synthesized images. It is worth emphasizing that DCT transform and inverse transform are real-time and lossless, so we fuse texture and depth images through DCT transform. Among the DCT

coefficients, the low-frequency coefficients mainly represent the information that changes gently in image intensity (brightness/grayscale), and the high-frequency coefficients mainly represent the detailed information of the image [33]. The low-frequency coefficients may contain noise information, and the high-frequency coefficients may contain geometric distortion information, both of which will degrade image quality [34]. As a result, we keep the low-frequency coefficients of the texture image and averagely fuse the high-frequency coefficients of the texture and depth image. Then the fused image is obtained by inverse DCT transform. The distortions of the texture image are directly transferred to the fused image, while the distortions of the depth image destroy the edge information of the fused image.

Because the Sobel operator is fast and accurate in edge positioning, we choose to use the Sobel operator to calculate the gradient image. The gradient image I_g of the spatial image I_d is calculated as follows:

$$I_g = \sqrt{G_x^2 + G_y^2}, \quad (1)$$

$$G_x = M * I_d, \quad (2)$$

$$G_y = M^T * I_d, \quad (3)$$

where $M = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix}$, T is the transpose operation, and $*$ is the convolution operation.

The fused image and the corresponding gradient image are shown in Figure 1. The gradient image can represent the edge information of the fused image well.

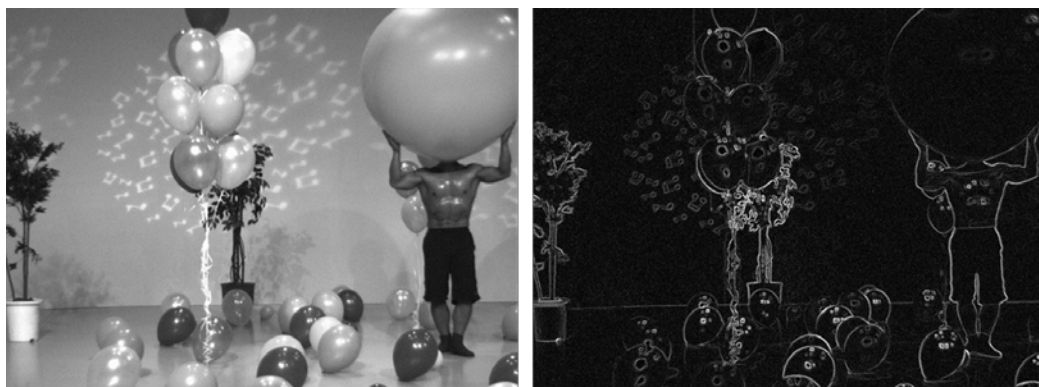


Figure 1. Fused image and corresponding gradient image.

3.2. Two-Channel Convolutional Neural Network Structure

To obtain more edge information, the proposed HEDIP model is designed as a Two-Channel Convolutional Neural Network structure, which can extract features in the image spatial and gradient domain, respectively. Among them, the edge features extracted by the gradient channel can effectively reflect the geometric distortions. The output of each layer in the proposed HEDIP model is shown in Table 1. To be specific, the network structure is shown in Figure 2, including Conv3 × 3, Residual block, Max pooling, Upsample block, Global average pooling, Add, Concatenate, and the Fully connected layer. Among them, the residual block can prevent gradient disappearance by reusing shallow features of the image. As shown in Figure 3a, the Residual block consists of Conv3 × 3, Conv1 × 1, and Conv3 × 3. Table 2 shows that the main function of Conv1 × 1 is to reduce the number of parameters. As shown in Figure 3b, the Upsample block is composed of Conv1 × 1 and Upsample. The function of Conv1 × 1 here is to change the number of channels, and the function of Upsample is to change the size of the deep features to match the shallow features. Notably, each convolutional layer is followed by a Rectified Linear Unit

(ReLU) [35] activation function $z = \max(0, \sum_i w_i a_i)$, where z , w_i , and a_i represent the output of the current layer and the weight and the output of the previous layer, respectively.

Table 1. Information of each layer of the two-channel convolutional neural network structure.

Layer Name	Output Size
Spatial/Gradient image patch	$128 \times 128 \times 1$
Conv3 \times 3/ReLU	$64 \times 64 \times 32 (F_{64 \times 64})$
Residual block	$32 \times 32 \times 32$
Max pooling	$16 \times 16 \times 32 (F_{16 \times 16})$
Conv3 \times 3/ReLU	$8 \times 8 \times 64$
Residual block	$4 \times 4 \times 64 (F_{4 \times 4})$
Upsample block	$16 \times 16 \times 32$
Concatenate	$16 \times 16 \times 64$
Upsample block	$64 \times 64 \times 32$
Concatenate	$64 \times 64 \times 64$
Global average pooling 1/2/3	$1 \times 1 \times 64$
Add	$1 \times 1 \times 64$
Concatenate	$1 \times 1 \times 128$
Fully connected layer (Score)	1

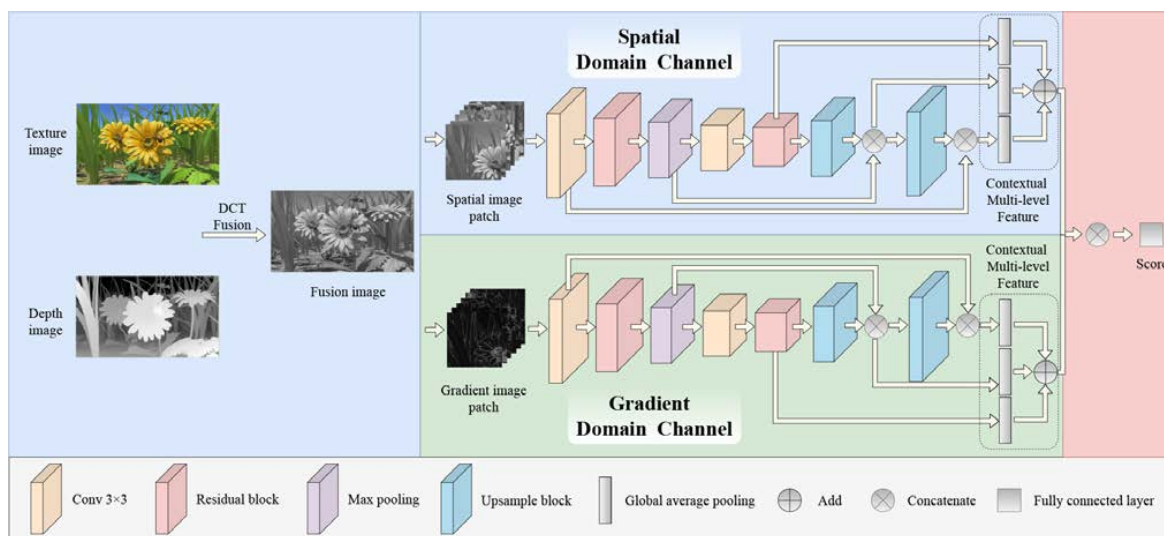


Figure 2. Two-channel convolutional neural network structure based on spatial-gradient domain.

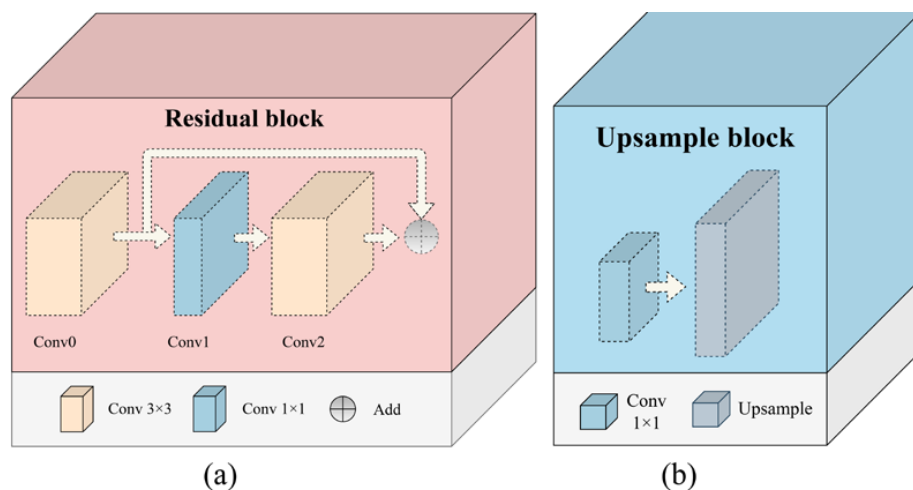


Figure 3. Residual block and Upsample block structure. (a) Residual block; (b) Upsample block.

Table 2. Information of each layer of the residual block.

Layer Name	Padding	Filter Size	Stride
Conv0/ReLU	1	3×3	2
Conv1/ReLU	0	1×1	1
Conv2/ReLU	1	3×3	1
Add	/	/	/

This paper denotes the spatial domain channel as $SDC(\cdot)$. The spatial domain feature is:

$$F_S = SDC(w_S, p_S), \quad (4)$$

where w_S and p_S denote the spatial domain channel weight and spatial image patch. Similar to the spatial domain channel, the gradient domain channel is denoted as $GDC(\cdot)$. The gradient domain feature is:

$$F_g = GDC(w_g, p_g), \quad (5)$$

where w_g and p_g represent the gradient domain channel weight and gradient image patch, respectively. Then, F_S and F_g are fused as:

$$F = \text{concat}(F_S, F_g), \quad (6)$$

where $\text{concat}(\cdot)$ represents the concatenating multiple features.

Finally, F is linearly regressed into the quality score by a fully connected layer.

3.3. Heterogeneous Distortion Perception Module

In DIBR-synthesized images, the overall distortion is different from the local distortion. From this point of view, the ground-truth score of the source image cannot be efficiently represented as the labels of each image patch during training.

To address this problem, we propose an HDP module, which is shown in Figure 4. The image patch and the corresponding source image are evaluated by the classic BRISQUE model to obtain scores a and b . Remarkably, unlike the ground-truth score of the source image, the evaluation standard of the BRISQUE model is that a large score corresponds to more serious distortion. If the quality of the image patch is lower relative to the quality of the source image, the score a of the image patch is larger than the score b of the source image. In this case, in order for the training label of the image patch to match the ground-truth score of the source image, i.e., the larger the score, the smaller the distortion, the HDP weight of the image patch is calculated as:

$$w = \frac{b}{a}, \quad (7)$$

where w represents the distortion of the image patch relative to the source image. When w is smaller, it indicates that the distortion of the image patch is more serious, and the corresponding score (training label) is smaller. Hence, the training label for the image patch can be computed as:

$$\hat{a} = \hat{b} * w, \quad (8)$$

where \hat{b} is the ground-truth score of the source image.

Figure 5a shows the visualization of local distortion. It can be seen from the figure that the HDP weight w of the image patch with severe distortion is smaller, and the corresponding color is darker. A visualization of the global distortion is shown in Figure 5b, where the distortion perception weight w is almost the same for each image patch. The HDP module can be easily observed to be suitable not only for images with local distortion, but also for images with global distortion.

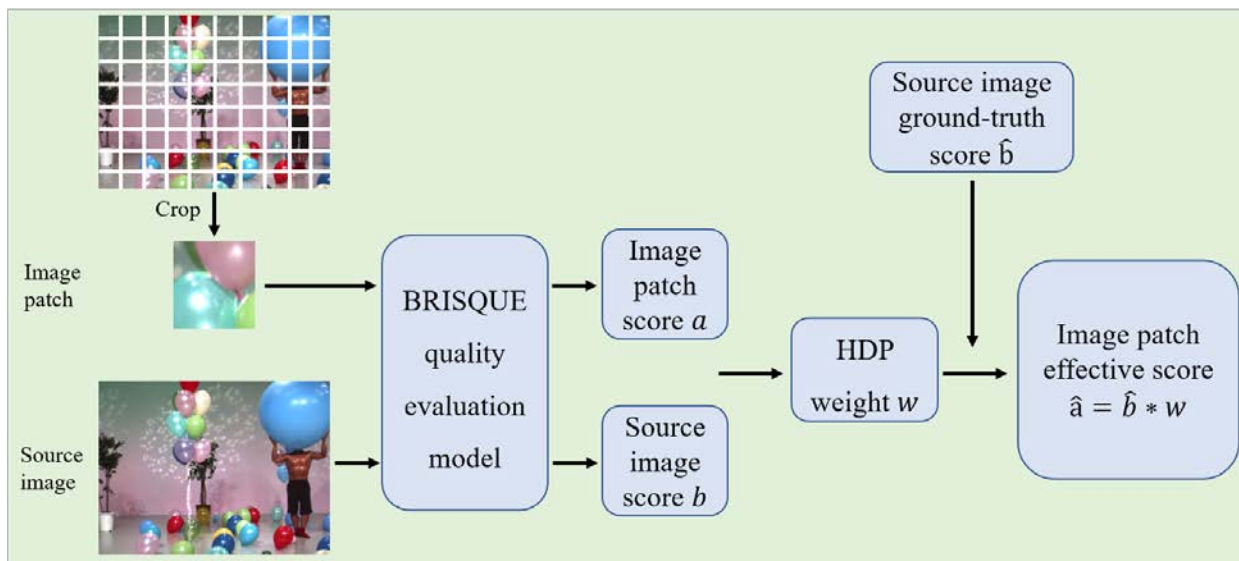


Figure 4. Heterogeneous distortion perception module.

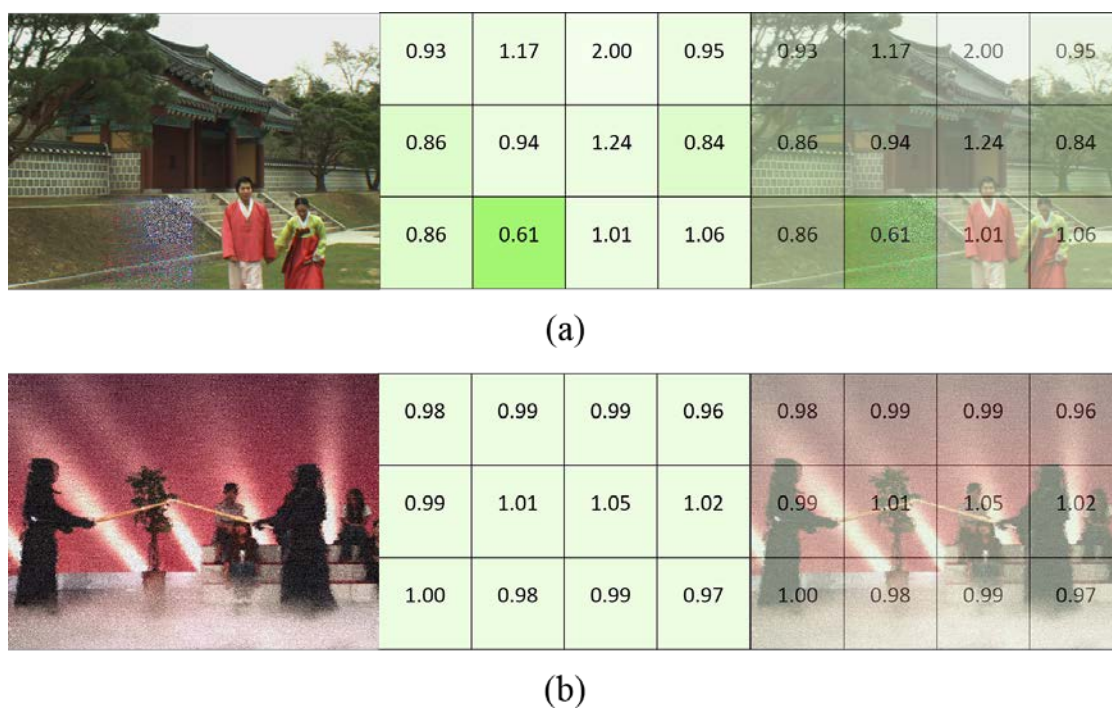


Figure 5. Visualization of HDP weights. The left image is a distorted image, the middle image is an HDP weight map, and the right image is an intuitive comparison of the distorted image and the HDP weight map: (a) case of local distortion; (b) case of global distortion.

3.4. Contextual Multi-Level Feature Fusion Module

To better describe the geometric distortions, we propose a contextual multi-level feature fusion module, which fuses shallow detail features and deep semantic features. Figure 2 shows the module, and the feature names required for operation are shown in Table 1. First, the feature $F_{4 \times 4}(4 \times 4 \times 64)$ is adjusted to the $F'_{4 \times 4}(16 \times 16 \times 32)$ through the Upsample block, and then we concatenate the $F'_{4 \times 4}$ and $F_{16 \times 16}(16 \times 16 \times 32)$ to obtain the $F'_{16 \times 16}(16 \times 16 \times 64)$. In addition, the $F'_{64 \times 64}(64 \times 64 \times 64)$ is obtained by operating the same steps as above for the $F'_{16 \times 16}$. Finally, the $GAP_{F_{64}}$, $GAP_{F_{16}}$, and GAP_{F_4} are obtained

by global average pooling the $F'_{64 \times 64}$, $F'_{16 \times 16}$, and $F_{4 \times 4}$, respectively. The weight of the i -th feature is recorded as:

$$p_i^* = \max(0, p_i) + \tau, \quad (9)$$

where τ is a stable constant, which can guarantee $p_i^* > 0$. Furthermore, the weights p_i^* are normalized to:

$$b_i = \frac{p_i^*}{\sum_j^{N_p} p_j^*}, \quad (10)$$

where N_p is equal to 3. Therefore, the feature F after fusion is calculated as:

$$F = p_1^* \text{GAP}_{F64} + p_2^* \text{GAP}_{F16} + p_3^* \text{GAP}_{F4}, \quad (11)$$

3.5. Training

We employ a window sliding strategy to divide the image into several 128×128 image patches to train our model. During the training phase, each image patch is provided with labels according to the designed HDP module. In the testing phase, the predicted score of the source image is obtained by averaging the predicted scores of all image patches in the source image. The mapping between extracted features and scores is achieved by minimizing the loss of predicted and ground-truth scores, so the loss function is designed as:

$$\min \frac{1}{N} \sum_{l=1}^N \|q_l - \hat{q}_l\|, \quad (12)$$

where N is the number of texture–depth image pairs in the training set, and q_l and \hat{q}_l denote the predicted score and training label of the l -th image patch, respectively. The proposed HEDIP model is implemented in Pytorch and runs on a Windows 10 system with a 3.70 GHz CPU and NVIDIA 2080 Ti GPU.

4. Experiments

4.1. Datasets and Evaluation Protocols

We conduct a series of experiments on the MCL-3D [36] and IST [37] databases to verify the performance of the proposed quality prediction metric for DIBR-based view synthesis. MCL-3D database [36]. The database consists of 684 synthesized image pairs and corresponding Mean Opinion Score (MOS) value. Among them, 648 image pairs are generated by the View Synthesis Reference Software (VSRS) [36] using the texture–depth image pairs. There are three combinations of texture and depth images for view synthesis: (1) distorted texture images and undistorted depth images, (2) undistorted texture images and distorted depth images, and (3) distorted texture images and distorted depth images. Six kinds of distortions are applied to the input color and/or depth images, namely, Gaussian blur, JPEG compression, downsampling blurring, additive white noise, JPEG2000, and transmission error. IST database [37]. The database consists of 180 synthesized image pairs and corresponding MOS values. Among them, 120 image pairs are synthesized by the VSIM [38] algorithm, and the remaining 60 image pairs are synthesized by the VSRS [36] algorithm. Moreover, both the texture and depth images suffer from compression artifacts to varying degrees. It is worth noting that the images are synthesized by the VSIM and VSRS algorithms, respectively, in the DIBR-based view synthesis process. Therefore, for this database, we conduct two sets of experiments, respectively, on the texture and depth images required in the synthesis process of the VSIM and VSRS algorithms.

The MOS values of the synthesized images in the above two databases can be used as the ground-truth scores of input texture–depth image pairs. Furthermore, we adopt the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank order Correlation Coefficient (SRCC) to evaluate model performance. PLCC is used to measure the performance of the model in terms of accuracy, and SRCC is used to measure the performance of the model in terms of monotonicity. The closer the PLCC and SRCC are to one, the better the model performance [24,39].

4.2. Performance Evaluation

We compare the proposed HEDIP model with state-of-the-art related models. Four general NR-IQA metrics are compared, namely, BRISQUE [15], NIQE [16], IL-NIQE [40], and M3 [41]. Quality evaluation metrics for view synthesis are compared, including MW-PSNR [42], MP-PSNR [43], LOGS [6], SET [30], Jakhetiya’s [44], and NIQSV [20]. In addition, the metric [23], which first proposed the idea of view synthesis quality prediction, is also compared. Depending on the scene, 80% of the image pairs are randomly selected for training, and the remaining 20% are used for testing. To avoid bias, the random split of the training test is repeated 10 times, and the average values are reported [45]. It should be noted that the metric [23] needs undistorted texture and depth images during the quality prediction, which are not provided in the IST dataset. Therefore, the PLCC and SRCC of the metric [23] on the IST database cannot be calculated.

The accuracy (PLCC) and monotonicity (SRCC) of the general quality evaluation, view synthesis quality evaluation, and view synthesis quality prediction models on the MCL-3D and IST databases are shown in Tables 3–5. The best result is highlighted in boldface, and the second best result is underlined. In Tables 3–5, ‘Post-DIBR’ indicates that the model uses DIBR synthesized images for quality evaluation, and ‘Pre-DIBR’ indicates that the model uses the texture and depth images to predict the quality of view synthesis. ‘GNR’ denotes the general no-reference quality metric and ‘VFR/VRR/VNR’ denotes the full-reference/reduced-reference/no-reference view synthesis quality metric. ‘T’ represents traditional methods, and ‘D’ represents deep learning methods. By comparison, it can be found from Table 3 that the proposed HEDIP model has the best performance in MCL-3D, in terms of both PLCC and SRCC. In addition, in terms of PLCC, the post-DIBR metric SET [30] has the second best performance. In terms of SRCC, the pre-DIBR metric [23] has the second best performance. For VSIM on the IST database (in Table 4), the HEDIP has the best PLCC as well as the second best SRCC. For VSRS on the IST database (in Table 5), the HEDIP delivers the best SRCC while also producing the second best PLCC (very close to the best SET [30]). In summary, the proposed HEDIP model achieves state-of-the-art overall performance. Moreover, as a pre-DIBR model, which is a deep learning model, the HEDIP outperforms the post-DIBR model.

To intuitively understand the performance of the proposed model, Figure 6 shows the texture–depth image pairs with different scenes and distortions, as well as the MOS values of the synthesized image and the predicted scores. From Figure 6a–e, it can be found that the predicted scores are very close to MOS values. Furthermore, when the MOS values increase, the predicted scores of the proposed model also increase. It can be seen that the prediction criteria of the proposed model are in line with the human scoring criteria.

Table 3. Performances of view synthesis quality metrics on the MCL-3D and IST database. The best result is highlighted in boldface, and the second best result is underlined.

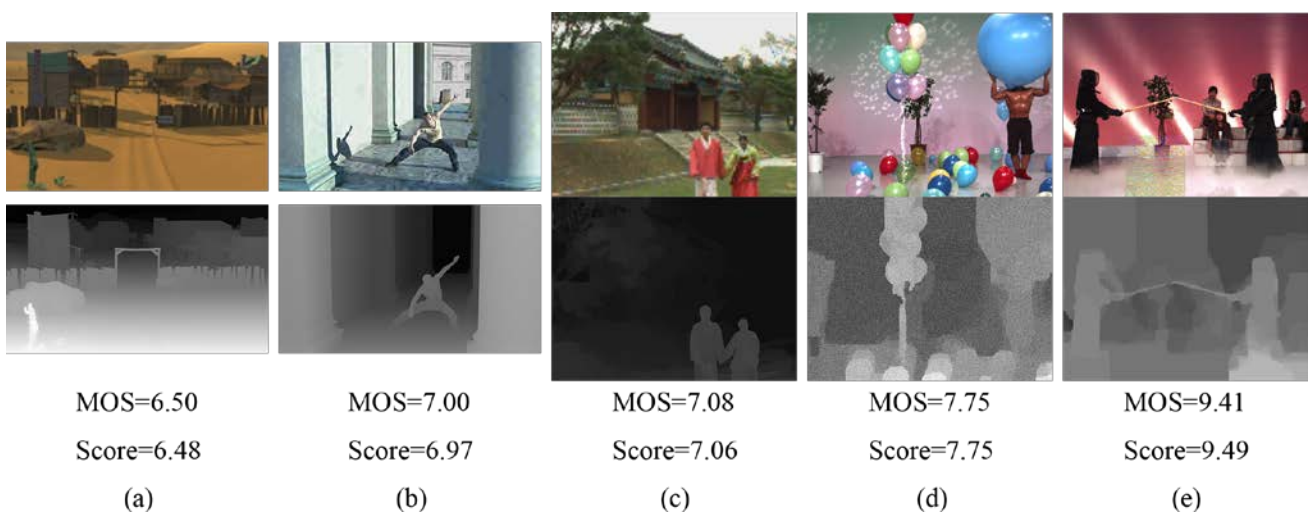
Category	Metric	Type	MCL-3D Database	
			PLCC	SRCC
Post-DIBR	NIQE [16]	GNR-T	0.754	0.710
	BRISQUE [15]	GNR-T	0.694	0.665
	IL-NIQE [40]	GNR-T	0.716	0.645
	M3 [41]	GNR-T	0.561	0.463
	MW-PSNR [42]	VRR-T	0.801	0.810
	MP-PSNR [43]	VRR-T	0.817	0.823
	LOGS [6]	VRR-T	0.726	0.661
	SET [30]	VNR-T	<u>0.918</u>	0.917
	Jakhetiya [44]	VNR-T	0.491	0.476
	NIQSV [20]	VNR-T	0.678	0.622
Pre-DIBR	Metric [23]	VFR-T	0.906	<u>0.918</u>
	HEDIP	VNR-D	0.934	0.927

Table 4. Performances of view synthesis quality metrics on the VSIM-based IST database. The best result is highlighted in boldface, and the second best result is underlined.

Category	Metric	Type	VSIM on IST Database	
			PLCC	SRCC
Post-DIBR	NIQE [16]	GNR-T	0.584	0.586
	BRISQUE [15]	GNR-T	0.651	0.588
	IL-NIQE [40]	GNR-T	0.396	0.379
	M3 [41]	GNR-T	0.662	0.612
	MW-PSNR [42]	VRR-T	0.684	0.677
	MP-PSNR [43]	VRR-T	0.722	0.727
	LOGS [6]	VRR-T	0.630	0.627
	SET [30]	VNR-T	0.815	0.803
	Jakhetiya [44]	VNR-T	0.357	0.367
	NIQSV [20]	VNR-T	0.377	0.359
Pre-DIBR	Metric [23]	VFR-T	/	/
	HEDIP	VNR-D	0.866	<u>0.787</u>

Table 5. Performances of view synthesis quality metrics on the VSRS-based IST database. The best result is highlighted in boldface, and the second best result is underlined.

Category	Metric	Type	VSRS on IST Database	
			PLCC	SRCC
Post-DIBR	NIQE [16]	GNR-T	0.640	0.620
	BRISQUE [15]	GNR-T	0.745	0.711
	IL-NIQE [40]	GNR-T	0.613	0.599
	M3 [41]	GNR-T	0.713	<u>0.719</u>
	MW-PSNR [42]	VRR-T	0.572	0.564
	MP-PSNR [43]	VRR-T	0.552	0.535
	LOGS [6]	VRR-T	0.634	0.608
	SET [30]	VNR-T	0.753	0.710
	Jakhetiya [44]	VNR-T	0.504	0.343
	NIQSV [20]	VNR-T	0.521	0.455
Pre-DIBR	Metric [23]	VFR-T	/	/
	HEDIP	VNR-D	<u>0.750</u>	0.767

**Figure 6.** Texture–depth image pairs for six different scenes; the top row shows the texture image, and the bottom row shows the depth image. Each image pair comes with the MOS value and the predicted score. (a–e) are different scenarios.

4.3. Performance on Different Distortions

The MCL-3D database includes six distortion types. In order to investigate the performance of the proposed HEDIP model on different distortion types, we test images of six distortion types, respectively. Figure 7a–f show the radar plots of the proposed model with different distortion types on the MCL-3D test set; the blue line is the MOS value, and the orange line is the predicted score. The closer the two lines are, the more accurate the model is. On the other hand, the more similar the shapes are, the more monotonic the model is. It can be intuitively found from the radar plots that the HEDIP model still has excellent accuracy and monotonicity under different distortion types. Further, the MOS value for each distorted image pair is very close to the ground truth (given in Figure 8).

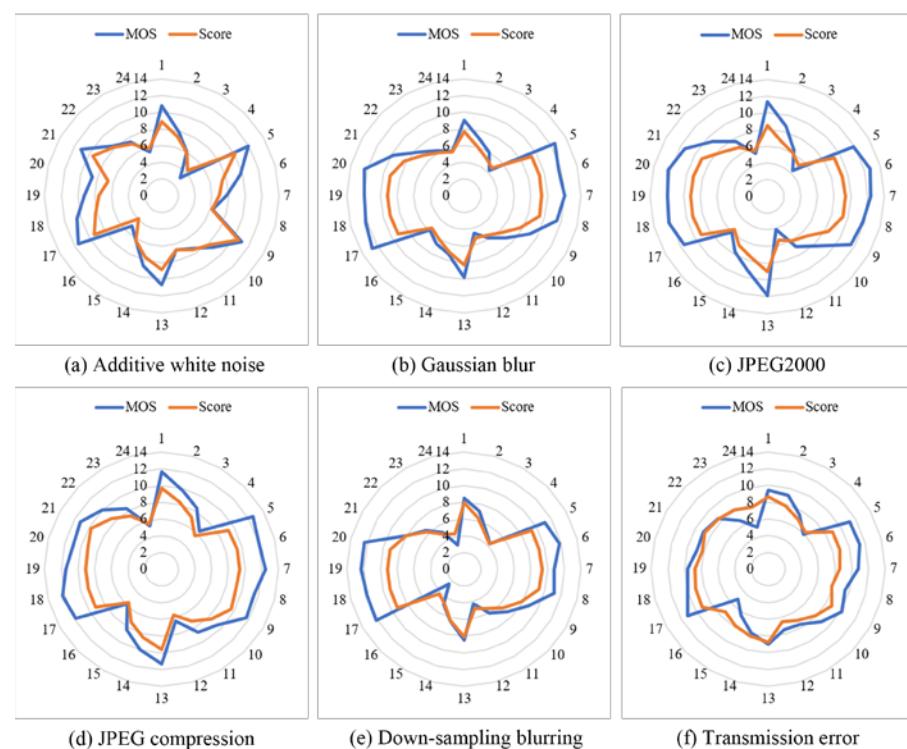


Figure 7. Radar plots of the MOS values and the predicted scores on the MCL-3D test set: (a) test result with additive white noise; (b) test result with Gaussian blur; (c) test result with JPEG2000; (d) test result with JPEG compression; (e) test result with downsampling blurring; (f) test result with transmission error.

4.4. Ablation Study

To further demonstrate the effectiveness of the proposed modules, we conduct a series of ablation experiments based on the MCL-3D database, which use the same environment configuration as before. We use TCCNN without any modules as the base model. Then, the CMLFF module and the HDP module are added to the base model in turn for experiments.

The experimental results are shown in Table 6. When CMLFF and HDP modules are added, the performance reaches the state of the art. From this result, we can see the importance and pertinence of each module. Moreover, it can be observed from Figure 9 that the basic TCCNN model outperforms most state-of-the-art view synthesis quality metrics.

Table 6. Ablation results on the MCL-3D database. The HEDIP is the model proposed in this paper.

Module	PLCC	SRCC
TCCNN	0.898	0.894
TCCNN + CMLFF	0.905	0.904
TCCNN + CMLFF + HDP (HEDIP)	0.934	0.927

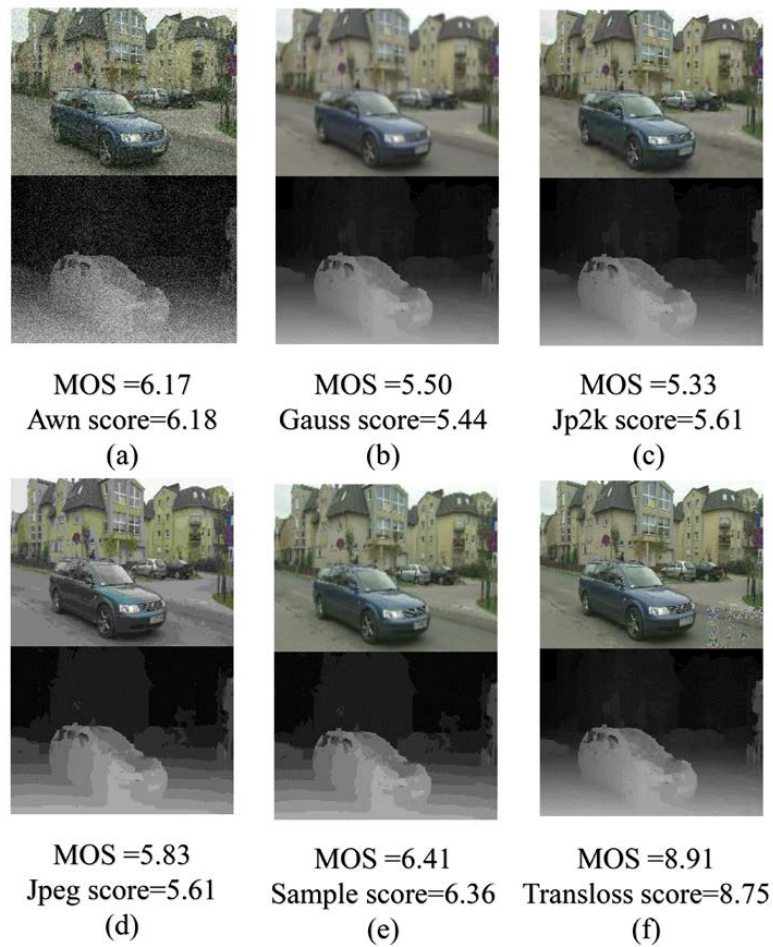


Figure 8. Texture–depth image pairs for six different distortions; each image pair comes with the MOS value and the predicted score: (a) additive white noise; (b) Gaussian blur; (c) JPEG2000; (d) JPEG compression; (e) downsampling blurring; (f) transmission error.

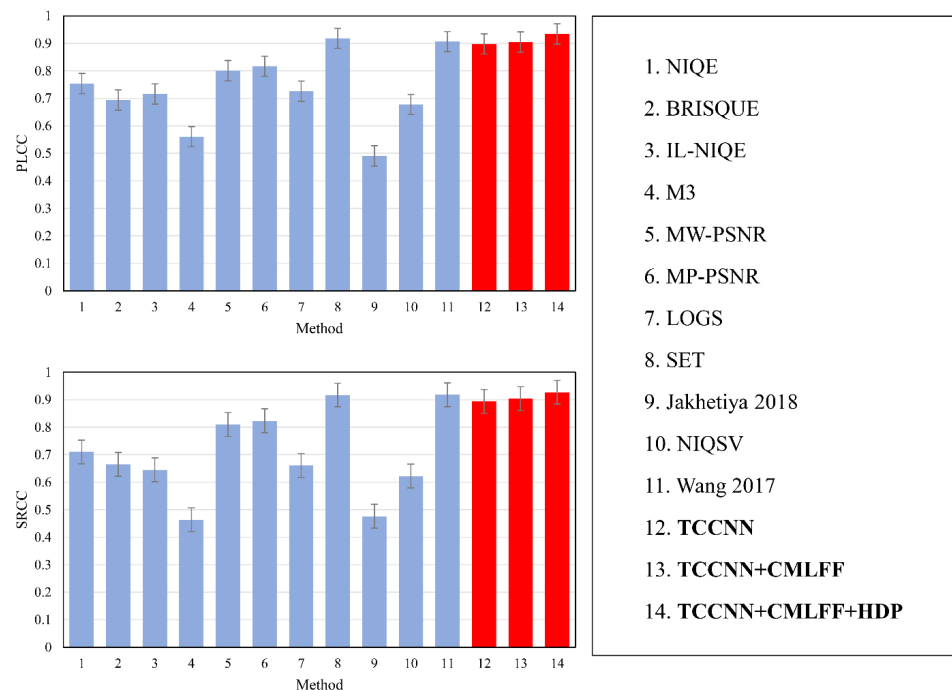


Figure 9. Performance and standard error bars of different methods on the MCL-3D database [23,44].

5. Discussion

The current quality assessment methods for view synthesis basically use hand-designed features. Due to the shallow feature extraction of hand-designed methods, the performance improvement of traditional methods is relatively slow. Inspired by the above efforts, we proposed a blind quality prediction model based on heterogeneous distortion perception, which predicts the image quality of view synthesis from pre-synthesis texture and depth images. The proposed deep learning model is a two-channel architecture that can extract features in the spatial and gradient domains. Furthermore, due to the presence of local distortion in the view synthesis image, we address a heterogeneous distortion perception module to provide effective training labels for each image patch. The experimental results demonstrate the effectiveness of the proposed model.

The quality prediction model can make the view synthesis system more flexible, considering that if the input color/depth images cannot generate satisfactory synthesized viewpoint (by prediction), their quality can be adjusted before sending to the time-consuming DIBR process. The current quality assessment methods for view synthesis basically use hand-designed features, while convolutional neural networks can learn more effective features, which may promote the development of quality assessment technology for view synthesis. Although our model achieves very high performance in predicting the quality of view synthesis, we believe that further improvements to the backbone network in future work may still have the potential to improve the overall performance of the model. The work in this paper mainly evaluates the quality of view synthesis of images. With the demand for high-quality visuals, evaluating the view synthesis quality of videos is a very promising direction. Therefore, in following work, we may extend from the two-dimensional quality evaluation to the three-dimensional quality evaluation; of course, this will be challenging.

6. Conclusions

The quality of synthesized images affects the development and application of DIBR-based view synthesis technology. Most of the current view synthesis quality metrics evaluate the image quality after DIBR-based view synthesis and use hand-crafted methods to extract features. On the one hand, the DIBR process is computationally expensive. On the other hand, shallower hand-crafted features may affect the performance improvement. To tackle these problems, we have proposed a blind quality prediction model based on heterogeneous distortion perception, which predicts the image quality of view synthesis from pre-synthesis texture and depth images. To the best of our knowledge, the proposed model is the first to apply deep learning in the field of view synthesis quality assessment, while predicting the synthesized images without the complex DIBR process. The proposed model has been designed as a two-channel convolutional neural network structure, which can extract spatial and gradient domain features separately. Furthermore, we have designed a heterogeneous distortion perception module, which can provide effective training labels for image patches in source images. Extensive experiments have been conducted on two public view synthesis image databases. The experimental results have demonstrated the superior performance of the proposed model.

The work of this paper is to predict the image quality after view synthesis without DIBR-based view synthesis, which will make the view synthesis system more sensitive. If the predicted synthesis quality is low before synthesis, it can be adjusted in time to avoid complex calculations. In future work, improving the backbone network of the proposed model can optimize the quality prediction performance. Due to the strong ability of deep learning to learn features, the wider application of convolutional neural networks in the field of quality evaluation of view synthesis may promote the development of this field.

Author Contributions: H.S. was responsible for the experimental design and execution; H.S. and G.W. edited and reviewed the manuscript; H.S., L.W. and G.W. contributed to the writing of the paper with the assistance of all authors; H.S. created the computer code and algorithms supporting it. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (grant No. 61972239 and 62071122), the Key Research and Development Program Projects of Shaanxi Province (grant No. 2020GY-024 and 2021GY-182).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The used datasets were obtained from publically open source datasets from: 1. MCL-3D: <http://mcl.usc.edu/mcl-3d-database/>, (1 September 2022); 2. IST: <https://github.com/jascenso/ISTSynthesizeDataset>, (1 September 2022).

Acknowledgments: We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this paper.

DIBR	Depth-Image-Based Rendering
TCCNN	Two-Channel Convolutional Neural Network
CMLFF	Contextual Multi-Level Feature Fusion
HDP	Heterogeneous Distortion Perception
IQA	Image Quality Assessment
FR	Full-Reference
RR	Reduced-Reference
NR	No-Reference
DCT	Discrete Cosine Transform
CNN	Convolutional Neural Network
MOS	Mean Opinion Score
VSRS	View Synthesis Reference Software
PLCC	Pearson Linear Correlation Coefficient
SRCC	Spearman Rank order Correlation Coefficient

References

1. Tanimoto, M.; Tehrani, M.P.; Fujii, T.; Yendo, T. Free-viewpoint TV. *IEEE Signal Process. Mag.* **2010**, *28*, 67–76. [[CrossRef](#)]
2. Tehrani, M.P.; Senoh, T.; Okui, M.; Yamamoto, K.; Inoue, N.; Fujii, T.; Nakamura, H. *Proposal to Consider a New Work Item and Its Use Case-Rei: An Ultra-Multiview 3D Display*; ISO/IEC JTC1/SC29/WG11/m30022; ISO: Geneva, Switzerland, 2013.
3. Bosc, E.; Pepion, R.; Le Callet, P.; Koppel, M.; Ndjiki-Nya, P.; Pressigout, M.; Morin, L. Towards a new quality metric for 3-D synthesized view assessment. *IEEE J. Sel. Top. Signal Process.* **2011**, *5*, 1332–1343. [[CrossRef](#)]
4. Huang, Y.; Li, L.; Zhu, H.; Hu, B. Blind quality index of depth images based on structural statistics for view synthesis. *IEEE Signal Process. Lett.* **2020**, *27*, 685–689. [[CrossRef](#)]
5. Valjarević, A.; Djekić, T.; Stevanović, V.; Ivanović, R.; Jandžiković, B. GIS numerical and remote sensing analyses of forest changes in the Toplica region for the period of 1953–2013. *Appl. Geogr.* **2018**, *92*, 131–139. [[CrossRef](#)]
6. Li, L.; Zhou, Y.; Gu, K.; Lin, W.; Wang, S. Quality assessment of DIBR-synthesized images by measuring local geometric distortions and global sharpness. *IEEE Trans. Multimed.* **2018**, *20*, 914–926. [[CrossRef](#)]
7. Pan, Z.; Yuan, F.; Lei, J.; Fang, Y.; Shao, X.; Kwong, S. VCRNet: Visual Compensation Restoration Network for No-Reference Image Quality Assessment. *IEEE Trans. Image Process.* **2022**, *31*, 1613–1627. [[CrossRef](#)]
8. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)]
9. Sheikh, H.R.; Bovik, A.C.; De Veciana, G. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Trans. Image Process.* **2005**, *14*, 2117–2128. [[CrossRef](#)]
10. Sheikh, H.R.; Bovik, A.C. Image information and visual quality. *IEEE Trans. Image Process.* **2006**, *15*, 430–444. [[CrossRef](#)]

11. Gu, K.; Zhai, G.; Yang, X.; Zhang, W. A new reduced-reference image quality assessment using structural degradation model. In Proceedings of the 2013 IEEE International Symposium on Circuits and Systems (ISCAS), Brijing, China, 19–23 May 2013; pp. 1095–1098.
12. Wang, Z.; Wu, G.; Sheikh, H.R.; Simoncelli, E.P.; Yang, E.-H.; Bovik, A.C. Quality-aware images. *IEEE Trans. Image Process.* **2006**, *15*, 1680–1689. [[CrossRef](#)]
13. Tang, Y.; Ren, F.; Pedrycz, W. Fuzzy C-means clustering through SSIM and patch for image segmentation. *Appl. Soft Comput.* **2020**, *87*, 105928. [[CrossRef](#)]
14. Moorthy, A.K.; Bovik, A.C. A two-step framework for constructing blind image quality indices. *IEEE Signal Process. Lett.* **2010**, *17*, 513–516. [[CrossRef](#)]
15. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. [[CrossRef](#)] [[PubMed](#)]
16. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a “completely blind” image quality analyzer. *IEEE Signal Process. Lett.* **2013**, *20*, 209–212. [[CrossRef](#)]
17. Kang, L.; Ye, P.; Li, Y.; Doermann, D. Convolutional neural networks for no-reference image quality assessment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1733–1740.
18. Yan, Q.; Gong, D.; Zhang, Y. Two-stream convolutional networks for blind image quality assessment. *IEEE Trans. Image Process.* **2019**, *28*, 2200–2211. [[CrossRef](#)]
19. Yan, B.; Bare, B.; Tan, W. Naturalness-aware deep no-reference image quality assessment. *IEEE Trans. Multimed.* **2019**, *21*, 2603–2615. [[CrossRef](#)]
20. Tian, S.; Zhang, L.; Morin, L.; Deforges, O. NIQSV: A no reference image quality assessment metric for 3D synthesized views. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 1248–1252.
21. Tian, S.; Zhang, L.; Morin, L.; Déforges, O. NIQSV+: A no-reference synthesized view quality assessment metric. *IEEE Trans. Image Process.* **2018**, *27*, 1652–1664. [[CrossRef](#)]
22. Gu, K.; Jakhetiya, V.; Qiao, J.-F.; Li, X.; Lin, W.; Thalmann, D. Model-based referenceless quality metric of 3D synthesized images using local image description. *IEEE Trans. Image Process.* **2018**, *27*, 394–405. [[CrossRef](#)]
23. Wang, J.; Wang, S.; Zeng, K.; Wang, Z. Quality assessment of multi-view-plus-depth images. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 85–90.
24. Shao, F.; Yuan, Q.; Lin, W.; Jiang, G. No-reference view synthesis quality prediction for 3-D videos based on color–depth interactions. *IEEE Trans. Multimed.* **2017**, *20*, 659–674. [[CrossRef](#)]
25. Li, L.; Huang, Y.; Wu, J.; Gu, K.; Fang, Y. Predicting the quality of view synthesis with color–depth image fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 2509–2521. [[CrossRef](#)]
26. Thornton, M.W.; Atkinson, P.M.; Holland, D. Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. *Int. J. Remote Sens.* **2006**, *27*, 473–491. [[CrossRef](#)]
27. Wang, M.; Shang, X. A fast image fusion with discrete cosine transform. *IEEE Signal Process. Lett.* **2020**, *27*, 990–994. [[CrossRef](#)]
28. Sandić-Stanković, D.D.; Kukolj, D.D.; Le Callet, P. Fast blind quality assessment of DIBR-synthesized video based on high-high wavelet subband. *IEEE Trans. Image Process.* **2019**, *28*, 5524–5536. [[CrossRef](#)] [[PubMed](#)]
29. Gu, K.; Qiao, J.; Lee, S.; Liu, H.; Lin, W.; Le Callet, P. Multiscale natural scene statistical analysis for no-reference quality evaluation of DIBR-synthesized views. *IEEE Trans. Broadcast.* **2019**, *66*, 127–139. [[CrossRef](#)]
30. Zhou, Y.; Li, L.; Wang, S.; Wu, J.; Gao, X. No-reference quality assessment for view synthesis using DoG-based edge statistics and texture naturalness. *IEEE Trans. Image Process.* **2019**, *28*, 4566–4579. [[CrossRef](#)]
31. Wang, G.; Wang, Z.; Gu, K.; Li, L.; Xia, Z.; Wu, L. Blind quality metric of DIBR-synthesized images in the discrete wavelet transform domain. *IEEE Trans. Image Process.* **2019**, *29*, 1802–1814. [[CrossRef](#)]
32. Li, L.; Zhou, Y.; Wu, J.; Li, F.; Shi, G. Quality index for view synthesis by measuring instance degradation and global appearance. *IEEE Trans. Multimed.* **2021**, *23*, 320–332. [[CrossRef](#)]
33. Ahmed, N.; Natarajan, T.; Rao, K.R. Discrete cosine transform. *IEEE Trans. Comput.* **1974**, *C-23*, 90–93. [[CrossRef](#)]
34. Huang, Y.; Meng, X.; Li, L. No-reference quality prediction for DIBR-synthesized images using statistics of fused color–depth images. In Proceedings of the 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Shenzhen, China, 6–8 August 2020; pp. 135–138.
35. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
36. Song, R.; Ko, H.; Kuo, C. MCL-3D: A database for stereoscopic image quality assessment using 2D-image-plus-depth source. *arXiv* **2014**, arXiv:1405.1403.
37. Rodrigues, F.; Ascenso, J.; Rodrigues, A.; Queluz, M.P. Blind quality assessment of 3-D synthesized views based on hybrid feature classes. *IEEE Trans. Multimed.* **2019**, *21*, 1737–1749. [[CrossRef](#)]
38. Farid, M.S.; Lucenteforte, M.; Grangetto, M. Depth image based rendering with inverse mapping. In Proceedings of the 2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSp), Cagliari, Italy, 30 September–2 October 2013; pp. 135–140.
39. Huang, Y.; Zhou, Y.; Hu, B.; Tian, S.; Yan, J. DIBR-synthesized video quality assessment by measuring geometric distortion and spatiotemporal inconsistency. *Electron. Lett.* **2020**, *56*, 1314–1317. [[CrossRef](#)]

40. Zhang, L.; Zhang, L.; Bovik, A.C. A feature-enriched completely blind image quality evaluator. *IEEE Trans. Image Process.* **2015**, *24*, 2579–2591. [[CrossRef](#)] [[PubMed](#)]
41. Xue, W.; Mou, X.; Zhang, L.; Bovik, A.C.; Feng, X. Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features. *IEEE Trans. Image Process.* **2014**, *23*, 4850–4862. [[CrossRef](#)]
42. Sandić-Stanković, D.; Kukolj, D.; Le Callet, P. DIBR synthesized image quality assessment based on morphological wavelets. In Proceedings of the 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), Pilos, Greece, 26–29 May 2015; pp. 1–6.
43. Sandic-Stankovic, D.; Kukolj, D.; Le Callet, P. Multi-scale synthesized view assessment based on morphological pyramids. *J. Electr. Eng.* **2016**, *67*, 3. [[CrossRef](#)]
44. Jakhetiya, V.; Gu, K.; Singhal, T.; Guntuku, S.C.; Xia, Z.; Lin, W. A highly efficient blind image quality assessment metric of 3-D synthesized images using outlier detection. *IEEE Trans. Ind. Inform.* **2018**, *15*, 4120–4128. [[CrossRef](#)]
45. Tang, Y.; Pan, Z.; Pedrycz, W.; Ren, F.; Song, X. Viewpoint-Based Kernel Fuzzy Clustering With Weight Information Granules. *IEEE Trans. Emerg. Top. Comput. Intell.* **2022**, *2022*, 1–15. [[CrossRef](#)]