entropy

MDPI

Article

No-Reference Quality Assessment for 3D Synthesized Images Based on Visual-Entropy-Guided Multi-Layer Features Analysis

Chongchong Jin [1], Zongju Peng [1,2,*], Wenhui Zou [1], Fen Chen [1,2], Gangyi Jiang [1] and Mei Yu [1]

[1] Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China; jinchongchong94@163.com (C.J.); zouwench@163.com (W.Z.); chenfen@cqut.edu.cn (F.C.); jianggangyi@126.com (G.J.); yumei@nbu.edu.cn (M.Y.)

[2] School of Electrical and Electronic Engineering, Chongqing University of Technology, Chongqing 400054, China

* Correspondence: pengzongju@126.com

Abstract: Multiview video plus depth is one of the mainstream representations of 3D scenes in emerging free viewpoint video, which generates virtual 3D synthesized images through a depth-image-based-rendering (DIBR) technique. However, the inaccuracy of depth maps and imperfect DIBR techniques result in different geometric distortions that seriously deteriorate the users' visual perception. An effective 3D synthesized image quality assessment (IQA) metric can simulate human visual perception and determine the application feasibility of the synthesized content. In this paper, a no-reference IQA metric based on visual-entropy-guided multi-layer features analysis for 3D synthesized images is proposed. According to the energy entropy, the geometric distortions are divided into two visual attention layers, namely, bottom-up layer and top-down layer. The feature of salient distortion is measured by regional proportion plus transition threshold on a bottom-up layer. In parallel, the key distribution regions of insignificant geometric distortion are extracted by a relative total variation model, and the features of these distortions are measured by the interaction of decentralized attention and concentrated attention on top-down layers. By integrating the features of both bottom-up and top-down layers, a more visually perceptive quality evaluation model is built. Experimental results show that the proposed method is superior to the state-of-the-art in assessing the quality of 3D synthesized images.

Keywords: 3D synthesized images; image quality assessment (IQA); no-reference; visual-entropy-guided; multi-layer features analysis

check for updates

Citation: Jin, C.; Peng, Z.; Zou, W.; Chen, F.; Jiang, G.; Yu, M. No-Reference Quality Assessment for 3D Synthesized Images Based on Visual-Entropy-Guided Multi-Layer Features Analysis. *Entropy* **2021**, *23*, 770. https://doi.org/10.3390/e23060770

Academic Editor: Ke Gu

Received: 30 April 2021
Accepted: 14 June 2021
Published: 18 June 2021

1. Introduction

With the advancement of video technologies, a free viewpoint video (FVV) system is gradually applied to various fields, such as distance education, medical service, and entertainment [1]. Compared with traditional 2D videos, users can interactively embody 3D scenes from arbitrary viewpoints in the FVV system. Unfortunately, limited by equipment and cost, capturing all views of FVV via camera is unrealistic and needs the existence of virtual synthesized viewpoints to enhance the scene switching continuity. Multiview video plus depth is one of the mainstream representations of 3D scenes, which generate virtual synthesized images through depth-image-based-rendering (DIBR) techniques [2]. At this stage, the inaccuracy of depth maps and imperfect DIBR techniques result in different geometric distortions which seriously deteriorate the users' visual perception. In addition, it is time-consuming and impracticable to screen the quality of massive synthesized images by humans. Hence, designing an effective image quality assessment (IQA) metric [3] via human visual simulation to measure the image quality deterioration and further determine the application feasibility of 3D synthesized views is a significant research topic.

So far, extensive IQA methods were designed for the traditional distortions in 2D images, such as JPEG/JPEG2K compression [4,5], Gaussian white noise [6], Gaussian blur [7], blocking [8], and fast fading channel errors [9]. Generally, these distortions globally distribute in entire 2D images. In contrast, the 3D synthesized geometric distortions appear in local areas, and seriously destroy the structural semantic information of synthesized views. Due to the particularity of synthetic distortions, the existing IQA methods for 2D traditional distortions, like [10–14], cannot measure the 3D synthesized distortions effectively. With this concern, some researchers have proposed IQA metrics targeting 3D synthesized images. These methods are mainly divided into two categories, full-reference (FR) [15–23] and no-reference (NR) [24–33].

Bosc et al. explored the necessity of designing synthesized IQA metric, and evaluated the image quality via pixel deviation [15]. Conze et al. designed an SSIM-based view synthesis quality assessment (VSQA) metric, which mainly researched the synthesized view quality degradation caused by shift artifacts [16]. Battisti et al. statistically analyzed the shift artifacts in the Haar wavelet sub-bands, and proposed a 3D synthesized view image quality metric (3DSwIM) [17]. Ling and Le Callet proposed a sketch-token-based synthesized IQA (ST-SIQA) metric [18] and elastic metric based IQA (EM-IQA) metric [19]. Both ST-SIQA and EM-IQA analyzed shift artifacts by calculating contour similarity between the reference and synthesized images. Sandić-Stanković et al. designed two IQA metrics, i.e., morphological wavelet peak signal-to-noise ratio (MW-PSNR) [20] and morphological pyramid peak signal-to-noise ratio (MP-PSNR) [21], in order to evaluate the quality of synthesized geometric distortions in a transform domain. Tian et al. matched the horizontal displacement between the reference and synthesized images to devise a shift-compensation-based IQA (SC-IQA) metric [22]. Li et al. presented an FR quality metric for visual views by simultaneously measuring local instance degradation and global appearance (IDEA), in which local distortions were detected by discrete orthogonal moments and global sharpness was measured by super-pixel representation [23]. However, FR synthesized IQA metrics are not suitable for real application because the reference images of synthesized view are usually unavailable in FVV systems.

Gu et al. proposed an NR autoregression-plus thresholding (APT) metric based on a natural scene statistical (NSS) model [24]. Lately, Gu et al. considered local and global distortion, and presented a multi-scale NSS-based (MNSS) metric [25]. Jakhetiya et al. counted outliers by a three sigma rule-based robust outlyingness ratio (OUT) to evaluate the quality of synthesized images [26]. Recently, Jakhetiya et al. further proposed a kernel-ridge-regression-based predictor for synthesized IQA, which detected the complete distortion surface with geometric distortions and estimated corresponding quality scores [27]. The NSS-based methods above are time consuming and basically designed for severe geometric distortions. In addition, the metrics based on transform domain are also considered. Sandić-Stanković et al. proposed an NR IQA metric for synthesized videos which combined a high frequency component in a morphological wavelet domain with threshold (NR_MWT) [28]. Wang et al. also extracted features of geometric distortion, global sharpness, and image complexity in a wavelet transform domain to evaluate the quality of 3D synthesized images [29]. These transform-domain-based metrics eliminate uninterested information of synthesized image and save calculation time but are still sensitive to limited geometric distortion types. Based on this, Zhou et al. analyzed synthesized images using Difference-of-Gaussian-based edge statistics and texture naturalness (SET) to measure different types of geometric distortions [30]. Tian et al. proposed an NR IQA of synthesized views (NIQSV), which measured the blurry and crumbling distortions by opening and closing operations [31]. Subsequently, Tian et al. further analyzed the hole and stretching distortions, and advanced the NIQSV to NIQSV+ [32]. Likewise, Yue et al. classified the distortions, and combined local and global features to measure 3D synthesized images (CLGM) [33]. These distortion-classification-based metrics targeted measure multiple distortion types and are more comprehensive. The pity is that the synthesized image degradation caused by weak geometric distortions has not received enough attention. Furthermore, few deep-learning-based metrics were exploratively used to evaluate the quality of 3D synthesized images. Ling et al. proposed a generative adversarial networks-based

NR metric (GANs-NRM) for synthesized images, which expanded the distortion sample through the GANs, then used a 'bag of distortion word' codebook to classify the distortion, and finally used the support vector machine to regress the quality score [34]. However, it only uses the network to expand the training samples and does not achieve end-to-end score learning. Wang et al. built a new synthesized database including 504 pictures to expand the ground-truth of training and utilized the local saliency to weight the predicted scores [35]. Unfortunately, the database samples proposed by this method are still limited. Thus, how to evaluate the synthesized images using an end-to-end deep learning model though the small database still remains an open problem.

In summary, the existing IQA metrics for 3D synthesized images still have some limitations. (1) The reference images are not accessible in the FVV system. (2) Most of the existing IQA metrics search geometric distortions though the entire image, which have difficulty measuring local-distributed distortions in synthesized images. (3) Although the performance of the distortion-classification-based IQA metrics is competitive, they have room for further improvement in terms of weak geometric distortion measurement.

In this paper, a novel NR IQA metric based on visual-entropy-guided multi-layer features analysis (MLFA) is proposed. Extensive experiments exhibit that MLFA has a better performance than the prevailing IQA metrics and strong robustness on different databases. The main contributions of MLFA are as follows:

(1) The metric elaborately classifies geometric distortions into bottom-up and top-down layers via visual entropy, and integrates multi-layer features to regress quality score.
(2) In the bottom-up layer, the strong geometric distortion is measured by calculating area proportion plus transition threshold.
(3) In the top-down layer, key regions of weak geometric distortions are extracted by the relative total variation model, and the features are measured by the interaction of decentralized attention (entropy, secondary Gaussian blur similarity, and horizontal pixels correlation) and concentrated attention (Gaussian mixture models).

The rest of this paper is organized as follows. The motivation of our method is detailed in Section 2. Section 3 describes the visual-entropy-guided MLFA method for synthesized images. Section 4 presents the experimental results. Finally, conclusions are drawn in Section 5.

2. Motivation

Figure 1 shows the visual comparison pair of geometric distortions, the left and right are the local areas of original and distorted images, respectively, and all subfigures are originated from the IRCCyN_IVC_DIBR_images database [36]. Different from traditional 2D distortions, 3D synthesized geometric distortions are mainly caused by inaccurate depth map and DIBR techniques.

Figure 1a shows the hole distortion. Occlusion and exposure are the main reasons for the hole generation. If one object is occluded in the real view and exposed in the virtual view, the corresponding region in virtual view cannot be warped from the real view. Consequently, a hole is generated. Most of the hole phenomenon occurs in the depth abrupt areas.

To tackle the hole problem, many scholars preprocessed the depth video. For instance, Fehn et al. [37] used a low-pass filter to smooth the depth information. By this method, the hole problem can be alleviated in synthesized images, but inaccurate depth information also brings the geometric distortions, curving, and object shifting, which are respectively visible in the chair and face of Figure 1b,c. One can see that the distortions are particularly perceptible in background and foreground transitions.

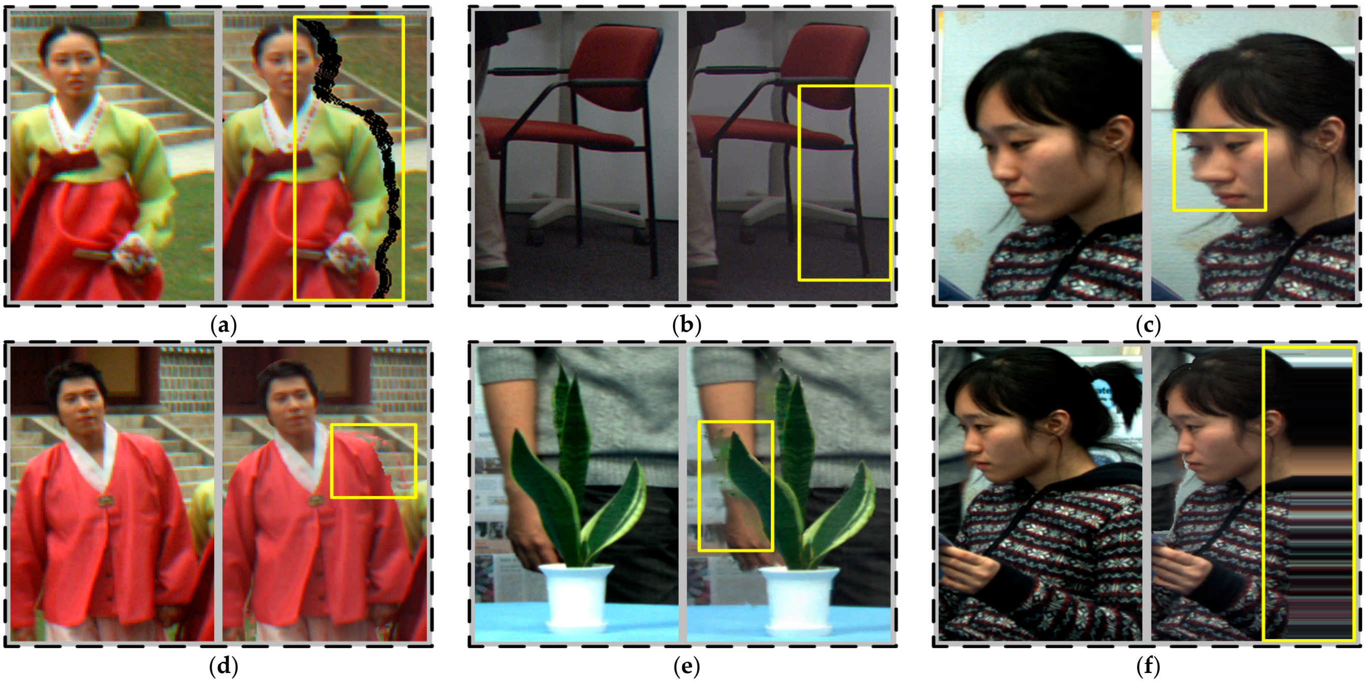


Figure 1. 3D synthesized geometric distortions, (**a**) hole; (**b**) curving; (**c**) object shifting; (**d**) ghosting; (**e**) blurry; (**f**) stretching.

In addition, the filling algorithms for hole regions also bring distortions. Figure 1d shows the ghosting phenomenon based on patch-based synthesis methods [38,39]. This distortion generates when the pixels from the optimal matching patch do not fit the actual scene. Figure 1e shows a rendered result when the hole area is filled using the methods in [40,41]. As we can observe, the in-painting method cannot effectively fill holes in complex texture areas, which result in blurry distortion at the boundaries of the potted plant and the man's arm. Additionally, the stretching distortion mainly occurs on the left/right side of image and is produced by a particular in-painting method [42], which fills holes with existing horizontal adjacent pixels, as shown in Figure 1f.

According to the observations of above synthesized distortions, we find that people cannot distinguish the specific distortion types without professional training, and can only roughly evaluate the degree of image quality degeneration. Therefore, the mess types of synthesized distortions are regulated for unified measurement. The distortions, caused by inaccurate depth information, i.e., curving and object shifting, are classified to 'deforming'. The distortions that are manifested as the pixel overflow and caused by the inaccurate filling algorithm, i.e., ghosting and blurry, are collectively called 'blurry'. Simultaneously, we find that the geometric distortion often occurs in the local areas of synthesized images, especially on the left/right side of images and the boundary areas of objects.

Biologically, visual stimuli enter the primary visual cortex for the short term and progress along two parallel hierarchical streams, i.e., the brain neurons are divided into two major regions to control the attention mechanism. The 'dorsal stream' mainly processes visual information in the posterior parietal cortex and is concerned with directing attention. The 'ventral stream' processes stimuli in the inferotemporal cortex, focusing on recognition capability [43]. The dorsal and ventral streams must interact to achieve good scene understanding. However, the fusion of two streams to process information is simple for the human brain, but challenging for the computer. Otherwise, implementing two streams at the same time has an obstacle that only small parts of visual stimuli are stored as short-term memory [44]. Hence, processing a large amount of sensory information in one step is unrealistic.

We focus our research by combining biological theories, and the distinction of energy entropy included in different stimuli (i.e., distortions) is huge, which may cause different distortions to be processed in different visual cortexes. This presumption is indeed verified by some studies—that there exists an approximately linear relationship between energy entropy and a visual attention mechanism [45,46]. Thus, a two-component framework for visual attention mechanism stimulated by stimuli energy entropy (short for visual entropy) was proposed to simulate the physiological structure of the human brain processing visual information [47]. The framework suggests a human selective attention scene though bottom-up and top-down mechanisms. The bottom-up mechanism means that a stimulus with high energy is sufficiently salient and can pop out of a visual scene, which will take 20–50 ms reaction time of human attention. On the contrary, for the top-down mechanism, like a task in which people need to move their eyes to find low energy scenes, such volitional attention will take 200 ms or more reaction time. Inspired by this theory, hole distortion can pop out in an image due to its obviousness, which tends to be a bottom-up mechanism. Other distortions are interfered by complex textures and require careful observation, which takes a longer reaction time and tends to be a top-down mechanism. In addition, the visual attention mechanism is affected by inhibition of return (the current attention will not be attended again), so both bottom-up and top-down mechanisms can operate in parallel.

Particularly, the performance of top-down attention is controlled by complex brain regions, such as the frontal lobes. Hence, it is difficult to express visual perception by integration of the various scene features. Treisman and Gelade proposed a feature integration theory [48], which came up with two visual attention mechanisms, decentralized attention and concentrated attention. The former is a decentralized search for different features of the scene (e.g., color, shape). The latter is mainly a concentrated search for the scene where various features are mixed. The decentralized attention is a single-dimensional feature extraction, which has strong pertinence and information dependence. By contrast, the concentrated attention is a multi-dimensional extraction of mixed features, which has strong robustness to information update. Therefore, we consider extracting the distortions of the top-down layer via feature integration theory (i.e., decentralized and concentrated attention) to achieve the maximum utilization of features.

Based on all distortion observations, biology and psychology theory, we divide the 3D synthesized geometric distortions into two visual-entropy-guided attention layers. Specifically, the hole distortion is divided into a bottom-up layer because of its eye-catching energy, and insignificant geometric distortions (i.e., deforming, blurry, and stretching) are assigned to a top-down layer. Further focusing on a top-down layer, the key distributed areas (i.e., left/right side of images and the boundary areas of objects) of weak geometric distortion are highlighted, and the decentralized and concentrated attention are combined to measure top-down features based on key areas. By integrating the features of bottom-up and top-down layers, a novel NR IQA metric for 3D synthesized images is built. Extensive experimental results demonstrate the effectiveness and robustness of the proposed method (MLFA).

3. The Proposed Visual-Entropy-Guided MLFA Method

Figure 2 shows the block diagram of the proposed visual-entropy-guided MLFA method, which contains three parts, feature extraction of bottom-up and top-down layers, and quality regression of random forest (RF). The details of each part will be introduced in Sections 3.1–3.3.

3.1. Feature Extraction of the Bottom-Up Layer

Figure 3 shows two kinds of black areas origination in the IRCCyN_IVC_DIBR_images database: natural black object and hole distortion, in which the natural black object does not affect the quality attenuation of the image. Therefore, we need to eliminate the interference caused by natural black objects (i.e., non-hole) when extracting the hole features. Specifically, the regions with a pixel value of 0 are calculated as the candidate areas, as shown in the second subfigures of Figure 3a,b.

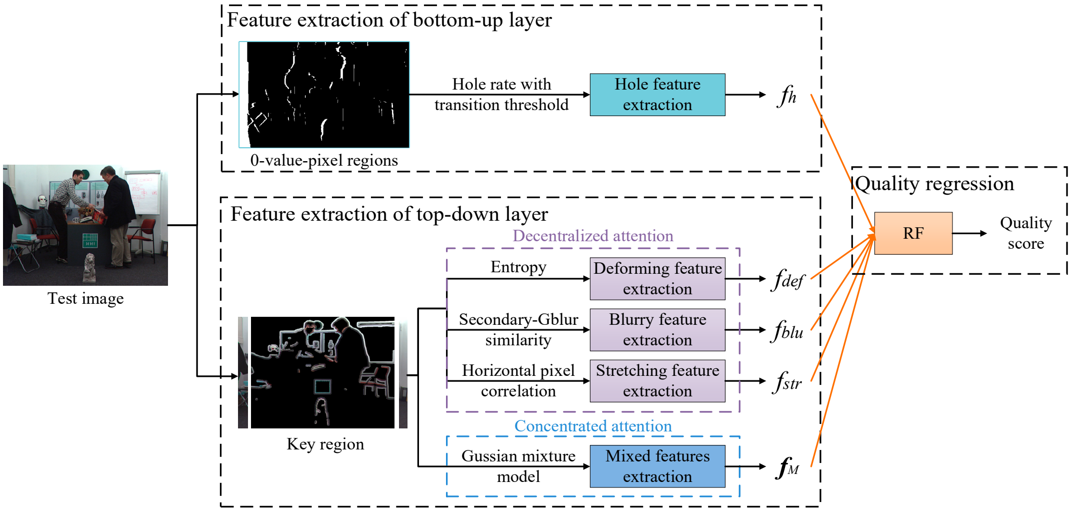
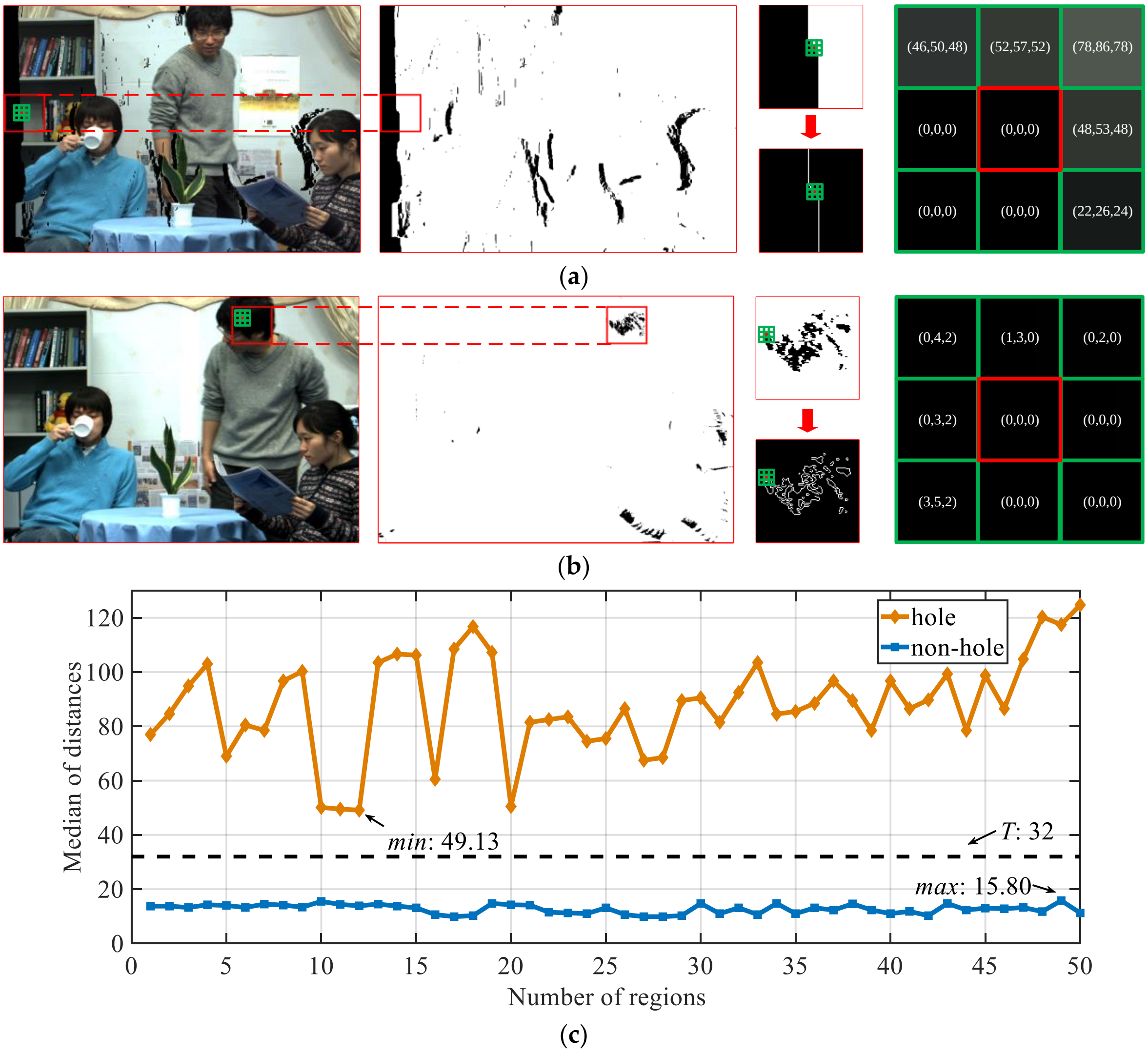

Feature extraction of bottom-up layer
Hole rate with transition threshold
Hole feature extraction
f_h
0-value-pixel regions
Test image
Quality regression
RF
Quality score
Feature extraction of top-down layer
Decentralized attention
Entropy
Deforming feature extraction
f_{def}
Secondary-Gblur similarity
Blurry feature extraction
f_{blu}
Horizontal pixel correlation
Stretching feature extraction
f_{str}
Key region
Concentrated attention
Gussian mixture model
Mixed features extraction
f_M

Figure 2. Block diagram of the visual-entropy-guided MLFA method.



(46,50,48)	(52,57,52)	(78,86,78)
(0,0,0)	(0,0,0)	(48,53,48)
(0,0,0)	(0,0,0)	(22,26,24)

(**a**)

(0,4,2)	(1,3,0)	(0,2,0)
(0,3,2)	(0,0,0)	(0,0,0)
(3,5,2)	(0,0,0)	(0,0,0)

(**b**)

(**c**)

Figure 3. Distinguish the hole and non-hole regions. (**a**) 3D synthesized image with 0-value-pixel regions (hole regions); (**b**) original image with 0-pixel regions (non-hole regions); (**c**) median of distances about the hole regions (orange line) and non-hole regions (blue line).

Subsequently, we find that, compared with non-hole 0-pixel regions, the boundary pixels of hole regions are more abrupt, such as the rightmost subfigures of Figure 3a,b shown. Thus, a statistical method of boundary-pixel transition is introduced. The 0-value-pixel boundaries are obtained by Sobel algorithm (visualized in the third column subfigures of Figure 3a,b, and the Euclidean distance *d* between the current boundary pixel and the predicted boundary pixel are calculated:

$$d(i,j) = \left| b(i,j) - \widetilde{b}(i,j) \right| \tag{1}$$

where *b*(*i*,*j*) represents 0-value boundary pixels at (*i*,*j*). $\widetilde{b}(i,j)$ means the pixel value predicted by the transition statistical method at (*i*,*j*):

$$\widetilde{b}(i,j) = \frac{1}{n \times n} \sum_{q(i,j) \in \Omega 1} q(i,j) \tag{2}$$

where *q*(*i*,*j*) belongs to **Ω1**, which are adjacent pixels surrounding (*i*,*j*) in the $n \times n$ patch.

After that, the same numbers of hole and non-hole 0-pixel regions are respectively selected to get their median of distances as shown in Figure 3c. Based on the size of database [36], the number of 0-pixel regions are set to 50, and the median performance of 1000 calculations is considered as the model to exclude outlier distances. Then, a transition threshold is defined as *T* = *Average* (*min* {hole}, *max* {non-hole}) to distinguish between hole and non-hole regions. Here, *T* is rounded to 32.

To the end, the hole rate is calculated as the feature of the bottom-up layer:

$$f_h = \begin{cases} \frac{Num(R_h)}{W \times H} & Median\{d(i,j)\} > T \\ 0 & otherwise \end{cases} \tag{3}$$

where $Num(\cdot)$ indicates the pixel number. $\mathbf{R}_h$ represents hole regions. *W* and *H* denote the width and height of the test image.

3.2. Feature Extraction of a Top-Down Layer

As mentioned by the APT metric [24], the performance of using the NSS model directly on unprocessed synthesized images is poor, and the histograms of different geometric distortions are quite close to each other as shown in Figure 4a. Thus, to avoid the global 'good quality' information affecting local distorted information, popping out the local distorted regions is indispensable. Due to the distortions that usually occur on the left/right side of the image and boundaries of objects, we consider extracting these two parts as the key region. Fundamentally, the test image is divided into the side region ($\mathbf{R}_s$) and middle region ($\mathbf{R}_m$) according to the image width (*W*):

$$\begin{cases} \mathbf{R}_m(i,j) = \{\mathbf{Y}(i,j) | T_l \cdot W \le i \le (1-T_r) \cdot W,\ 0 \le j \le H\} \\ \mathbf{R}_s(i,j) = \mathbf{Y}(i,j) - \mathbf{R}_m(i,j) \end{cases} \tag{4}$$

where **Y**(*i*,*j*) is the pixel value at (*i*,*j*) in the test image. T_l and T_r are width proportion thresholds, which determine the left and right sides of $\mathbf{R}_m$ in the image.

Further to $\mathbf{R}_m$, inspired by the fact that the image semantic contains structure and texture information [49], we combine image structure extraction with morphological operations to extract the boundaries of objects. Specifically, a relative total variation model is used to extract the structure image **S**:

$$\underset{\mathbf{s}(i,j)}{\operatorname{argmin}} \sum_{i,j} (\mathbf{S}(i,j) - \mathbf{R}_m(i,j))^2 + \lambda \cdot \left(\frac{W_{s,x}(i,j)}{W_{f,x}(i,j) + \varepsilon} + \frac{W_{s,y}(i,j)}{W_{f,y}(i,j) + \varepsilon} \right) \tag{5}$$

where the first term aims to make **S**(*i*,*j*) and $\mathbf{R}_m$(*i*,*j*) similar. λ is a weight which determines the blur degree of the structure image. ε is a small constant to avoid the situation of division-by-zero. $W_{s,x}(i,j)$ and $W_{s,y}(i,j)$ are the values measured by sliding window in x and y directions:

$$\begin{cases} W_{s,x}(i,j) = \sum\limits_{k(i,j)\in\Omega 2} g_{(i,j),k(i,j)} \cdot \left|(\partial_x S)_{k(i,j)}\right| \\ W_{s,y}(i,j) = \sum\limits_{k(i,j)\in\Omega 2} g_{(i,j),k(i,j)} \cdot \left|(\partial_y S)_{k(i,j)}\right| \end{cases} \tag{6}$$

where *k*(*i*,*j*) belongs to **Ω2**, the 3 × 3 neighboring pixels centered at (*i*,*j*). ∂x and ∂y are partial derivatives. $g_{(i,j),k(i,j)}$ is a weighting function, which is proportional to the exponent:

$$g_{(i,j),k(i,j)} \propto \exp(-\frac{(x_{(i,j)} - x_{k(i,j)})^2 + (y_{(i,j)} - y_{k(i,j)})^2}{2\sigma_s{}^2}) \tag{7}$$

where σ_s dominates the scale of the window and controls the scale of the texture element.

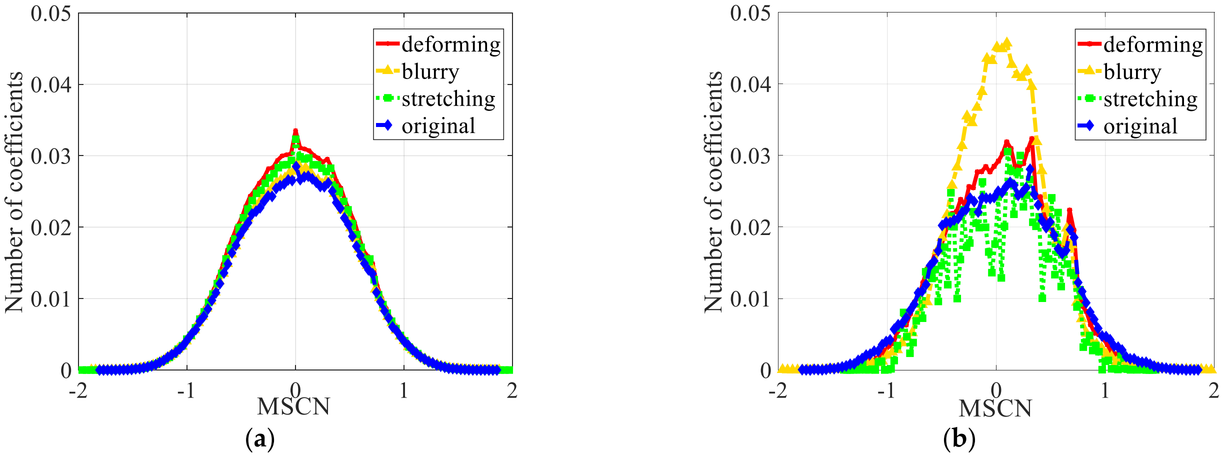


Figure 4. The histograms of the MSCN coefficients of the original image and its corresponding top-down distorted versions. (**a**) MSCN in unprocessed images; (**b**) MSCN in key region $\mathbf{R}_k$.

Similarly, $W_{f,x}(i,j)$ and $W_{f,y}(i,j)$ are measured by a fixed window in Equation (5). They are defined as:

$$\begin{cases} W_{f,x}(i,j) = \left|\sum\limits_{k(i,j)\in\Omega 2} g_{(i,j),k(i,j)} \cdot (\partial_x S)_{k(i,j)}\right| \\ W_{f,y}(i,j) = \left|\sum\limits_{k(i,j)\in\Omega 2} g_{(i,j),k(i,j)} \cdot (\partial_y S)_{k(i,j)}\right| \end{cases} \tag{8}$$

Different from the formula in Equation (6), the value obtained by a fixed window does not include the modulus. Thus, the sum of $\partial_{(\cdot)}S$ directly decides the gradient consistency.

In short, the structure and texture information of **S** depends on two parameters: λ and σ_s. When λ and σ_s are small, **S** contains complex texture information. Otherwise, details of **S** are lost too much to capture object boundaries reasonably. Here, λ and σ_s are experimentally set as 0.02 and 4.

Figure 5 presents the visualized results of the relative total variation model with morphological processing. Specifically, Figure 5a shows the acquired structure image **S**. The structure edge image $\mathbf{S}_e$ and structure mask image $\mathbf{S}_m$ are obtained by the Sobel algorithm and dilation operation, as shown in Figure 5b,c. Figure 5d shows the structure distortion image $\mathbf{S}_d$, which is obtained by **S** × $\mathbf{S}_m$. Finally, the key region $\mathbf{R}_k$ is stitched by $\mathbf{R}_s$ (red boxes in Figure 5e) and $\mathbf{S}_d$ (green box in Figure 5e). In addition, the original edge image $\mathbf{O}_e$, the original mask image $\mathbf{O}_m$, and the original distortion image $\mathbf{O}_d$ are calculated for comparison as shown in Figure 5f–h. It can be found that $\mathbf{O}_d$ is more complicated and chaotic than $\mathbf{S}_d$, which proves that the extracted $\mathbf{R}_k$ can effectively highlight object boundaries' regions with geometric distortion and is consistent with the subjective perception of real synthesized distortion regions.

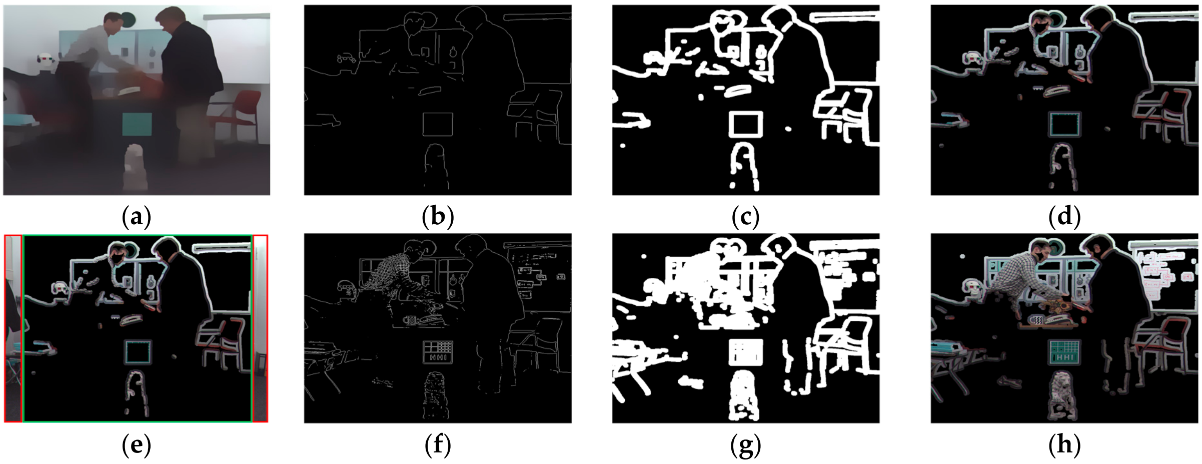


Figure 5. Illustration and visual comparison of key region extraction. (**a**) structure image $\mathbf{S}$; (**b**) structure edge image $\mathbf{S}_e$; (**c**) structure mask image $\mathbf{S}_m$; (**d**) structure distortion image $\mathbf{S}_d$; (**e**) key region $\mathbf{R}_k$; (**f**) original edge image $\mathbf{O}_e$; (**g**) original mask image $\mathbf{O}_m$; (**h**) original distortion image $\mathbf{O}_d$.

After the $\mathbf{R}_k$ is extracted, the feature integration theory (i.e., decentralized and concentrated attentions) is applied to measure the geometric distortions on the top-down layer. On the one hand, the features of geometric distortions on top-down layer are independently extracted by decentralized attention.

For the deforming distortion, it can be observed from Figure 1b,c that the regular pixel arrangement turns to a disorderly distribution after deforming. As a universal cognition, image entropy is a quantity that expresses the degree of disorder of the pixels state. Therefore, we use image entropy to extract the feature of deformation:

$$f_{def} = -\sum_{a}\sum_{b} P_{a,b} \lg P_{a,b} \tag{9}$$

where a is the gray value of the pixel, and b is the average gray value in the 3 × 3 neighborhood. $p = f(a,b)/Num(\mathbf{R}_k)$ expresses the frequency that the gray feature group $f(a, b)$ in $\mathbf{R}_k$.

For the blurry distortion, since its distortion appearance is similar to Gaussian blur (Gblur), a secondary Gblur plus structural similarity (SSIM) [11] is calculated as the feature:

$$f_{blu} = S\big(\mathbf{R}_k(i,j), \mathbf{R}''_k(i,j)\big) = \frac{2\mathbf{R}_k(i,j)\cdot\mathbf{R}''_k(i,j)+\varepsilon}{\mathbf{R}^2_k(i,j)+\mathbf{R}''^2_k(i,j)+\varepsilon} \tag{10}$$

where $\mathbf{R}''_k(i,j) = \mathbf{R}_k(i,j)\cdot w(i,j)$ is the secondary Gblur image; among this, the value of Gaussian weight $w(i,j) = \frac{1}{2\pi\sigma_b{}^2}\exp(-\frac{i^2+j^2}{2\sigma_b{}^2})$, and $\sigma_b = 1.5$.

For the stretching distortion, the horizontal pixel correlation is analyzed. Specifically, we detect the value equality of current pixel and its horizontal neighboring pixels. If the pixel satisfies the relevance condition, the numbers of pixels are counted:

$$f_{str} = \begin{cases} \frac{Num(x(i,j))}{R_k(i,j)} & \sum\limits_{l=1}^{2} \|x(i+l,j)-x(i,j)\|_1 = 0 \\ 0 & otherwise \end{cases} \tag{11}$$

where $x(i,j)$ denotes the pixel value at pixel coordinates (i,j).

On the other hand, the mixed multi-dimensional features are concentratively extracted by the Gaussian mixture model. The image is normalized:

$$\mathbf{R}'_k(i,j) = \frac{\mathbf{R}_k(i,j) - \mu_k(i,j)}{\sigma_k(i,j) + \varepsilon} \tag{12}$$

where $\mu_k(i,j)$ and $\sigma_k(i,j)$ are the mean and contrast value of $\mathbf{R}_k(i,j)$, which are calculated by a Gaussian kernel with a size of 3 × 3. $\mathbf{R}'_k(i,j)$ represents the mean subtracted contrast normalized (MSCN) coefficient.

Figure 4b plots a histogram of MSCN coefficients for an original image and different top-down geometric distorted versions to visualize how the MSCN coefficient distributions change as a function of geometric distortions. Compared with Figure 4a, the MSCN coefficients can explicitly distinguish different top-down distortions within a certain range in key region $\mathbf{R}_k$, which further verify the effectiveness of the above-mentioned key region extraction strategy.

In addition, the MSCN coefficients of adjacent pixels also have similar statistical characteristics. The MSCN coefficients of the present pixel and its four adjacent pixels (horizontal, vertical, main-diagonal, and secondary-diagonal) are calculated. Then, the Gaussian mixture model, which consists of generalized Gaussian distribution (GGD) and asymmetric GGD (AGGD), is used to extract mixed multi-dimensional features [50]. The mixed feature f_M is a set with $\mathbf{f}_{GGD}$ and $\mathbf{f}_{AGGD}$:

$$\mathbf{f}_{GGD(x;\alpha,\sigma^2)} = \frac{\alpha\sqrt{\Gamma(3/\alpha)}}{2\sigma\sqrt{\Gamma^3(1/\alpha)}} \exp(-\left(\frac{|x|}{\sigma} \cdot \sqrt{\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}})^{\alpha}\right) \tag{13}$$

$$\mathbf{f}_{AGGD(x;\beta,\sigma_l,\sigma_r)} = \begin{cases} \frac{\beta\sqrt{\Gamma(3/\beta)}}{(\sigma_l+\sigma_r)\sqrt{\Gamma^3(1/\beta)}} \exp(-\left(\frac{-x}{\sigma_l}\sqrt{\frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}})^{\beta}\right) & x < 0 \\ \frac{\beta\sqrt{\Gamma(3/\beta)}}{(\sigma_l+\sigma_r)\sqrt{\Gamma^3(1/\beta)}} \exp(-\left(\frac{-x}{\sigma_r}\sqrt{\frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}})^{\beta}\right) & x \geq 0 \end{cases} \tag{14}$$

where $\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}dt, x > 0$, α and σ^2 are the parameters of *GGD*, which reflect the shape and variance features of the current pixel distribution. β, σ_l, σ_r, and η are four parameters that affect *AGGD*. The *AGGD* gets the best performance when $\eta = (\sigma_r - \sigma_l)\frac{\Gamma(2/\beta)}{\sqrt{\Gamma(3/\beta)}\sqrt{\Gamma(1/\beta)}}$.

In addition, owing to the human perception for scenes being multi-scale [25], we build the feature extraction model on original and down-sampled images. Therefore, the Gaussian mixture model generates 36-dimensional features, which includes α and σ^2 in *GGD* and β, σ_l, σ_r, and η in *AGGD* with four adjacent directions and two image scales, i.e., f_M = [4$\mathbf{f}_{GGD}$, 32$\mathbf{f}_{AGGD}$].

3.3. Quality Regression

In this part, we use the regression function $H_m(\cdot)$ to map the extracted features to objective scores Q, which are expressed as:

$$Q = H_m(\mathbf{f}_{total}) \tag{15}$$

where $H_m(\cdot)$ is obtained by machine learning, and $\mathbf{f}_{total}$ = [f_h, f_{def}, f_{blu}, f_{str}, f_M] are the total feature vectors.

RF shows favorable accuracy and has few over-fitting problems in regression operator. Therefore, we use the RF to learn the function $H_m(\cdot)$ and achieve the predication of objective quality scores. In specific experiments, the 3D synthesized images in databases are divided into two non-overlapping parts randomly, 80% are used for training and the rest 20% are used for testing. The process of 'training-testing' is repeated for 1000 times, and the median performance is selected as the final model to eliminate performance bias.

4. Experimental Results and Analysis

This section mainly evaluates the performance of the visual-entropy-guided MLFA method. Firstly, we introduce the databases and performance evaluation criteria used in experiments. Secondly, the parameters are determined for achieving the best performance. Then, the performance of the visual-entropy-guided MLFA method is compared with other state-of-the-art metrics. Finally, generalization ability, impact of training percentages, multi-layer strategy, key region extraction strategy, and feature ablation experiments are implemented to prove the effectiveness of the visual-entropy-guided MLFA method.

4.1. Databases and Evaluation Criteria

Two databases, IRCCyN_IVC_DIBR_images database [36] and IETR DIBR image database [51], are used for the experiment in this study. The IRCCyN_IVC_DIBR_images database contains 96 images (84 3D synthesized images, 12 original images) and subjective quality scores. The database has three sequences, each sequence has four virtual views synthesized by a neighboring viewpoint using seven DIBR algorithms [37–42]. The IETR DIBR image database contains 150 images (140 3D synthesized images, 10 original images) and subjective quality scores. The database uses seven updated DIBR algorithms [52–58] to synthesize visual views and excludes some old-fashioned distortions (e.g., hole).

Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Correlation Coefficient (SRCC), and Root Mean Square Error (RMSE) are used to evaluate the difference between objective scores from metrics and subjective scores. The higher value of PLCC and SRCC, and the lower value of RMSE, mean that objective scores predicted by metrics are more similar to the subjective scores.

4.2. Parameters Determination

Three parameters, n, T_l, and T_r in the MLFA method, are determined in the IRCCyN_IVC_DIBR_images database.

Table 1 lists n and the corresponding discrimination performance of hole and non-hole regions. Specifically, we set $n = \{3, 5, 7, 9, 11\}$ and select 50 hole and non-hole regions respectively to calculate each median of distances. Next, the standard deviation is used to compare the stability of 50 regions. The smaller value of standard deviation means more stable performance. Then, T and computational time are calculated in different values of n. The experimental results show that, with the increase of n, the standard deviation of the hole increases slightly, but the standard deviation of non-hole increases dramatically. This unstable trend decreases the distinction between hole and non-hole areas, and eventually T cannot be obtained. In addition, the method also becomes time-consuming with the increase of n. In short, the experimental data verify that expanding the value of rectangle adjacent pixels will destruct the autocorrelation of transition statistics between hole and non-hole regions and increase computational complexity. Hence, in the MLFA method, we assign n as equal to 3.

Table 1. Influence of n on the performance of hole and non-hole distinction.

n	**3**	**5**	**7**	**9**	**11**
Standard deviation of hole	17.87	17.16	18.32	18.33	18.55
Standard deviation of non-hole	1.67	2.06	4.56	7.63	11.74
T	32	33	42	47	-
Computational time (s)	2.83	3.71	4.40	4.74	4.92

Figure 6 shows the impact of different width thresholds T_l and T_r on SRCC performance. The optimal thresholds are determined by comparing the SRCC values when the ranges of T_l and T_r are 0 to 10. Form 3D surface of SRCC performance, the T_l and T_r, with largest SRCC (SRCC = 0.8579), are 6% and 5%, respectively. Hence, we set T_l as 6%, and T_r as 5% in the proposed MLFA method.

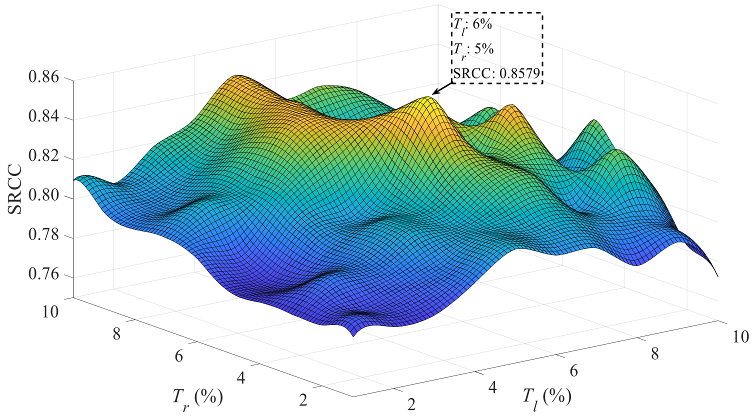


Figure 6. SRCC with different T_l and T_r.

4.3. Performance Comparison

Table 2 illustrates the PLCC, SRCC, and RMSE performance comparison of MLFA with state-of-the-art metrics on the IRCCyN_IVC_DIBR_images database in which the best performance is highlighted with bold font. Specifically, the PSNR and SSIM [11] are traditional IQA metrics. The VSQA [16], 3DSwIM [17], ST-SIAQ [18], EM-IQA [19], MW-PSNR [20], MP-PSNR [21], SC-IQA [22], and IDEA [23] are FR IQA metrics for 3D synthesized images. The APT [24], MNSS [25], OUT [26], NR-MWT [28], SET [30], NIQSV [31], NIQSV+ [32], CLGM [33], GANs-NRM [34], and Wang [35] are NR metrics designed for 3D synthesized images. In the experimental results, we can obtain three potential conclusions:

(1) The traditional metrics, like PSNR and SSIM, are not effective for 3D synthesized images. The performance of PSNR and SSIM is poor because they have not been conceived for dealing with the local specificity of geometric distortions (e.g., the PLCC is less than 0.5).

(2) The performance of metrics designed for 3D synthesized images is better than traditional metrics, but not sufficient. The metrics, VSQA, 3DSwIM, ST-SIAQ, EM-IQA, and NIQSV, are mainly designed for the object shifting and blurry distortions (parts of the top-down layer). The metrics, MW-PSNR, MP-PSNR, APT, MNSS, and OUT, are mainly sensitive to hole distortion. The above-mentioned metrics ignore the diversity of geometric distortions. Among them, the MNSS metric shows the best performance, and PLCC, SRCC, and RMSE are 0.7700, 0.7850, and 0.4120. A few metrics consider multiple distortions, such as SC-IQA, IDEA, NR-MWT, SET, NIQSV+, and CLGM. However, the weak geometric distortions are inadequately and ambiguously classified, and merely measured via decentralized attention. In addition, only a few metrics (e.g., IDEA) emphasize the utilization of local distortion distribution characteristics. These limitations lead these metrics to fail to effectively estimate weak distortions. Even for SET, the best among these metrics, the corresponding PLCC, SRCC, and RMSE are 0.8586, 0.8109, and 0.3015, and can be further improved. The performance of deep-learning-based metrics, such as GANs-NRM and Wang, is also unsatisfactory due to the limitation of insufficient training samples.

(3) The proposed method MLFA is superior to the state-of-the-art metrics, and PLCC, SRCC, and RMSE are 0.8757, 0.8579, and 0.4106. It affirms the effectiveness of MLFA method for 3D synthesized images.

Table 2. Performance comparison of the proposed method with state-of-the-art metrics on the IRCCyN_IVC_DIBR images database.

Category	Distortion Type	Metric	PLCC	SRCC	RMSE
FR	2D traditional distortion	PSNR	0.4515	0.4589	0.5527
		SSIM [11]	0.4850	0.4368	0.5823
FR	3D synthesized distortion	VSQA [16]	0.5742	0.5233	0.5451
		3DSwIM [17]	0.6584	0.6156	0.5011
		ST-SIAQ [18]	0.6914	0.6746	0.4812
		EM-IQA [19]	0.7430	0.6282	0.4455
		MW-PSNR [20]	0.5622	0.5757	0.5506
		MP-PSNR [21]	0.6174	0.6227	0.5238
		SC-IQA [22]	0.8496	0.7640	**0.3511**
		IDEA [23]	0.7796	0.6652	0.3533
NR	3D synthesized distortion	APT [24]	0.7307	0.7157	0.4546
		MNSS [25]	0.7700	0.7850	0.4120
		OUT [26]	0.7243	0.7010	0.4591
		NR-MWT [28]	0.7343	0.5169	0.4520
		SET [30]	0.8586	0.8109	0.3015
		NIQSV [31]	0.6346	0.6167	0.5146
		NIQSV+ [32]	0.7114	0.6668	0.4679
		CLGM [33]	0.6750	0.6528	0.4620
		GANs-NRM [34]	0.8262	0.8072	0.3861
		Wang [35]	0.8112	0.7520	0.3820
		MLFA	**0.8757**	**0.8579**	0.4106

Figure 7 shows the scatter plots of subjective DMOS and objective scores in SSIM, MP-PSNR, APT, MNSS, NIQSV+, and MLFA on the IRCCyN_IVC_DIBR_images database. The points of the MLFA method aggregate on the fitting line. By contrast, the scattered plots of the comparative metrics present vertical point distribution, i.e., objective scores of the vertical distributed points are similar, while the subjective scores are different. By validating the corresponding image of each point, we found that the comparative metrics can roughly distinguish specific distortions but fail to effectively estimate weak geometric distortions. For instance, the NIQSV+ metric can roughly distinguish three kinds of distortions, hole, stretching, and blurry distortion. Correspondingly, the scatter points present three clusters with different objective scores. However, due to the insufficiency of mixed weak distortions estimation, the corresponding objective scores of scatter points are close in each cluster, as shown in the subfigure of Figure 7. Hence, the objective scores calculated by the MLFA method can achieve higher consistency with human subjective perception.

Table 3 shows PLCC, SRCC, and RMSE of MLFA with state-of-the-art IQA metrics on the IETR DIBR image database, where the best results are highlighted in boldface. One can see that the performance of some representative metrics, such as NIQSV+ and CLGM, is poor. These metrics mainly measure limited and salient distortion types via decentralized attention. Thus, the defects, i.e., poor robustness for update 3D synthesized scenes, are easily exposed on the database without old-fashioned distortions. Among these metrics, the performance of SC-IQA metric is the best. However, its PLCC, SRCC, and RMSE are only 0.6856, 0.6423, and 0.1805. Comparatively, the MLFA method obtains the best performance on this database, i.e., PLCC, SRCC, and RMSE are 0.7378, 0.7036, and 0.1899. It validates that the MLFA method is effective and robust for various distorted scenes, especially including weak geometric distortions.

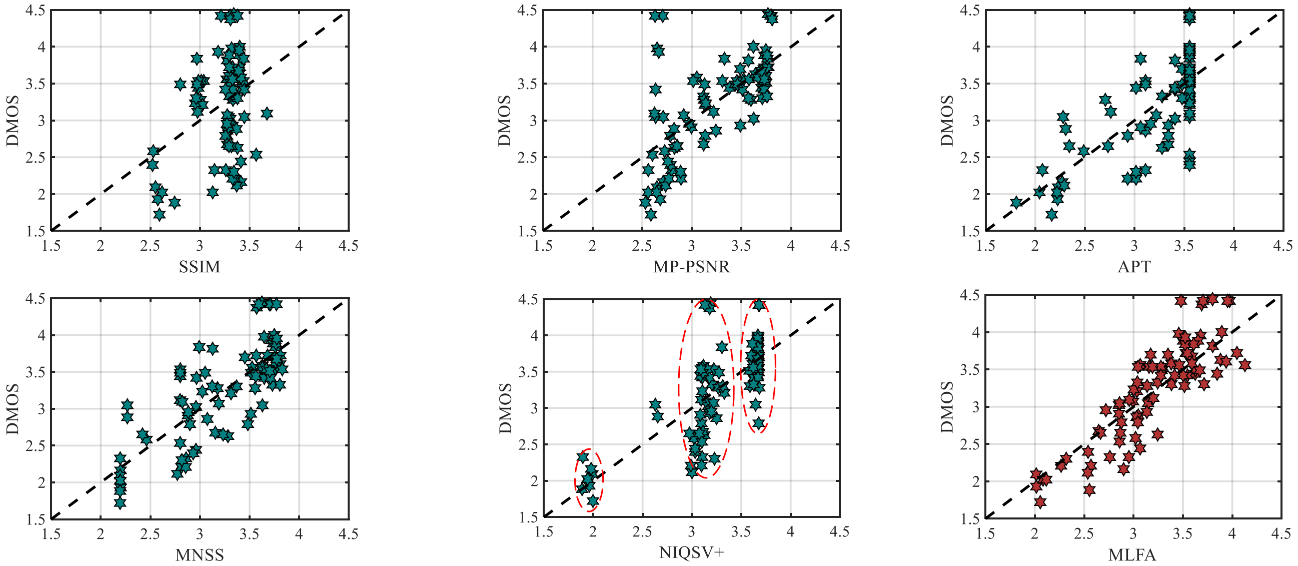


Figure 7. Scatter plots of subjective DMOS and objective scores on the IRCCyN_IVC_DIBR_images database.

Table 3. Performance comparison of the proposed method with state-of-the-art metrics on the IETR DIBR image database.

Category	Distortion Type	Metric	PLCC	SRCC	RMSE
FR	2D traditional distortion	PSNR	0.6012	0.5356	0.1985
		SSIM [11]	0.4016	0.2395	0.2275
FR	3D synthesized distortion	VSQA [16]	0.5576	0.4719	0.2062
		ST-SIAQ [18]	0.3345	0.4232	0.2336
		EM-IQA [19]	0.5627	0.5670	0.2020
		MW-PSNR [20]	0.5301	0.4845	0.2106
		MP-PSNR [21]	0.5753	0.5507	0.2032
		SC-IQA [22]	0.6856	0.6423	**0.1805**
NR	3D synthesized distortion	APT [24]	0.4225	0.4187	0.2252
		MNSS [25]	0.3387	0.2281	0.2333
		OUT [26]	0.2007	0.1924	0.2429
		NR-MWT [27]	0.4769	0.4567	0.2179
		NIQSV [31]	0.1759	0.1473	0.2446
		NIQSV+ [32]	0.2095	0.2190	0.2429
		CLGM [33]	0.1146	0.0860	0.2463
		MLFA	**0.7378**	**0.7036**	0.1899

4.4. Generalization Ability

As a train-test-based quality model, the generalization ability is a persuasive robustness criterion. Therefore, we verify the generalization ability of our visual-entropy-guided MLFA method through a cross-experiment, where the best results are also marked in bold. Specifically, (1) the IETR DIBR image database is used when training the model, and the IRCCyN_IVC_DIBR_images database is used to test. (2) The IRCCyN_IVC_DIBR_images database is adopted to train and the IETR DIBR image database is used to test. Table 4 shows the performance comparison of our MLFA method and the other NR state-of-the-art synthesized IQA metrics. One can see that the proposed MLFA method acquires the best performance among these metrics. In addition, the performance of training on the IETR DIBR image database and testing on the IRCCyN_IVC_DIBR_images database is better than training on the IRCCyN_IVC_DIBR_images database and testing on the IETR DIBR image database. This is because the distortions of IRCCyN_IVC_DIBR_images database are old-fashioned, while the distortions of IETR DIBR image database are upgraded and more meticulous.

Table 4. Cross-validation of the proposed MLFA method and the NR state-of-the-art metrics on IETR DIBR image database and IRCCyN_IVC_DIBR_images database.

Training Database	Testing Database	Method	PLCC	SRCC	RMSE
IETR DIBR image	IRCCyN_IVC_DIBR_images	APT	0.6745	0.5817	0.4916
		MNSS	0.6539	0.6147	0.5037
		NIQSV	0.4989	0.0889	0.5494
		NIQSV+	0.5921	0.2680	0.5365
		MLFA	**0.8645**	**0.8562**	**0.3945**
IRCCyN_IVC_DIBR_images	IETR DIBR image	APT	0.3838	0.2198	0.2249
		MNSS	0.2829	0.2196	0.2335
		NIQSV	0.1216	0.0839	0.2416
		NIQSV+	0.0292	0.0569	0.2433
		MLFA	**0.7046**	**0.6720**	**0.2181**

4.5. Impact of Training Percentages

To research how the amount of training data affects the performance of MLFA method, we execute the experiment via adopting different proportions of two DIBR image databases with 10% steps to train the model. Mainly, the image percentages of database used to train the model are set to five levels, i.e., 90%, 80%, 70%, 60% and 50%. All of the training–testing processes are operated 1000 times to get the median value, and the results are shown in Table 5. With the cut back of training data, the performance of the model gradually decreases. However, even with the lowest 50% training in the IRCCyN_IVC_DIBR_images database, we still get relatively good performance compared to most state-of-the-art synthesized IQA metrics, i.e., PLCC reaches 0.83. Moreover, on IETR DIBR image database, with only 50% training, our method outperforms the state-of-the-art metrics. These experiment results verify that our proposed MLFA method can still achieve better performance even if it uses less data for training.

Table 5. Performances of the proposed MLFA method with different training percentages.

Database	Training–Testing	PLCC	SRCC	RMSE
IRCCyN_IVC_DIBR_images	90–10%	0.8895	0.8585	0.2967
	80–20%	0.8757	0.8579	0.3106
	70–30%	0.8620	0.8330	0.3871
	60–40%	0.8467	0.8073	0.4339
	50–50%	0.8303	0.7970	0.5010
IETR DIBR image	90–10%	0.7473	0.7158	0.1642
	80–20%	0.7378	0.7036	0.1899
	70–30%	0.7180	0.6845	0.1928
	60–40%	0.7055	0.6644	0.2027
	50–50%	0.6899	0.6473	0.2092

4.6. Performance Analysis of a Multi-Layer Strategy

To illustrate the superiority of the visual-entropy-guided multi-layer strategy proposed in our method, we conduct a comparative experiment S1 with a single-layer strategy. In S1, the key region of the test image is firstly extracted. Then, the features of hole, deforming, blurry, and stretching are measured at the same level. Here, the MLFA method based on multi-layer strategy is denoted as S2.

Table 6 lists PLCC, SRCC, and RMSE of S1 and S2 on IRCCyN_IVC_DIBR_images and IETR DIBR image databases. Both S1 and S2 show good performance, which validate the integral effectiveness of feature extraction algorithms in our method. However, the performance of S1 is not poor but worse than S2, e.g., the PLCC value of S2 is about 0.02 higher than S1 on two databases. It suggests that putting the distortions of different visual stimuli at the same level will affect the accuracy of feature extraction to a certain extent.

Therefore, the multi-layer strategy is conducive to further improving the performance of 3D synthesized IQA metrics.

Table 6. Performance comparison of S1 and S2 on two databases.

Scheme	**IRCCyN_IVC_DIBR_Images**			**IETR DIBR Image**		
	PLCC	**SRCC**	**RMSE**	**PLCC**	**SRCC**	**RMSE**
S1	0.8558	0.8004	0.4269	0.7133	0.6861	0.2061
S2	0.8757	0.8579	0.4106	0.7378	0.7036	0.1899

4.7. Performance Analysis of Key Region Extraction

Table 5 shows the PLCC performance of f_h, f_{def}, f_{blu}, f_{str}, and f_M with or without key region extraction. Specifically, we perform a comparative experiment denoted by S3. In S3, the process of key region extraction is canceled. Meanwhile, the scheme with key region extraction (i.e., MLFA method) is named S4. From Table 7, the PLCC results of various features in S3 and S4 on different databases can reflect the following two conclusions:

(1) S3 and S4 have similar PLCC performance on the bottom-up layer (i.e., f_h). However, the performance of S3 is reduced on the top-down layer, especially for f_M. Theoretically, most regions of the 3D synthesized images are not geometrically distorted. In S3, the features are extracted throughout the entire image, and the local geometric distortions are too subtle to be extracted. However, S4 adopts key region extraction, which highlights the regions of weak geometric distortion. Hence, the interference of most non-geometric distortion regions is effectively eliminated. The experimental data indeed verifies this theoretical explanation, i.e., the PLCC of S4 is nearly twice as high as S3 in f_M on two databases.

(2) Different from f_M, the PLCC performance of f_{def}, f_{blu}, and f_{str} on the top-down layer is slightly affected by key region extraction. f_M is a multi-dimensional feature and is obtained by concentrated attention. By contrast, f_{def}, f_{blu}, and f_{str}, are single-dimensional features, and extracted from corresponding distortions via decentralized attention. Thus, the latter features are more distortion-specific, and insensitive to the regional interference in different scenes. The analysis is validated by the experimental results, which the PLCC of S3 slightly decreases within 0.04 compared to S4 in terms of f_{def}, f_{blu}, and f_{str}.

Table 7. PLCC comparison with or without key region extraction.

Database	**Scheme**	f_h	f_{def}	f_{blu}	f_{str}	f_M
IRCCyN_IVC_DIBR_images	S3	0.5409	0.5760	0.6248	0.3956	0.3535
	S4	0.5416	0.6108	0.6358	0.4331	0.6906
IETR DIBR image	S3	0.4278	0.2972	0.4092	0.3367	0.2094
	S4	0.4271	0.3365	0.4165	0.3681	0.4544

In short, the experimental results on both two databases verify the effectiveness of the key region extraction on the top-down layer. In particular, the strategy of key region extraction plays a decisive role in the performance of f_M, which means that the key region extraction is close relative to concentrated attention, and is a potential reason for the superiority of the overall model.

4.8. Feature Ablation Experiments

To analyze the contribution of the feature component, we also perform feature ablation experiments on the IRCCyN_IVC_DIBR_images database in which f_h, f_{def}, f_{blu}, f_{str}, and f_M are permuted and combined into 17 models. The experimental results are listed in Table 8, and the best results are marked in bold. From extensive experimental results, two arguments can be made.

(1) M1–M5, composed of one feature, have poor performance, i.e., PLCC is below 0.7 roughly. In M6–M11, the feature components reach three, and the PLCC ranges from 0.7465 to 0.8174. For M12–M16, the feature components are increased to four, and the PLCC is further improved and stabilized in 0.8294 to 0.8538. In M17, PLCC is the best and equals 0.8757, when the feature components are five. The experimental data show that the performance increases in steps and gradually stabilizes with the addition of feature components. Hence, each feature is an essential part of the MLFA method and can significantly increase the accuracy and stability of the IQA model.

(2) Among these models, M12 and M17 are emphatically compared. In M12, the features are merely obtained by decentralized attention (as traditional distortion-classification-based 3D synthesized IQA metrics do). In M17, the features are acquired via feature integration theory, i.e., the interaction of decentralized attention and concentrated attention (as the MLFA method does). Obviously, the performance of M17 is better than M12, i.e., PLCC and SRCC are 0.0353 and 0.0988 higher than M12, and RMSE is 0.0145 lower than M12. The performance comparison demonstrates that the MLFA method, which uses the strategy of feature integration theory, achieves higher feature utilization and improves the consistency with the subjective scores.

Table 8. Performance of different feature components on the IRCCyN_IVC_DIBR images database.

Models	**Features**					**IRCCyN_IVC_DIBR_Images**		
	f_h	f_{def}	f_{blu}	f_{str}	f_M	**PLCC**	**SRCC**	**RMSE**
M1	√					0.5416	0.3670	0.7050
M2		√				0.6108	0.3543	0.6248
M3			√			0.6358	0.5375	0.6420
M4				√		0.4331	0.3829	0.7605
M5					√	0.6906	0.5639	0.6022
M6	√	√	√			0.8174	0.7558	0.4518
M7	√	√		√		0.7465	0.6680	0.5366
M8	√	√			√	0.8029	0.7301	0.4856
M9	√		√	√		0.8103	0.7085	0.4698
M10	√		√		√	0.7895	0.7088	0.4850
M11	√			√	√	0.7781	0.6887	0.5049
M12	√	√	√	√		0.8404	0.7591	0.4251
M13	√	√	√		√	0.8378	0.7735	0.4353
M14	√	√		√	√	0.8294	0.7511	0.4537
M15	√		√	√	√	0.8373	0.7598	0.4497
M16		√	√	√	√	0.8538	0.7997	0.4146
M17	√	√	√	√	√	**0.8757**	**0.8579**	**0.4106**

Figure 8 shows four images in the IRCCyN_IVC_DIBR_images database, which includes different geometric distortions. We extracted their features of the single-dimensional channel (i.e., $f_h, f_{def}, f_{blu}, f_{str}$) separately and listed the results. It can be seen that the proposed feature extraction method acquires the largest value in their corresponding images, as shown in bold. Moreover, if there are several kinds of distortions in an image, the MLFA model can still work very well. For example, the 'blurry' images also include some stretching distortion. The value of stretching feature is extracted as 0.0422 but less than 0.0553 of the 'stretching' image. This further verifies that the relationship between the proposed feature extraction model and their corresponding distortion is highly consistent, and the feature extraction model can reflect image distortion levels pretty well.

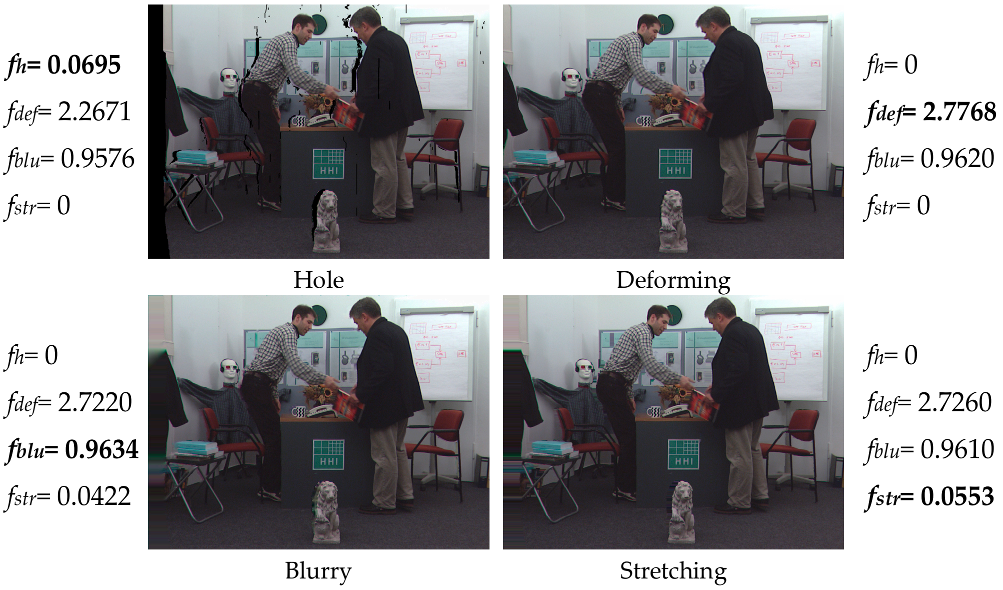


Figure 8. One-dimensional feature extraction of four images with different geometric distortions.

5. Conclusions

In this paper, we have proposed an NR IQA metric based on visual-entropy-guided MLFA for 3D synthesized images. Taking into account the stimulation of energy entropy to the human visual attention mechanism, different geometric distortions are divided into bottom-up layer and top-down layer. The ratio of 0-value pixels and the transition threshold are combined to calculate the hole feature on the bottom-up layer. In the meantime, based on key distorted region extraction, we adopt the interaction of decentralized and concentrated attentions to obtain the features of insignificant geometric distortions on the top-town layer. The final objective scores are obtained by regressing the features on multiple visual attention layers through RF. Extensive experiments have demonstrated that, compared with classical and state-of-the-art metrics, our MLFA method achieves better performance both on two public synthesized image databases and has a higher consistency with human subjective perception.

Author Contributions: C.J. designed and completed the algorithm and drafted the manuscript. Z.P. co-designed the algorithm and polished the manuscript. W.Z. proofread the manuscript. F.C. analyzed the experimental results. G.J. co-designed the algorithm. M.Y. polished the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant Nos. 61771269, 61620106012, 61871247, 62071266, and 61931022, the Natural Science Foundation of Zhejiang Province under Grant No. LY20F010005, the Natural Science Foundation of Ningbo under Grant No. 2019A610107, the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant No. KJZD-K202001105, and the Scientific Research Foundation of Chongqing University of Technology.

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

References

1. Buisine, J.; Bigand, A.; Synave, R.; Delepoulle, S.; Renaud, C. Stopping Criterion during Rendering of Computer-Generated Images Based on SVD-Entropy. *Entropy* **2021**, *23*, 75. [CrossRef] [PubMed]
2. Qiao, Y.; Jiao, L.; Yang, S.; Hou, B.; Feng, J. Color Correction and Depth-Based Hierarchical Hole Filling in Free Viewpoint Generation. *IEEE Trans. Broadcast.* **2019**, *65*, 294–307. [CrossRef]
3. Lin, Y.; Yu, M.; Chen, K.; Jiang, G.; Chen, F.; Peng, Z. Blind Mesh Assessment Based on Graph Spectral Entropy and Spatial Features. *Entropy* **2020**, *22*, 190. [CrossRef] [PubMed]
4. Li, B.; Tian, M.; Zhang, W.; Yao, H.; Wang, X. Learning to Predict the Quality of Distorted-then-Compressed Images via a Deep Neural Network. *J. Vis. Commun. Image Represent.* **2021**, *76*, 103004. [CrossRef]

5. Cui, X.; Peng, Z.; Jiang, G.; Chen, F.; Yu, M. Perceptual Video Coding Scheme Using Just Noticeable Distortion Model Based on Entropy Filter. *Entropy* **2019**, *21*, 1095. [CrossRef]
6. Deng, C.; Wang, S.; Bovik, A.C.; Huang, G.; Zhao, B. Blind Noisy Image Quality Assessment Using Sub-Band Kurtosis. *IEEE Trans. Cybern.* **2020**, *50*, 1146–1156. [CrossRef]
7. Guan, X.; He, L.; Li, M.; Li, F. Entropy based Data Expansion Method for Blind Image Quality Assessment. *Entropy* **2020**, *22*, 60. [CrossRef]
8. Zhan, Y.; Zhang, R. No-Reference JPEG Image Quality Assessment Based on Blockiness and Luminance Change. *IEEE Signal Process. Lett.* **2017**, *24*, 760–764. [CrossRef]
9. Soltani, M.; Pourahmadi, V.; Mirzaei, A.; Sheikhzadeh, H. Deep Learning-Based Channel Estimation. *IEEE Commun. Lett.* **2019**, *23*, 652–655. [CrossRef]
10. Sheikh, H.R.; Sabir, M.F.; Bovik, A.C. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.* **2006**, *15*, 3440–3451. [CrossRef]
11. 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] [PubMed]
12. 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]
13. Saha, A.; Wu, Q. Full-reference image quality assessment by combining global and local distortion measures. *Signal Process.* **2016**, *128*, 186–197. [CrossRef]
14. Zhang, Y.; Mou, X.; Chandler, D.M. Learning No-Reference Quality Assessment of Multiply and Singly Distorted Images with Big Data. *IEEE Trans. Image Process.* **2020**, *29*, 2676–2691. [CrossRef] [PubMed]
15. Bosc, E.; Callet, P.L.; Morin, L.; Pressigout, M. An edge-based structural distortion indicator for the quality assessment of 3D synthesized views. In Proceedings of the Picture Coding Symposium (PCS), Krakow, Poland, 7–9 May 2012; pp. 249–252.
16. Conze, P.-H.; Robert, P.; Morin, L. Objective view synthesis quality assessment. In Proceedings of the International Society for Optical Engineering (SPIE), Burlingame, CA, USA, 27 February 2012; pp. 8256–8288.
17. Battisti, F.; Bosc, E.; Carli, M.; Callet, P.L.; Perugia, S. Objective image quality assessment of 3D synthesized views. *Signal Process. Image Commun.* **2015**, *30*, 78–88. [CrossRef]
18. Ling, S.; Callet, P.L. Image quality assessment for free viewpoint video based on mid-level contours feature. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 79–84.
19. Ling, S.; Callet, P.L. Image quality assessment for DIBR synthesized views using elastic metric. In Proceedings of the 17th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1157–1163.
20. Sandić-Stanković, D.; Kukolj, D.; Callet, P.L. DIBR synthesized image quality assessment based on morphological wavelets. In Proceedings of the 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), Pylos-Nestoras, Greece, 26–29 May 2015; pp. 1–6.
21. Sandić-Stanković, D.; Kukolj, D.; Callet, P.L. Multi-scale synthesized view assessment based on morphological pyramids. *J. Elect. Eng.* **2016**, *67*, 9–11. [CrossRef]
22. Tian, S.; Zhang, L.; Morin, L.; Déforges, O. SC-IQA: Shift compensation based image quality assessment for DIBR-synthesized views. In Proceedings of the IEEE International Conference on Visual Communication and Image Processing (VCIP), Taiwan, China, 7–10 October 2018; pp. 1–4.
23. 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]
24. Gu, K.; Jakhetiya, V.; Qiao, J.; 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]
25. Gu, K.; Qiao, J.; Lee, S.; Liu, H.; Lin, W.; Callet, P.L. Multiscale Natural Scene Statistical Analysis for No-Reference Quality Evaluation of DIBR-Synthesized Views. *IEEE Trans. Broadcast.* **2020**, *66*, 127–139. [CrossRef]
26. Jakhetiya, V.; Gu, K.; Singhal, T.; Guntuku, S.C.; Xia, Z.; Lin, W. A highly efficient blind image quality assessment metric of 3d synthesized images using outlier detection. *IEEE Trans. Ind. Inform.* **2019**, *15*, 4120–4128. [CrossRef]
27. Jakhetiya, V.; Gu, K.; Jaiswal, S.P.; Singhal, T.; Xia, Z. Kernel-Ridge Regression-Based Quality Measure and Enhancement of Three-Dimensional-Synthesized Images. *IEEE Trans. Ind. Electron.* **2021**, *68*, 423–433. [CrossRef]
28. Sandić-Stanković, D.D.; Kukolj, D.D.; Callet, P.L. Fast Blind Quality Assessment of DIBR-Synthesized Video Based on High-High Wavelet Subband. *IEEE Trans. Image Process.* **2019**, *28*, 5524–5536. [CrossRef]
29. 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.* **2020**, *29*, 1802–1814. [CrossRef]
30. Zhou, Y.; Li, L.; Wang, S.; Wu, J.; Fang, Y.; 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. Tian, S.; Zhang, L.; Morin, L.; Déforges, 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), Los Angeles, CA, USA, 5–9 March 2017; pp. 1248–1252.
32. 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]

33. Yue, G.; Hou, C.; Gu, K.; Zhou, T.; Zhai, G. Combining Local and Global Measures for DIBR-Synthesized Image Quality Evaluation. *IEEE Trans. Image Process.* **2019**, *28*, 2075–2088. [CrossRef]
34. Ling, S.; Li, J.; Wang, J.; Callet, P.L. Gans-nqm: A generative adversarial networks based no reference quality assessment metric for RGB-D synthesized views. *arXiv* **2019**, arXiv:1903.12088.
35. Wang, X.; Liang, X.; Yang, B.; Li, F. No-reference synthetic image quality assessment with convolutional neural network and local image saliency. *Comput. Vis. Media* **2019**, *5*, 193–208. [CrossRef]
36. Bosc, E.; Pepion, R.; Callet, P.L.; 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]
37. Fehn, C. Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV. In Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE), San Jose, CA, USA, 21 May 2004; pp. 93–104.
38. Ndjiki-Nya, P.; Koppel, M.; Doshkov, D.; Lakshman, H.; Merkle, P.; Muller, K.; Wiegand, T. Depth Image-Based Rendering with Advanced Texture Synthesis. In Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME), Suntec City, Singapore, 19–23 July 2010; pp. 424–429.
39. Köppel, M.; Ndjiki-Nya, P.; Doshkov, D.; Lakshman, H.; Merkle, P.; Müller, K.; Wiegand, T. Temporally consistent handling of disocclusions with texture synthesis for depth-image-based rendering. In Proceedings of the 2010 IEEE International Conference on Image Processing (ICIP), Hong Kong, China, 26–29 September 2010; pp. 1809–1812.
40. Telea, A. An image inpainting technique based on the fast marching method. *J. Graph. Tools* **2004**, *9*, 23–34. [CrossRef]
41. Mori, Y.; Fukushima, N.; Yendo, T.; Fujii, T.; Tanimoto, M. View generation with 3D warping using depth information for FTV. *Signal Process. Image Commun.* **2009**, *24*, 65–72. [CrossRef]
42. Müller, K.; Smolic, A.; Dix, K.; Merkle, P.; Kauff, P.; Wiegand, T. View synthesis for advanced 3D video systems. *EURASIP J. Image Video Process.* **2009**, *2008*, 1–11. [CrossRef]
43. Miller, E.K. The prefrontal cortex and cognitive control. *Nat. Rev. Neurosci.* **2000**, *1*, 59–65. [CrossRef]
44. Crick, F.; Koch, C. Constraints on cortical and thalamic projections: The no-strong-loops hypothesis. *Nature* **1998**, *391*, 245–250. [CrossRef]
45. Friston, K. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **2010**, *11*, 127–138. [CrossRef] [PubMed]
46. Varga, D. No-Reference Image Quality Assessment Based on the Fusion of Statistical and Perceptual Features. *J. Imaging* **2020**, *6*, 75. [CrossRef]
47. Itti, L.; Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2001**, *2*, 194–203. [CrossRef] [PubMed]
48. Treisman, A.M.; Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **1980**, *12*, 97–136. [CrossRef]
49. Xu, L.; Yan, Q.; Xia, Y.; Jia, J. Structure Extraction from Texture via Relative Total Variation. *ACM Trans. Graph.* **2012**, *31*, 139:1–139:10. [CrossRef]
50. Cui, Y. No-Reference Image Quality Assessment Based on Dual-Domain Feature Fusion. *Entropy* **2020**, *22*, 344. [CrossRef]
51. Tian, S.; Zhang, L.; Morin, L.; Deforges, O. A benchmark of DIBR synthesized view quality assessment metrics on a new database for immersive media applications. *IEEE Trans. Multimed.* **2018**, *21*, 1235–1247. [CrossRef]
52. Tanimoto, M.; Fujii, T.; Suzuki, K.; Fukushima, N.; Mori, Y. Reference softwares for depth estimation and view synthesis. In *ISO/IEC JTC1/SC29/WG11*; MPEG: Archamps, France, 2008; doc. M15377.
53. Zhu, C.; Li, S. Depth image based view synthesis: New insights and perspectives on hole generation and filling. *IEEE Trans. Broadcast.* **2016**, *62*, 82–93. [CrossRef]
54. Criminisi, A.; Perez, P.; Toyama, K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process.* **2004**, *13*, 1200–1212. [CrossRef] [PubMed]
55. Luo, G.; Zhu, Y.; Li, Z.; Zhang, L. A hole filling approach based on background reconstruction for view synthesis in 3-D video. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, CA, USA, 27–30 June 2016; pp. 1781–1789.
56. Solh, M.; AlRegib, G. Hierarchical hole-filling for depth-based view synthesis in FTV and 3-D video. *IEEE J. Sel. Top. Signal Process.* **2012**, *6*, 495–504. [CrossRef]
57. Jantet, V.; Guillemot, C.; Morin, L. Object-based layered depth images for improved virtual view synthesis in rate-constrained context. In Proceedings of the 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011; pp. 125–128.
58. Ahn, I.; Kim, C. A novel depth-based virtual view synthesis method for free viewpoint video. *IEEE Trans. Broadcast.* **2013**, *59*, 614–6268. [CrossRef]