Multimedia Tools and Applications (2021) 80:11557–11581
https://doi.org/10.1007/s11042-020-10196-x

Disocclusion-type aware hole filling method for view synthesis

Xiaodong Chen[1,2] · Haitao Liang[1,2] · Huaiyuan Xu[1,2] · Siyu Ren[1,2] · Huaiyu Cai[1,2] · Yi Wang[1,2]

Received: 13 November 2019 / Revised: 8 September 2020 / Accepted: 24 November 2020 /
Published online: 6 January 2021

Abstract

Depth-image-based rendering is an efficient way to produce content for 3D video and free viewpoint video. However, since the background that is occluded by the foreground objects in the reference view may become visible in the synthesized view, disocclusions are produced. In this paper, a disocclusion-type aware hole filling method is proposed for disocclusion handling. Disocclusions are divided into two types based on the depth value of their boundary pixels: foreground-background (FG-BG) disocclusion and background-background (BG-BG) disocclusion. For FG-BG disocclusion, the depth values of the associated pixels are optimized in the reference image to ensure the removal of ghosts and adaptively divide the disocclusion into some small holes. For BG-BG disocclusion, a foreground removal method is applied to remove the corresponding foreground objects. The removed regions are filled with the surrounding background textures so that the BG-BG disocclusion in the synthesized image can be eliminated. Experimental results indicate that the proposed method outperforms the other methods in the objective and subjective evaluations.

Keywords Depth-image-based rendering · 3D video · Single view synthesis · Disocclusion filling · Foreground removal

1 Introduction

3D video and free viewpoint video (FVV) have generated great interest in the past few years due to their realistic and interactive experience [38]. Compared with the traditional 2D video,

✉ Yi Wang
koala_wy@tju.edu.cn

[1] School of Precision Instrument & Opto-electronics Engineering, Tianjin University, Tianjin, China

[2] Key Laboratory of Opto-electronics Information Technology, Ministry of Education, Tianjin University, Tianjin 300072, China

the addition of depth perception helps viewers to better distinguish the occlusion relationship between objects in the scene. Due to the limitation of transmission bandwidth, it is not realistic to generate FVV through a large number of video capture devices [20, 23]. One practical approach is to generate a series of virtual views at the receiving end based on one or more key reference views. In this case, video-plus-depth or multiview video-plus-depth is the common format for transmitting 3D video, because a depth image requires less bandwidth than a color image of the same size [41]. In view synthesis techniques, depth-image-based rendering (DIBR) is widely used to produce content for 3D video and FVV, which only requires a single reference view and its associated depth image [10]. In the DIBR, all pixels in the reference image are projected to the world coordinate system based on their respective depth values. Then the virtual image is synthesized by projecting these points into the target image plane. This process is called 3D warping [25].

In fact, during the 3D warping process, a critical problem arises along with the movement of the view is that the background occluded by the foreground objects in the reference view may become visible in the synthesized view. As no pixels from the reference image are projected into these regions, holes are created, called disocclusions (shown in Fig. 1) [7]. Figure 1(a) and 1(b) are the synthesized results for small baseline and large baseline, respectively. It can be seen that the disocclusion becomes larger as the baseline increases (marked in white). In addition, since the depth image can reflect the distance from the object to the camera [14], the larger depth discontinuity between adjacent pixels at the foreground and background junctions, the greater their distance in the virtual image, resulting in larger disocclusion. Therefore, the reasonable filling of these disocclusions is critical to the quality of the virtual image. In addition to disocclusions, there are other types of artifacts in the virtual image, such as ghosts, cracks, and out-of-field regions (OFRs) [26], as shown in Fig. 1. Ghosts mean that the pixels on the foreground boundary are assigned depth values of adjacent backgrounds during depth acquisition, which are incorrectly projected into the background region and blended with the background texture. Cracks are usually 1–2 pixels wide, which are caused by the rounding of target pixel position. Moreover, In the case where the virtual view exceeds the imaging range of reference view, OFRs appear at the edges of the virtual image. Since these artifacts are difficult to avoid in the original 3D warping, especially for disocclusions, how to fill them in a visually plausible way is a challenging task.

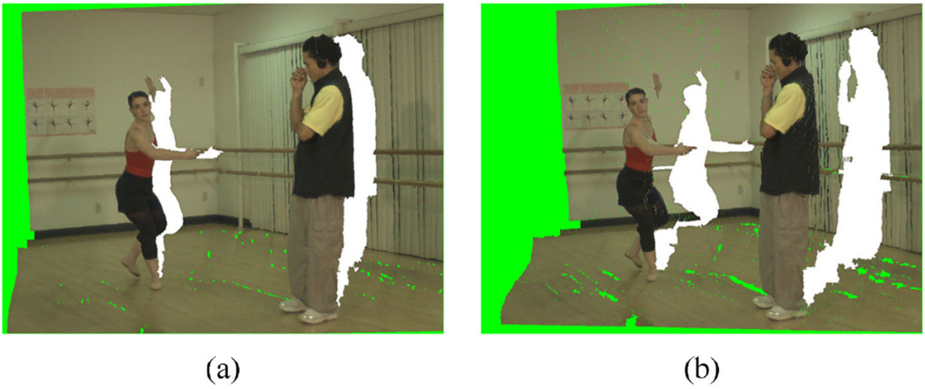

Fig. 1 Original 3D warping result. **a** Small baseline, and **b** large baseline

In this paper, to fill the disocclusions properly and improve the visual quality of virtual view, a disocclusion-type aware hole filling method is proposed. Our main contributions are as follows: 1) We divide the disocclusions into two types based on the depth information of their edge pixels: foreground-background (FG-BG) disocclusion and background-background (BG-BG) disocclusion. 2) For the former, an adaptive depth image preprocessing method is introduced to decompose the disocclusions so that they can be easily filled in the virtual view. For the latter, the corresponding foreground objects are removed, so that the disocclusion can be filled with the reliable background texture. 3) A modified depth-based inpainting method is proposed in the filling process to increase the credibility of the results. The rest of the paper is organized as follows. The related work is introduced in Section 2. The proposed disocclusion handling approach is described in Section 3. The experimental results are presented in Section 4. Finally, Section 5 discusses the limitations and concludes the paper.

2 Related work

Under the condition of single view rendering, in general, the disocclusion handling methods can be divided into two categories. One is to preprocess the depth image before 3D warping. Depth image can be obtained by depth camera or stereo matching algorithm. Accurate depth value is important for identifying foreground and background pixels [35, 42]. Since the disocclusion is caused by depth discontinuity, some methods apply low pass filter to smooth the depth image. Symmetric Gaussian filters have been proposed in [37] to remove isolated noise pixels in the depth image and reduce the area of disocclusion. This method usually leads to the rubber sheet effect, which means the geometric distortion of the object. On this basis, Zhang et al. [44] apply an asymmetric Gaussian filter to filter the depth discontinuity vertically more than horizontally. Based on the anisotropic nature, this filter prevents the disocclusion while avoiding the rubber sheet effect. However, the regions which do not cause disocclusions are also smoothed, resulting in degradation of the depth layers and visual quality. To overcome this problem, Chen et al. [2] propose an edge-dependent Gaussian filter, which only smooths the edges in the horizontal direction and prevents the geometric distortion of non-hole regions. Liu et al. [21] apply the structure-aided filter to process the depth image in the transformed domain. Adaptive smoothing helps prevent the generation of disocclusion. However, smoothing of the edges may result in blurring of the foreground boundaries, and the above methods are suitable for small baseline configurations or small depth discontinuity. Targeted to large baseline, Lei et al. [15] propose a divide-and-conquer depth image preprocessing method that decomposes disocclusion by reducing the depth discontinuity in the foreground edge. This method can achieve good results under the condition of simple background, but for complex background, especially when the vertical texture is rich, the change of depth value may lead to the distortion of structure.

The other type of method is to fill the disocclusions with surrounding textures in the spatial or temporal domain. In the spatial domain, due to the good performance in recovering the unknown region, the image inpainting method is introduced into the disocclusion filling. The exemplar-based inpainting method proposed by Criminisi et al. [5] is widely used for hole filling. In this approach, the confidence and texture of the hole boundary pixels are combined to determine the priority, and the hole is filled with the most similar patch in the source region. In [33], the graphcut algorithm and an example-based inpainting technique are combined to fill the holes. Shen et al. [34] proposed gradient-based image completion algorithm, which could reconstruct the image from the gradient maps by solving a Poisson equation. In [43], Deep learning scheme is used for image content prediction and aesthetic assessment. In fact, as the disocclusions are originated from background region, they should be filled with the background texture. However, this classic method treats the foreground and background boundaries equally, causing some foreground textures to be sampled into the disocclusion. To prevent this problem, some improved methods modify the inpainting method by introducing the depth information. In [6], additional depth term is used in the calculation of priority, and the patch with lower depth variance is assigned higher priority. Moreover, the calculation of the matching cost takes into account the similarity of color and depth information. This method fills the disocclusion under the condition that the depth image of the virtual view is known, which may not hold in practical applications. Ahn et al. [1] generate the depth image of the virtual view during the 3D warping. Disocclusions in the color image and the depth image are then simultaneously filled by the depth-based image inpainting method. Kao [13] optimizes the priority calculation in [6] and uses the depth-based gray-level distance to measure the similarity of patches. However, when the depth value of the foreground edge is inaccurate, ghosts appear at the edge of the disocclusion, causing the penetration of foreground pixels. In order to reduce the interference of foreground pixels on the hole filling, Luo et al. [22] remove the foreground object from the reference image based on the depth information, and then apply the improved Criminisi's method to generate the background image, which is used to fill the disocclusion after 3D warping. Han et al. [12] propose a layered 3D warping method, in which foreground objects are segmented, and the 3D perception and visual comfort are balanced by disparity control. However, these methods are strongly dependent on the accuracy of foreground extraction. In the case where the depth image contains multiple layers, it is difficult to apply the edge detection or the threshold method such as the Otsu thresholding method [29] to segment the foreground object. It may even be necessary to manually specify the foreground object, which affects the robustness of the method.

In the temporal domain, as the foreground object moves, the background it occludes is also changing. Therefore, some methods try to fill the disocclusion by exploiting reliable background pixels from other frames. Sun et al. [36] use the switchable Gaussian model to establish an online background model to adapt to different scenes. Lie et al. [18] propose a background sprite model, which can combine time and spatial information to remove foreground. In [28], a background modeling method based on novelty-view domain is presented to fill the disocclusion. But this method needs to generate several novel viewpoints, which increases the computational complexity. In [8], the improved Gaussian mixture model (GMM) is used to establish an adaptive background model independently for each pixel in the reference image. However, these background models can achieve good results under the assumption of moving foreground objects. For still foreground objects or a single still image, the foreground pixels may not be completely removed. In this case, there is still a risk of foreground pixels penetration.

For view synthesis with two or multiple reference views, more reliable pixels can be used to fill the disocclusion due to the increase of views [32]. Zhu et al. [47] explore the mechanism of disocclusion generation in view interpolation, and filled the disocclusion based on visible and invisible background information. In [16], the experimental results show that multi-view rendering can effectively reduce the area of the disocclusion in view interpolation and extrapolation. Most of the disocclusions can be filled by view merging, which reduces the difficulty of the postprocessing. However, multi-view rendering requires more image acquisition equipment and higher transmission bandwidth [4]. In addition, there are still some holes in the merging result that need to be filled, especially for viewpoint extrapolation. For special application, such as in 2D-to-3D system, only single view can be used. Therefore, exploring the disocclusion filling method for single view synthesis is still necessary.

In order to reduce the penetration of the foreground texture and select high-reliability background pixels to fill the disocclusion, we apply the Laplacian of Gaussian (LOG) operator to generate the laplacian image of the reference depth image, which can be used to identify the pixels on the disocclusion boundary, thereby dividing the disocclusion into two types: FG-BG disocclusion and BG-BG disocclusion. For the former, an adaptive depth image preprocessing method is introduced to remove the ghosts and decompose the disocclusions based on the complexity of the surrounding background texture. For the latter, the valid pixels on the boundary are marked in the reference image. The corresponding foreground object that occludes the background is extracted under the guidance of the depth information, so that the disocclusion can be filled with the reliable background texture. In addition, a postprocessing approach is applied to deal with the remaining artifacts after merging.

3 Proposed disocclusion handling approach

The flowchart of the proposed method is shown in Fig. 2. Firstly, the reference Laplacian image is generated by applying the LOG operator to the depth image. Disocclusions in the original virtual image are marked and divided into two types. Then, the FG boundary leading to the FG-BG disocclusion and its surrounding BG texture are adaptively optimized in the depth image. The ghosts are also removed in this step. Thirdly, based on the valid pixels located on the BG-BG disocclusion boundaries, the associated foreground objects are detected and removed according to the depth features. The removed regions are filled by applying the depth-based inpainting method. The optimized reference image and its depth image are used as input to 3D warping, and the output is merged with the original virtual image. Finally, the remaining other artifacts are handled during the postprocessing. The pseudo code description

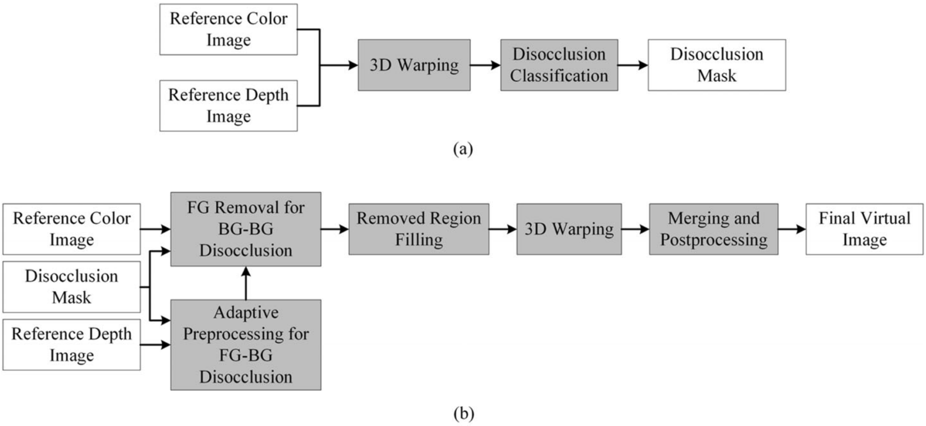

Fig. 2 Flowchart of the proposed method. **a** Disocclusion classification, and **b** view synthesis module with disocclusion mask for hole filling

of the proposed method is shown in Algorithm 1. In the following, the core steps of the proposed method will be described in detail.

Algorithm 1 The pseudo code description of the proposed method

```
Input: reference image $I_r$, depth image $d_r$, camera parameters of reference view and virtual
       view $\mathbf{K}_r$, $\mathbf{K}_v$, $\mathbf{R}$ and $\mathbf{t}$
Output: virtual image $I_v$
Begin
  origin virtual color image $I_{v1}$ and depth image $d_{v1}$ $\leftarrow$ 3D warping ($I_r$, $d_r$, $\mathbf{K}_r$, $\mathbf{K}_v$, $\mathbf{R}$, $\mathbf{t}$);
  disocclusion boundary $\delta\Omega$ $\leftarrow$ edge detection ($d_{v1}$);
  FG-BG disocclusion ($i=1,\ldots,m$) and BG-BG disocclusion ($j=1,\ldots,n$) $\leftarrow$ classify and
  count the disocclusions ($\delta\Omega$);
  locate the disocclusion boundary in the reference image;
  for $i=1,\ldots,m$
    mark the FG boundary of each FG-BG disocclusion;
    analyze the texture complexity of surrounding BG regions;
    adaptive process the depth value of adjacent BG regions;
  end
  for $j=1,\ldots,n$
    mark the valid pixels on the BG-BG disocclusion boundary;
    replace the depth value of the marked pixels with the depth value of the neighboring FG
    pixel;
    calculate the depth difference between the neighboring pixel and the marked FG pixels
    within the minimum bounding rectangle of disocclusion boundary;
    mark the pixels with similar depth values;
    update the FG mask and perform the next iteration until all FG pixels within the
    minimum bounding rectangle are marked;
  end
  remove the marked FG pixels;
  fill the removed regions using the surrounding BG texture and depth value;
  render the filled result and the removed FG to virtual view;
  merge the FG and BG image and handle the remaining artifacts;
  output the final virtual image $I_v$;
end
```

3.1 Classification and marking of disocclusions

Figure 1 shows that the depth value distribution of the pixels on the disocclusion boundary is affected by the baseline setup and depth discontinuity. In the case of small baseline and small depth discontinuity, small disocclusion is generated by the movement of the viewpoint. When the FG object has a certain width, this occlusion hole will only appear between the FG and the BG, that is, its two sides belong to the FG and the BG respectively, as shown in Fig. 1(a). On the contrary, with the increase of baseline and depth discontinuity, the difference between FG and BG displacement in the virtual image becomes larger. Therefore, the area of the disocclusion gradually increases until the entire FG is projected into the new BG. Therefore, the whole occluded BG is exposed in the virtual view, as shown in Fig. 1(b). In this case, most of disocclusion edge pixels belong to the BG. In general, disocclusions occur on the right side of the foreground for the right synthesized view, and vice versa for the left synthesized view. Ghosts appear in the adjacent BG region of the disocclusion, appearing as a mixture of FG and BG textures. Accordingly, this paper takes the disocclusion handling for right virtual view as an example. The similar process can be applied on the rendering of other virtual views. Based on the depth value distribution of the boundary pixels, we classify the disocclusions and apply different handling strategies depending on the nature of them.

The study in [19] shows that the Laplacian is sensitive to the disocclusion boundary and has directional invariance. Therefore, the Laplacian operator is used to identify the valid pixel along the disocclusion boundary as FG boundary or BG boundary. In order to increase the robustness against noise, the LOG operator is introduced. In the case of single view rendering, based on the feature of the depth image, pixels along the disocclusion boundary are identified and labeled as follows:

$$L(u,v)=\begin{cases}\delta\Omega_{FG}, \text{if } (\Delta d)_w(u,v)<0\\ \delta\Omega_{BG}, \text{if } (\Delta d)_w(u,v)>0\end{cases}, \text{for } (u,v)\in\delta\Omega \tag{1}$$

where $\delta\Omega_{FG}$ and $\delta\Omega_{BG}$ donate the FG boundary and BG boundary respectively. $d$is the input depth image and Δd is the Laplacian of d. $(\Delta d)_w$ donates the warped Laplacian of the original depth image. $\delta\Omega$is the boundary of the disocclusion Ω. It should be noted that if the Laplacian value is equal to zero, it means that there is no depth discontinuity at this position and no disocclusion is generated. Therefore, the zero Laplacian value is note mentioned in Eq. (1). The classification result of the disocclusion boundary pixels is shown in Fig. 3, where the FG boundary is marked in red and the BG boundary is marked in green. In this case, disocclusions can be divided into FG-BG disocclusion (Ω_{FB}) and BG-BG (Ω_{BB}) disocclusion according to the proportion of foreground pixels on the boundary. Since the disocclusion is caused by depth discontinuity between the FG and the BG, it has at least one side background boundary. In this paper, the classification threshold is set to 35% for FG ratio. A disocclusion is classified as FG-BG disocclusion if at least 35% of the pixels around the boundary are FG. For other cases, it is classified as BG-BG disocclusion. After the classification, the boundary pixels of the two types of disocclusions are marked in the reference depth image as follows:

$$L_{FB}(u,v)=\begin{cases}1, \text{if } (u,v)_w\in\delta\Omega_{FB}\\ 0, \text{otherwise}\end{cases} \tag{2}$$

$$L_{BB}(u,v)=\begin{cases}1, \text{if } (u,v)_w\in\delta\Omega_{BB}\\ 0, \text{otherwise}\end{cases} \tag{3}$$

where L_{FB} donates the mask of the FG-BG disocclusion boundary and L_{BB} donates the mask of the BG-BG disocclusion boundary, as shown in Fig. 3 (c) and (d). $(u,v)_w$ donates the corresponding pixel of (u,v)in the virtual image. In the subsequent steps, only the marked pixels and their neighbors are processed, so that the computational complexity and the degradation of visual quality can be reduced.

3.2 Adaptive preprocessing for FG-BG disocclusion

Figure 3(c) indicates that for the right virtual view, the FG-BG disocclusion boundary pixels are mainly distributed to the right side of the FG object in the reference image. Due to the depth discontinuity caused by occlusion, adjacent pixels in the reference image are separated in the virtual image, resulting in disocclusions. The accuracy of depth image is very important for view synthesis. Meanwhile, the boundaries of objects in depth image may be mismatched with that of the color image because of the limitations of stereo matching algorithm and hardware equipment. Some FG pixels are assigned BG depth values, and separated from the FG object. They are projected into the BG region and blended with BG texture, called ghosts. If not processed, the subsequent hole filling algorithm will spread these FG textures, resulting in

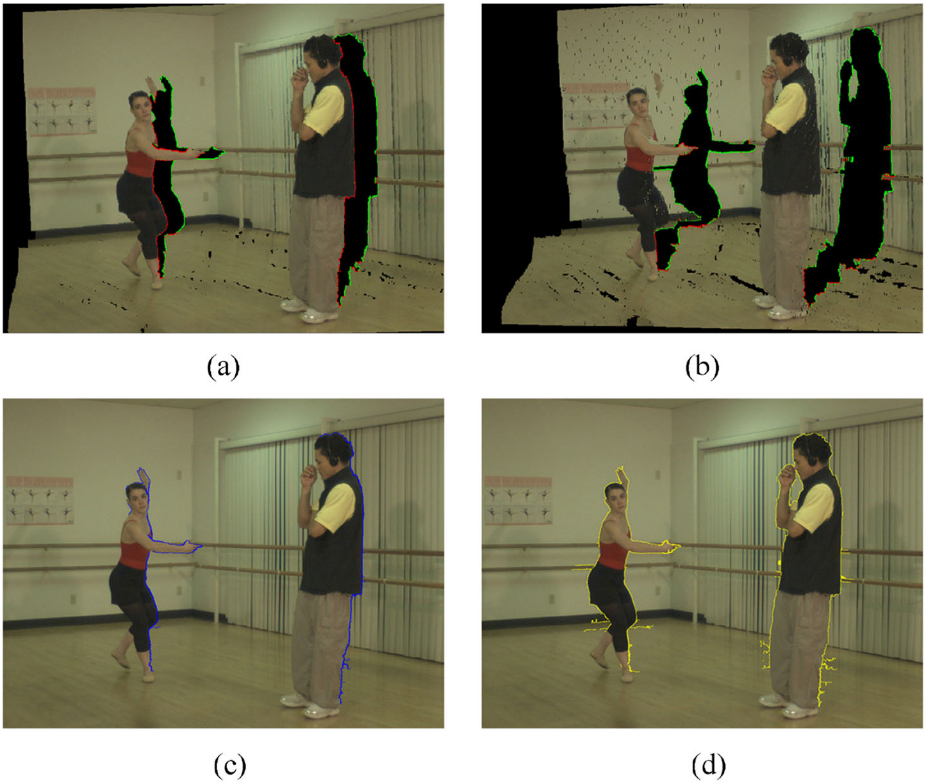

Fig. 3 Classification and marking of the disocclusion. **a** FG-BG disocclusion, **b** BG-BG disocclusion, **c** FG-BG disocclusion boundary in reference image, and **d** BG-BG disocclusion boundary in reference image

incorrect results. Therefore, an adaptive depth image preprocessing method is presented in this section to remove ghosts and decompose the FG-BG disocclusion.

For the right virtual view, the FG-BG disocclusion appear in the depth image where the depth value is changed from high to low. It should be noted that in this paper, the depth value is uniformly described as the pixel value in the depth image. The larger depths are represented by smaller values in the depth image, so that the depth value of the FG is higher than the BG, as shown in Fig. 4. In this case, the FG boundary that may create FG-BG disocclusion can be detected according to depth discontinuity and marked as:

$$E_{FG}(u,v)=\begin{cases}1,\text{if } d(u,v)-d(u+1,v)>T_1\\0,\text{otherwise}\end{cases}\tag{4}$$

where T_1 is the segmentation threshold between the FG and BG. Its value is set based on the depth distribution of the scene in order to extract a more complete FG boundary. Since the FG-BG disocclusion boundary pixels is marked in the reference image and its depth image, the FG boundary pixels extracted by Eq. (4) are further filtered. The FG boundary for preprocessing is limited to pixels adjacent to the FG-BG disocclusion boundary, and is marked as:

$$Mask(u,v)=\begin{cases}1,\text{if } E_{FG}(u,v)=1\&\&N(u,v)\cap L_{FB}\neq\varnothing\\0,\text{otherwise}\end{cases}\tag{5}$$

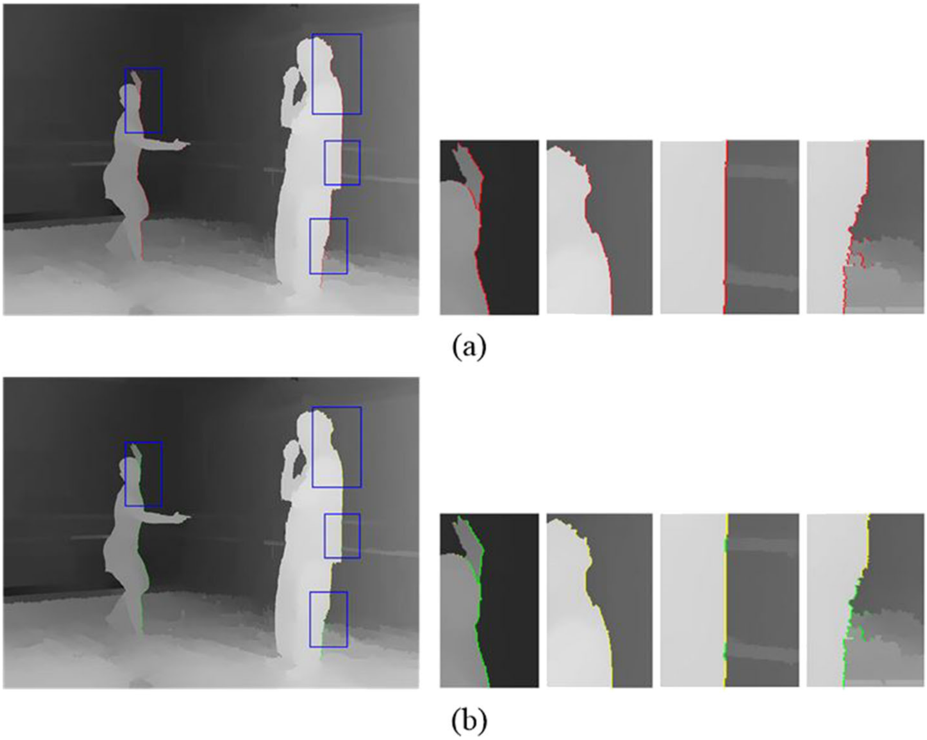

Fig. 4 Marking and classification of FG boundary. **a** Marking result, and **b** classification result based on texture complexity (R_1 marked in green, and R_2marked in yellow)

where $N(u, v)$ donates the set of (u, v) and its immediate neighbors. Figure 4(a) shows the filtered FG boundary mask. Based on the marked pixels, the depth values of their neighbor BG are optimized. Since the texture of image is not immutable, texture analysis is performed in these BG regions. For the identified BG region with flat texture, the distinct depth discontinuity is replaced by a gradually decreasing depth change, so that the FG-BG disocclusion can be decomposed into multiple small holes through the preprocessing of the depth image. After decomposition, the hole becomes smaller and its surrounding pixels belong to the BG texture, so it is easy to be filled. In order not to cause geometric distortions, we only decompose disocclusions with flat BG textures. For the remaining region with complex texture, its depth value is preserved to avoid geometric distortion in the virtual image. Moreover, in order to remove ghosts, all of the marked FG boundaries are extended to the BG region in the depth image, so as to correct the error in depth acquisition and prevent the penetration of foreground pixels in subsequent inpainting process.

As gradient information can reflect the complexity of image texture, we apply gradient operator to achieve texture analysis. The marked FG boundary is classified into two regions as:

$$\begin{cases} R_1 = \{(u,v)|Mask(u,v) = 1\&\&\lambda\cdot|G_u(u+k,v)| + (1-\lambda)\cdot|G_v(u+k,v)| < T_2, k = \{2,3,\cdots W_i\}\} \\ R_2 = \{(u,v)|Mask(u,v) = 1\&\&(u,v)\notin R_1\} \end{cases} \tag{6}$$

Where $G_u(u, v)$ and $G_v(u, v)$ represent the gradient values in the horizontal and vertical directions at (u, v), respectively. The human visual system obtains depth cues from disparity mainly from horizontal differences rather than vertical differences [44]. Compared with the vertical direction, the geometric distortion in the horizontal direction is more intolerable, and the visual quality of the virtual image is significantly reduced. Therefore, the horizontal direction should be given higher weight. In addition, the preprocessing is performed in the horizontal direction, so the weight coefficient λ is set to 0.7 in the experiment. $W_i$represents the average width of the FG-BG disocclusion i. T_2 represents the gradient threshold. For depth image, the gradient difference means the change in pixel value in an 8-bit single channel image where the value represents the disparity in pixels between the outermost views. Large gradient differences reflect large texture changes. In order to reduce geometric distortion, the threshold T_2 should be set to a small value. Considering that the color image has three channels, so T_2 is set to 30 in the experiment. The classification result of R_1 and R_2 is shown in Fig. 4(b). Therefore, the distortion can be reduced in the case of disocclusion decomposition, especially for linear structure, and the depth value of the adjacent FG can also be preserved.

For the region R_1 with flat texture, as the ghosts are usually 1–2 pixels wide, the marked FG boundary is firstly extended in the depth image to remove the ghosts as:

$$d(u+1,v)=d(u+2,v)=d(u,v),\text{if }(u,v)\in R_1 \tag{7}$$

In the case of the fixed baseline and scene, the area of FG-BG disocclusion is proportional to the depth discontinuity between adjacent pixels in the reference depth image, meaning that the larger depth discontinuity results in a larger disocclusion. Due to the additive nature of the error, the reliability of the inpainting method becomes lower at the center of the disocclusion. Therefore, the core idea of disocclusion decomposition is to reduce the depth discontinuity between adjacent pixels. For the BG region adjacent to the marked FG boundary, except for the extended region whose depth value is replaced with the FG depth value, a linear descent process is used in the horizontal direction to optimize the depth value. Large depth discontinuity between FG and BG is modified to a slow depth drop among several pixels. In order to make an even width of each hole after decomposition, a fixed depth drop parameter is set based on the depth distribution of scene. For the FG boundary belonging to R_1, the linear drop process for each row can be expressed as:

$$\begin{aligned} d(u+k+1,v)=d(u+k,v)-m,\text{if }(u,v)\in R_1\&\&d(u+k,v)-d(u+k+1,v)>m,k\\ =\{2,3,\cdots W_i\} \end{aligned} \tag{8}$$

In the experiment, it is necessary to ensure that the value of $m\cdot(W_i-1)$ is greater than the maximum depth discontinuity of the disocclusion i. Through decomposition process, the flat background texture can be evenly distributed around each hole, thus preventing the FG texture penetration. In addition, the reduction of the hole area helps to improve the accuracy of the inpainting result.

For the region R_2 with rich texture, as the contour of FG is irregular, the decomposition process may result in different depth values in the same linear texture, evolving into the stretching and distortion of the texture in the virtual image. Thus, only the ghost removal process is performed on this region, as follows:

$$d(u+1,v)=d(u+2,v)=d(u,v),\text{if }(u,v)\in R_2 \tag{9}$$

The depth image preprocessing result for the FG-BG disocclusion is shown in Fig. 5(a). Note that the boundary of the FG is extended to the BG region. The result in Fig. 5(b) shows that the

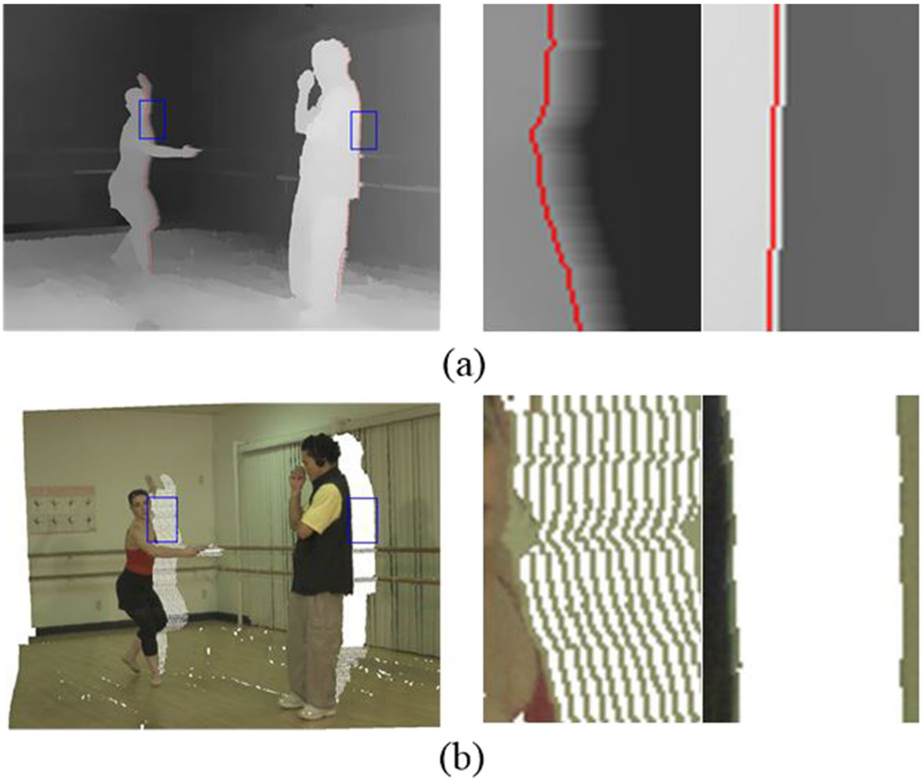

Fig. 5 Processing result of the proposed method. **a** Depth image preprocessing result (original FG boundaries marked in red), and **b** 3D warping result with adaptive preprocessing

proposed method can adaptively implement the segmentation transition from FG to BG, thereby effectively decomposing the disocclusion and improving the filling accuracy.

3.3 FG removal for BG-BG disocclusion

In general, disocclusions belong to the BG region that is occluded by the FG object in the reference view. These regions become visible in the virtual view, but the associated pixels cannot be retrieved from the reference image, thus causing holes. For view synthesis, as the viewpoint moves, different depth values mean different displacements of the pixels. Compared with the BG object, the parallax of the FG object is larger under the same baseline. In this case, when the FG object occludes a new BG region, that is, the FG and BG visible in the reference view overlap in the virtual image, the occluded BG in the reference view is completely exposed, forming the BG-BG disocclusion that both sides belong to the BG region. GMM-based methods can realize the detection of FG objects and establish the BG model by BG subtraction. However, such methods are only applicable to the case of moving FG objects, and still FG objects are usually considered part of the BG and remain in the BG model. For the FG extracting methods based on the edge detection, threshold selection is a difficult task.

In this paper, features of the depth image are introduced to extract BG objects that occlude BG-BG disocclusion region in the reference view. As the BG-BG disocclusion boundary pixels are located on the boundary of FG object in reference image, the edge of FG object can be extracted by detecting these pixels. Since adjacent pixels in the same FG object have similar depth values, extraction and removal of the specific FG objects can be achieved by an iterative process. Furthermore, in order to remove artifacts and completely extract FG objects, the depth values of the disocclusion boundary pixels are corrected before the iterative process.

After the valid pixels on the BG-BG disocclusion boundary are marked, these pixels are re-projected into the reference image, as shown in Fig. 6(a). A threshold is assigned to each disocclusion boundary based on the depth information to assist in the extraction of the FG object. In order to remove outliers, for the BG-BG disocclusion i, the FG-BG segmentation threshold T_i is set as the average depth value of the pixels along its boundary after removing the 10% highest and lowest values. For each disocclusion, a minimum bounding rectangle (MBR) is established in the reference depth image based on the distribution of its boundary pixels. The iterative process takes place within the rectangle. Due to the existence of ghosts, although most of the marked boundary pixels have BG depth values, they belong to the foreground object in the color image. Thus, the depth values of these pixels are firstly corrected. According to the condition under which disocclusion occurs, there must be pixels with FG depth value in the neighborhood of the boundary pixels. Therefore, the depth values of these marked pixels are replaced with the highest depth value within their neighborhood, that is, they are re-marked as FG pixels. Based on the marked FG boundary, the complete FG object is gradually extracted by judging whether its immediate neighbors belong to the FG. For the pixels x on the boundary F_i of FG i, y is an unmarked neighborhood pixel of x. According to the characteristics of depth image, if y belongs to the FG, it should have the similar depth value with x, and its depth value is higher than the BG it occluded. Therefore, the condition that y is appended to FG i can be expressed as follows:

$$y \in F_i \,||d(x)-d(y)| < \alpha \&\& d(y) > T_i, y \in N(x) \cap \mathrm{MBR}_i \tag{10}$$

where α is a small value and is set to 3 in our experiment. When all the pixels of F_i have been traversed, the new state of the FG object is used as the input to the next iteration. The iterative process continues until all the pixels in the MBR are traversed, and the extraction of the FG object is achieved. In addition, a morphological expansion operation is performed on the extracted FG region to extend the boundary outward so that all ghost pixels are included. The extracted foreground mask is shown in the Fig. 6(d), and then the corresponding FG objects are removed from the reference image and its depth image, as shown in Fig. 6(e) and (f). Note that some parts of FG object are still not removed. This is because their depth values are very close to the neighborhood BG pixels, so that no disocclusion is generated in the virtual, and it will not affect the disocclusion filling.

3.4 Removed region filling

As mentioned in Section 1, Criminisi's algorithm is an effective image inpainting method. This method takes into account the texture and structure information of the image. It determines the fill order based on the priority term, and search the most similar patch from the source region for hole filling. Compared with pixel-based methods, it does not introduce blurred artifacts. In

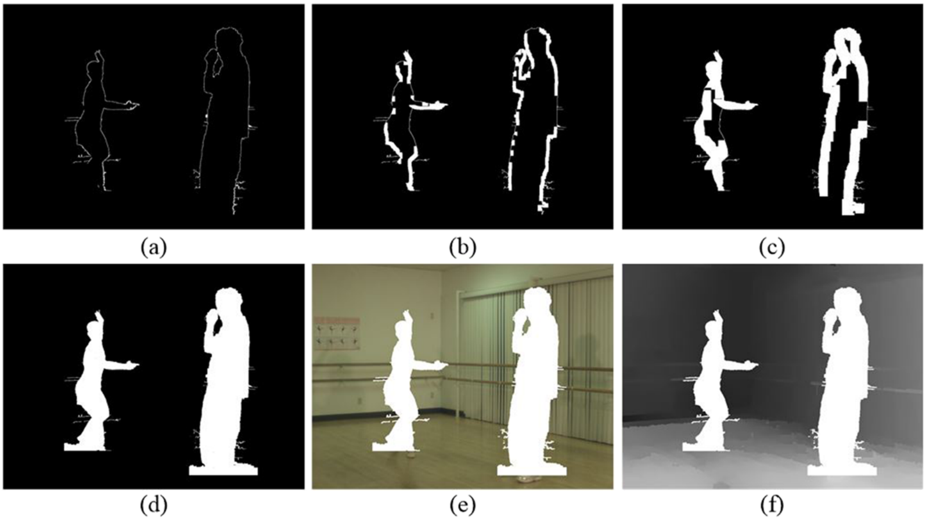

Fig. 6 FG removal for BG-BG disocclusion. **a** Disocclusion edge mask in reference image, **b** FG extraction result of the 10th iteration, **c** FG extraction result of the 30th iteration, **d** FG extraction result, **e** FG removal result in reference image, and **f** FG removal result in depth image

this paper, we extend the Criminisi's algorithm to fill the removed regions in the reference image and its depth image.

Image inpainting starts from the boundary of removed regions. In the classic Criminisi's algorithm, for a input image I, Ω is the region to be filled and Φ represents the remaining source region ($\Phi = I - \Omega$). For pixel p on the boundary of Ω, assume that the square patch Ψ_p centered at p is to be filled, and its priority is defined as:

$$P(p) = C(p) \cdot D(p) \tag{11}$$

where $C(p)$ and $D(p)$ are the confidence term and data term, respectively. $C(p)$ donates the percentage of valid pixels in Ψ_p. $D(p)$ represents strength of isophotes on the boundary, encouraging the linear structure to be inpainted first. Once all priorities on the boundary are computed, the pixel $\widehat{p}$ with highest priority and its corresponding patch $\Psi_{\widehat{p}}$ is found. Then, the most similar patch $\Psi_{\widehat{q}}$ in source region is searched to fill the hole in $\Psi_{\widehat{p}}$. The similarity between the two patches is computed by using the Sum of Squared Difference (SSD) of the valid pixels, so $\Psi_{\widehat{q}}$ is determined as follows:

$$\Psi_{\widehat{q}} = \arg \min_{\Psi_q \in \Phi} \mathrm{SSD}\left(\Psi_{\widehat{p}}, \Psi_q\right) \tag{12}$$

As the hole filling is an iterative process, the confidence term is updated before the next iteration as follows:

$$C(p) = C\left(\widehat{p}\right), \forall p \in \Psi_{\widehat{p}} \cap \Omega \tag{13}$$

It is noted that for removed region filling, as the FG objects and ghosts are removed, the pixels around the hole belong to the BG region. Compared with the disocclusion filling in virtual image, the texture in the reference image is taken from the real scene. Therefore, the filling process performed in the reference image can prevent the spread of artifacts which are introduced in 3D warping. Although the hole filling no longer starts from the FG boundary, the guidance of depth information is still of great significance for the image with multiple depth layers. In the proposed method, the modified priority term can be expressed as:

$$P(p) = C(p) \cdot D(p) \cdot Z(p) \tag{14}$$

where $Z(p)$ is the introduced depth term, indicating the average depth value of the valid pixels in Ψ_p, given by:

$$Z(p) = \frac{d_{\max} - \frac{\sum_{q \in \Psi_p \cap \Phi'} d(q)}{\left|\Psi_p \cap \Phi'\right|}}{d_{\max} - d_{\min}} \tag{15}$$

where $d_{\max}$ and $d_{\min}$ are the highest and lowest nonzero depth values in the depth image, respectively. Φ' is the newly defined source region with size $N \times N$ centered at $\widehat{p}$, where $N$is set to twice the length of the long side of MBR$_i$ to ensure the exploration of the suitable source patch and texture spatial locality. The depth term gives higher priority to the patch with lower depth value, thereby ensuring that the filling starts from a local BG region in the case of multiple depth layers. Once all the priorities are computed, $\Psi_{\widehat{q}}$ is searched in the Φ'. Since adjacent patches in the BG region usually have similar depth values, depth information is introduced in the similarity computation, and $\Psi_{\widehat{q}}$ is determined as:

$$\Psi_{\widehat{q}} = \arg\min_{\Psi_q \in \Phi'} \left[\mathrm{SSD}_{\mathrm{color}}\left(\Psi_{\widehat{p}}, \Psi_q\right) + \mathrm{SSD}_{\mathrm{depth}}\left(\Psi_{\widehat{p}}, \Psi_q\right) \right] \tag{16}$$

Moreover, in order to preserve the update of depth term, the removed region filling is simultaneously performed in the reference image and its depth image. The filling result is shown in Fig. 7, which can be used to fill BG-BG disocclusions in the virtual image.

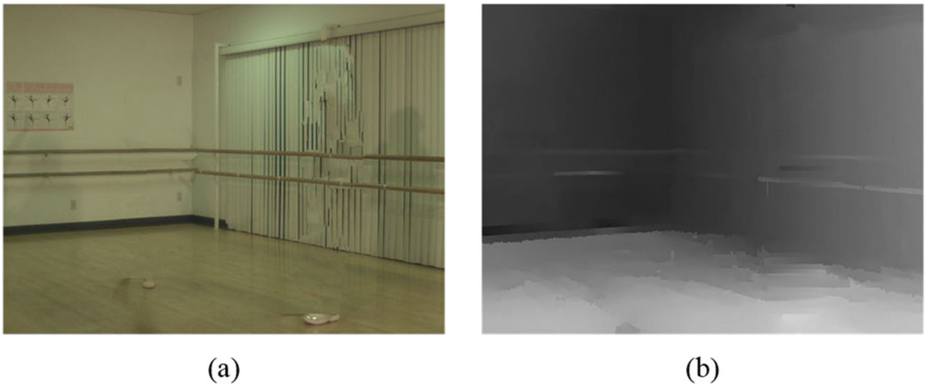

Fig. 7 Filling result of the removed region. **a** Filling result in color image, and **b** filling result in depth image

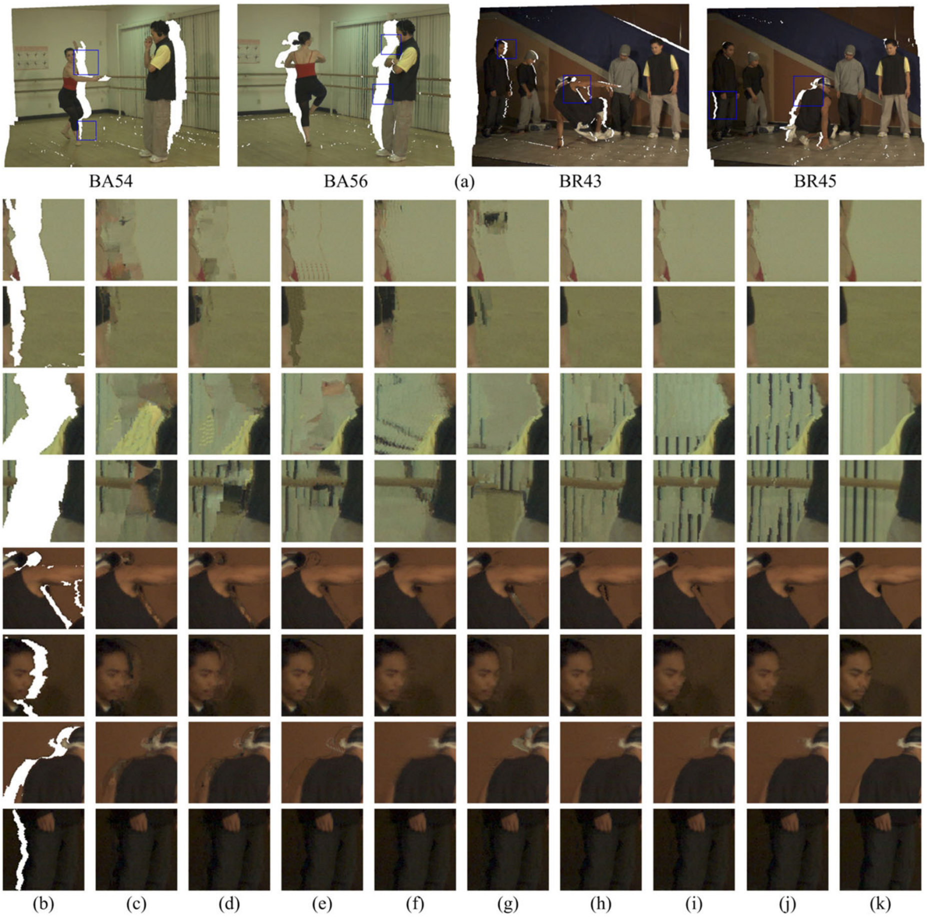

Fig. 8 Visual quality comparison of disocclusions for small baseline. **a** Warped virtual image, **b** magnified parts of (**a**), **c** Criminisi's method, **d** Daribo's method, **e** Ahn's method, **f** Lei's method, **g** Kao's method, **h** Chen's method, **i** Oliveira's method, **j** proposed method, and **k** ground truth

3.5 Merging and postprocessing

After the depth image preprocessing and corresponding FG removal, the optimized reference image and its depth image are warped to the target view. As the corresponding FG objects are removed, BG-BG disocclusions will no longer appear in the rendered image, and the FG-BG disocclusion is adaptively divided into several small holes. Note that these holes are almost surrounded by the BG pixels, so the proposed inpainting method can effectively fill them with reliable BG textures. Since some FG objects may be removed from the optimized reference image, a subsequent merging process is necessary, which combines the original DIBR result with the optimized virtual image. In order to preserve the correct occlusion relationship between the FG and BG, Z-buffer algorithm [11] is applied to process the pixels at the same location.

However, there are still some other types of artifacts in the original DIBR result, such as cracks and OFRs. They are not caused by the occlusion between the FG and BG objects, therefore, our method mentioned above do not detect and handle them. After merging, an

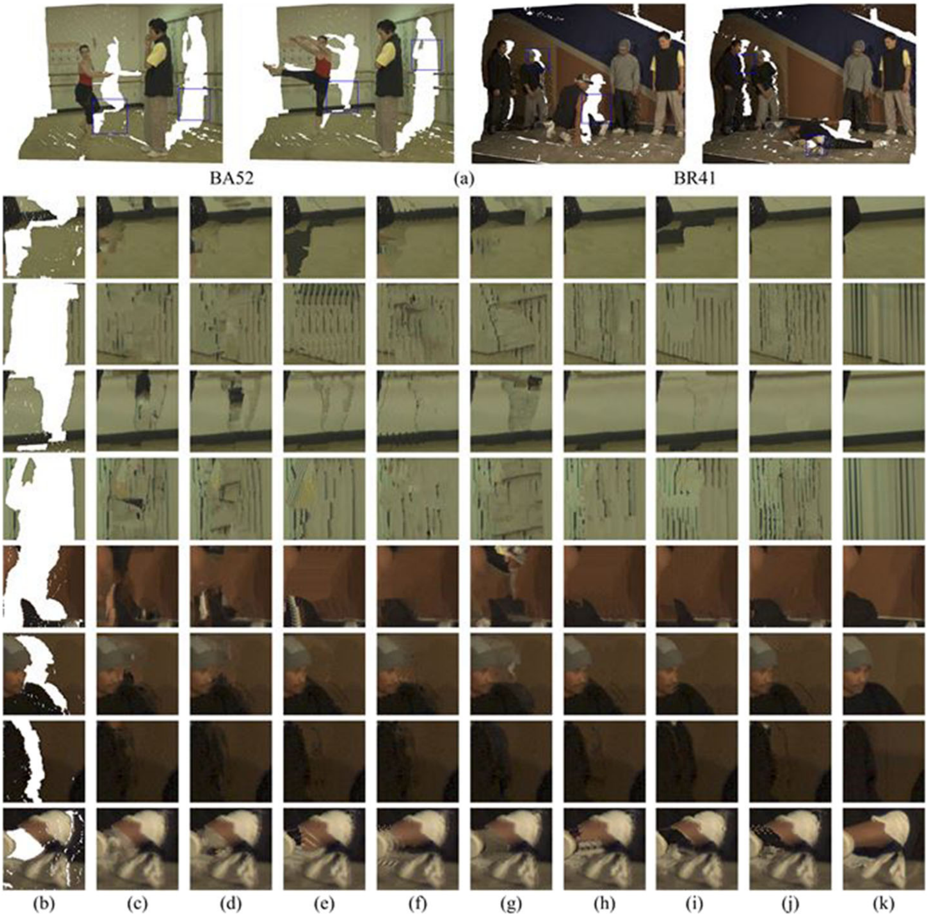

Fig. 9 Visual quality comparison of disocclusions for large baseline. **a** warped virtual image, **b** magnified parts of (**a**), **c** Criminisi's method, **d** Daribo's method, **e** Ahn's method, **f** Lei's method, **g** Kao's method, **h** Chen's method, **i** Oliveira's method, **j** proposed method, and **k** ground truth

additional postprocessing is applied to deal with these remaining artifacts. As proposed in our previous work [3], cracks caused by the rounding of the coordinate values are filled by the optimized DIBR algorithm, and the reference view extension approach is used to fill OFRs. Moreover, for other small holes caused by depth errors, since they are usually inside the FG or BG objects, satisfactory result can be obtained by applying a simple inpainting method in the postprocessing.

4 Experimental results and discussion

In order to validate the proposed approach, two Multiview Video-plus-Depth (MVD) sequences "Ballet" and "Breakdancers" are used in our experiment [48]. Each sequence contains 8 viewpoints, which have a resolution of 1024 × 768 pixels and 100 frames long. The "Ballet" sequence has large depth discontinuity between the FG and BG objects, and the FG object is

Table 1 PSNR Comparison for MVD sequences

Test sequence	PSNR (dB)							
	[5]	[6]	[1]	[15]	[13]	[3]	[7]	Proposed
BA54	26.16	29.13	29.57	29.60	28.56	30.25	31.16	31.42
BA52	23.95	24.82	24.89	24.33	23.97	25.93	25.71	26.29
BA56	25.85	27.54	27.69	27.95	27.79	28.25	28.19	28.64
BR43	28.27	28.81	29.63	28.99	29.68	29.96	30.58	30.53
BR41	25.67	26.40	26.59	26.41	26.43	26.95	27.47	27.62
BR45	28.19	29.55	29.57	29.63	29.74	30.02	30.36	30.86

far from another one. The "Breakdancers" sequence contains a series of FG objects with similar depth layers and the overlap of FG objects occurs in some frames. Small depth discontinuities result in small disocclusions, even in the case of large baseline. In addition, five public image-plus-depth sequences from the Middlebury Stereo Data Sets [31] are used to evaluate the performance of our method. In these data sets, the ground truth of the depth image is obtained using high-precision structured light. Camera parameters are known, including internal and external parameters. The synthesized view is named after the video sequence and the projection configuration, i.e., BA54 donates the sequence warped from view 5 to view 4 of "Ballet". To evaluate the performance of the proposed method, we measure the subjective and objective quality of the synthesized virtual view at different projection configurations and compare the results with the previous methods. The depth distribution of the scene affects the setting of some parameters. In our experiment, T_1 is set to 25 for "Ballet" and 15 for "Breakdancers" and Middlebury data sets. m is set to 8 for "Ballet" and 5 for "Breakdancers" and Middlebury data sets.

7 previous methods are chosen for subjective and objective quality comparisons, including Criminisi's classic inpainting method [5], Daribo's depth-aided inpainting method [6], Ahn's inpainting-based method [1], Lei's divide-and-conquer hole filling method [15], Kao's depth-based inpainting method [13], our previous work in [3] and Oliveira's filling method [7]. It should be noted that Daribo's method is under the assumption that the depth image of virtual view is known, while other methods and the proposed method need only a single reference view and its depth image for view synthesis.

The visual quality comparison results of disocclusions for small baseline are shown in Fig. 8, including BA54, BA56, BR43 and BR45. Influenced by the baseline distance and depth discontinuity between the FG and adjacent BG objects, all disocclusions belong to the FG-BG type. It can be seen that the proposed method outperforms others in terms of the synthesized

Table 2 SSIM Comparison for MVD sequences

Test sequence	SSIM							
	[5]	[6]	[1]	[15]	[13]	[3]	[7]	Proposed
BA54	0.7850	0.7937	0.8081	0.8015	0.7912	0.8102	0.8159	0.8229
BA52	0.7306	0.7415	0.7448	0.7372	0.7341	0.7484	0.7410	0.7546
BA56	0.7855	0.7913	0.7930	0.7922	0.7894	0.7933	0.8043	0.8074
BR43	0.7828	0.7840	0.7884	0.7842	0.7864	0.7907	0.7946	0.7928
BR41	0.7424	0.7455	0.7487	0.7481	0.7499	0.7532	0.7546	0.7566
BR45	0.7831	0.7857	0.7899	0.7885	0.7892	0.7917	0.7922	0.7996

Table 3 FSIMc Comparison for MVD sequences

Test sequence		FSIMc						
	[5]	[6]	[1]	[15]	[13]	[3]	[7]	Proposed
BA54	0.9269	0.9338	0.9409	0.9387	0.9292	0.9507	0.9516	0.9551
BA52	0.8238	0.8317	0.8345	0.8227	0.8235	0.8475	0.8359	0.8527
BA56	0.9206	0.9273	0.9367	0.9323	0.9337	0.9427	0.9395	0.9439
BR43	0.9492	0.9535	0.9528	0.9534	0.9530	0.9557	0.9567	0.9574
BR41	0.9011	0.9092	0.9123	0.9106	0.9009	0.9182	0.9096	0.9202
BR45	0.9576	0.9618	0.9612	0.9623	0.9607	0.9626	0.9623	0.9635

appearance and looks more plausible, while the filling results of other methods contain some artifacts or FG textures. In Criminisi's method [5], although the classic inpainting method uses patches to propagate textures and prevent the blurring effects, it does not take into account the fact that disocclusions usually belong to the BG region. Simultaneous inpainting from the FG and BG boundaries causes the FG texture to be incorrectly introduced into the synthesized image, as shown in Fig. 8(c). In Daribo's method [6], depth variance is introduced into the computation of priority and patch distance, but the presence of ghosts makes some artifacts appear in the disocclusion region, as shown in Fig. 8(d). Moreover, this method requires the depth image of virtual view, which is difficult to implement in practical applications. Ahn's method [1] solves this problem by generating the virtual depth image during 3D warping, so that the disocclusion is filled simultaneously in the virtual image and its depth image. The introduction of depth information reduces the priority of FG pixels, but when the boundary of FG objects in color image are mismatched with that of the depth image, some unexpected defects can be generated in the filling results, as shown in Fig. 8(e). the results of Lei's method [15] are shown in Fig. 8(f), which can produce satisfactory results for regions with simple background textures. But for background regions with complex textures, modified depth values cause the structural distortion of BG objects. Subsequent inpainting increases the error, resulting in a decline in the visual quality of the virtual view. Moreover, some artifacts occur along the boundary of FG object. In Kao's method [13], depth image preprocessing is applied before 3D warping, and the width of FG object is extended in the depth image. However, for FG-BG disocclusion, ghosts only occur on the BG side, so the extension of the other side causes some BG textures to be incorrectly projected into the virtual image. The priority computation based on the depth variance is not sensitive to the recognition of FG and BG,

Table 4 VSI Comparison for MVD sequences

Test sequence		VSI						
	[5]	[6]	[1]	[15]	[13]	[3]	[7]	Proposed
BA54	0.9775	0.9842	0.9879	0.9847	0.9833	0.9903	0.9908	0.9920
BA52	0.9577	0.9607	0.9611	0.9602	0.9604	0.9656	0.9640	0.9689
BA56	0.9791	0.9823	0.9851	0.9845	0.9855	0.9878	0.9870	0.9889
BR43	0.9913	0.9918	0.9916	0.9923	0.9915	0.9925	0.9926	0.9927
BR41	0.9819	0.9839	0.9840	0.9842	0.9825	0.9854	0.9839	0.9869
BR45	0.9924	0.9933	0.9935	0.9939	0.9934	0.9938	0.9939	0.9941

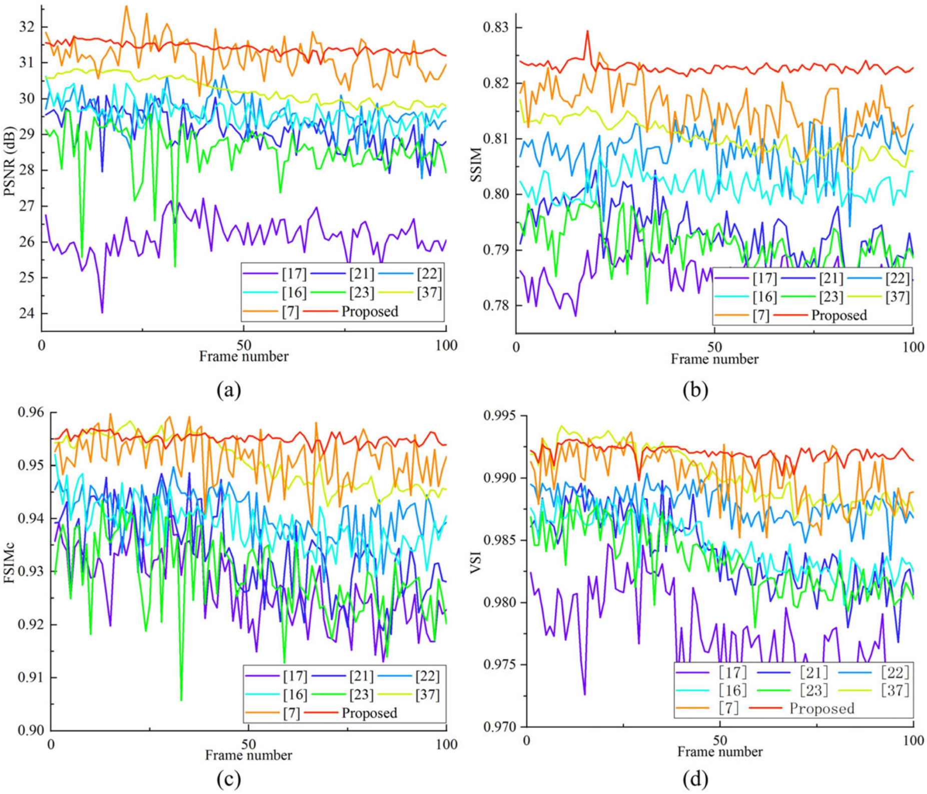

Fig. 10 Objective results for BA54. **a** PSNR results, **b** SSIM results, **c** FSIMc results, and **d** VSI results

especially for scenes with multiple depth layers. This causes some defects and erroneous textures to be produced into the synthesized results, as shown in Fig. 8(g). Figure 8(h) shows the results of our previous work [3]. Local background term and depth term are introduced in the priority computation. This method selects the appropriate BG texture to fill the disocclusion in most cases, but in regions with low confidence, some wrong textures may be sampled. In Oliveira's method [7], Confidence term and data term are replaced by depth term and background term in priority calculation. This makes the disocclusion filling start from the BG side. However, the texture information is ignored, resulting in that the linear structure cannot be preferentially extended, as shown in Fig. 8(i). The proposed method decomposes the disocclusions based on the texture complexity, which reduces the number of low-confidence fillings, and shows reasonable textures in Fig. 8(j).

In the case of large baseline, the area of the disocclusion increases, and the type to which it belongs is related to the depth distribution of the scene. For "Ballet" sequence, disocclusions generated in the virtual image are converted to the BG-BG type, while most of the disocclusions in the virtual image of "Breakdancers" are still maintain the FG-BG type, as shown in Fig. 9(a). Increased disocclusion area makes the filling process more challenging. The comparison results for disocclusion handling are shown in Fig. 9(c)-(j). The synthesized textures in the proposed method look more plausible than others. Especially for BG-BG disocclusion, as the associated FG objects are completely removed, the hole can be filled with

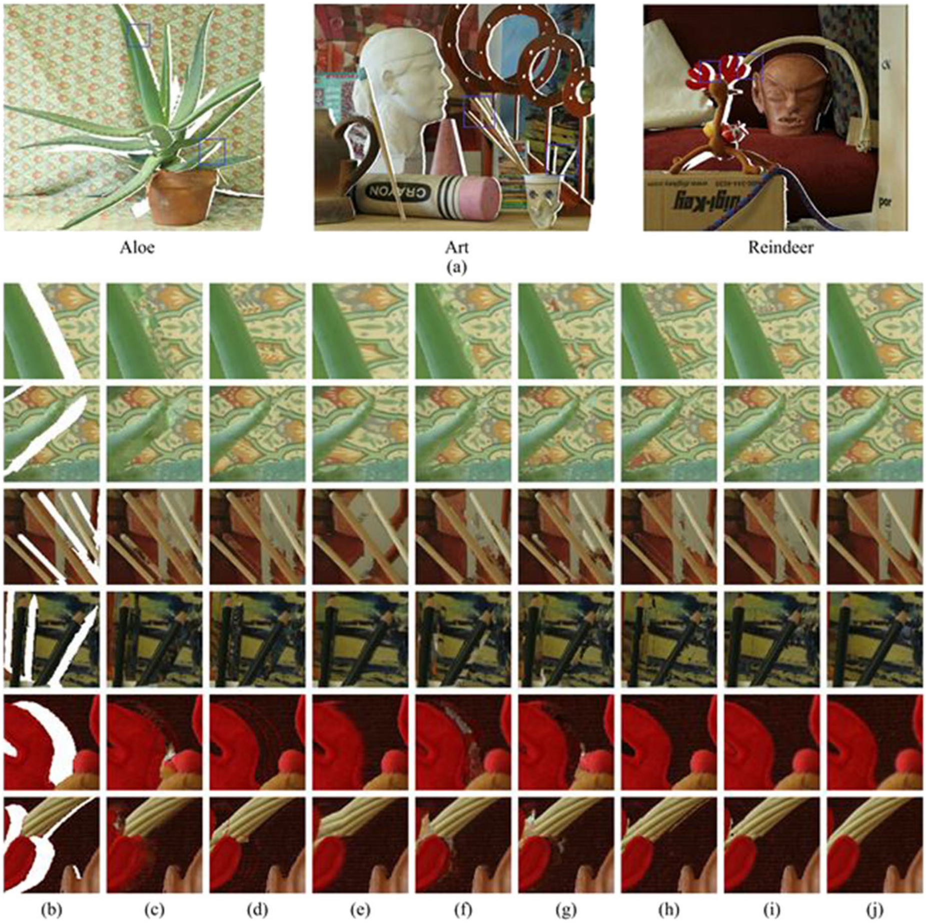

Fig. 11 Visual quality comparison of disocclusions for image sequence rendering. **a** warped virtual image, **b** magnified parts of (**a**), **c** Criminisi's method, **d** Ahn's method, **e** Lei's method, **f** Kao's method, **g** Chen's method, **h** Oliveira's method, **i** proposed method, and **j** ground truth

appropriate background textures around it, so the result looks the more likely the ground truth, while other methods contain some artifacts and unrealistic textures.

To objectively evaluate the performance of the proposed method, in our experiment, peak signal to noise ratio (PSNR), structural similarity (SSIM) [40], feature similarity index (color)

Table 5 PSNR Comparison for still image data sets

Data set	PSNR (dB)						
	[5]	[1]	[15]	[13]	[3]	[7]	Proposed
Aloe	28.13	27.91	29.11	29.28	29.63	29.41	29.74
Art	26.81	27.41	28.32	27.78	28.52	29.77	30.14
Baby1	32.65	33.07	33.14	33.42	33.64	33.70	33.85
Lampshade2	31.36	33.78	35.26	35.49	35.99	36.42	36.81
Reindeer	32.17	33.06	32.85	32.93	33.27	33.57	33.75

Table 6 SSIM Comparison for still image data sets

Data set	SSIM						
	[5]	[1]	[15]	[13]	[3]	[7]	Proposed
Aloe	0.9007	0.8943	0.9046	0.9069	0.9058	0.9063	0.9082
Art	0.8979	0.9049	0.9055	0.9094	0.9127	0.9160	0.9156
Baby1	0.9306	0.9289	0.9330	0.9322	0.9327	0.9336	0.9343
Lampshade2	0.9426	0.9456	0.9574	0.9563	0.9566	0.9578	0.9584
Reindeer	0.9230	0.9251	0.9291	0.9256	0.9260	0.9294	0.9319

(FSIMc) [45], and visual saliency-induced index (VSI) [46] are used to measure the similarity between the synthesized image and ground truth. PSNR measures the squared intensity difference and SSIM measures the perceptual visual quality, while FSIMc and VSI focus on the gradient feature and color feature. The higher PSNR value means the higher similarity between the two images. The values of SSIM, FSIMc and VSI are normalized to 0–1, where 1 donates the highest similarity, i.e., the two images are identical. The objective comparison results for MVD sequences are given in Tables 1, 2,3 and 4. The proposed method achieves the best overall results. The frame-wise objective results for BA54 are shown in Fig. 10. In most cases, the proposed method outperforms the other methods. Objective comparison results indicate that the proposed algorithm would effectively handle the disocclusions and is robust to scene and baseline distance changes.

For image sequence rendering, the subjective comparison results are shown in Fig. 11. Since the reference depth image is obtained by structured light, it has high accuracy and is conducive to the view synthesis. As the depth image of the virtual view in the Middlebury data sets is not provided, the performance of algorithm [6] is not evaluated. Subjective comparison results show that the proposed method has better results in most cases. Our method implements effective measures for different types of disocclusion. During the filling process, the edge of the FG object is maintained and errors caused by artifacts are effectively prevented. For BG-BG disocclusions, as the corresponding FG is removed, the FG texture would not be sampled into the disocclusion region. The objective comparison results for Middlebury data sets are shown in Tables 5, 6, 7 and 8. It can be seen that the proposed method achieves the best overall results.

In addition to the visual quality results of the proposed method, the performance in terms of computational complexity is analyzed. Compared with the original DIBR, the computation cost of the proposed method is higher. In MATLAB environment, for the synthesis of first

Table 7 FSIMc Comparison for still image data sets

Data set	FSIMc						
	[5]	[1]	[15]	[13]	[3]	[7]	Proposed
Aloe	0.9554	0.9596	0.9707	0.9736	0.9739	0.9742	0.9752
Art	0.9394	0.9567	0.9646	0.9671	0.9733	0.9738	0.9740
Baby1	0.9716	0.9723	0.9778	0.9793	0.9796	0.9787	0.9801
Lampshade2	0.9534	0.9726	0.9827	0.9799	0.9837	0.9826	0.9864
Reindeer	0.9747	0.9819	0.9845	0.9780	0.9856	0.9863	0.9904

Table 8 VSI Comparison for still image data sets

Data set	VSI						
	[5]	[1]	[15]	[13]	[3]	[7]	Proposed
Aloe	0.9887	0.9891	0.9916	0.9933	0.9919	0.9924	0.9929
Art	0.9809	0.9901	0.9916	0.9907	0.9925	0.9930	0.9933
Baby1	0.9955	0.9953	0.9958	0.9963	0.9962	0.9966	0.9968
Lampshade2	0.9914	0.9932	0.9964	0.9975	0.9973	0.9977	0.9982
Reindeer	0.9941	0.9966	0.9969	0.9957	0.9968	0.9972	0.9976

frame of BA54, the original DIBR algorithm takes 1542 ms, and the proposed method takes 10,314 ms. The ratio of elapsed time is 6.69. As the proposed method preforms the 3D warping process multiple times and introduces a patch-based inpainting method, the improvement of the visual quality leads to an increase in computational complexity. In addition, the proposed method is implemented in MATLAB with an unoptimized code. The running time cost can be greatly reduced by optimizing the proposed method in a more efficient language such as C++. GPU-based parallel computing or related hardware such as FPGA is a viable measure to improve computational efficiency. In some steps of the proposed method, the computation for each pixel to be processed is independent, such as 3D warping, detection and marking of FG or BG pixels, and the computation of priority and similarity in the inpainting process. Therefore, parallel computing in software or hardware can be introduced into these processes to reduce the computation cost.

At present, deep learning technology has been widely used in the field of image processing, such as deep visual attention prediction [39], fast tracking [17], and image segmentation, etc. Based on a large amount of training data, the established deep learning framework can quickly output accurate results and is very robust for known scenes. In terms of virtual view synthesis, the scenes used are common and available, and deep networks can be used for moving foreground object tracking, foreground segmentation, and image inpainting. Therefore, in future work, we will try to combine our method with the deep network. Migrate the existing deep network to the scene of view synthesis, and introduce depth information to train and improve the network framework, so that the accuracy and efficiency of the view synthesis method can be further improved.

5 Conclusion

In this paper, the underlying mechanisms of disocclusion generation and the factors affecting the distribution of pixels around it are investigated. Based on this, a disocclusion-type aware hole filling method is proposed. Disocclusions in the virtual image are identified and divided into two types. FG-BG disocclusions are divided into several small holes through the texture-based depth image preprocessing. BG-BG disocclusions are filled by extracting the corresponding FG objects in the reference image and replacing them with the surrounding BG textures. Further, additional post-processing is introduced to handle the holes that still exist after the merging. Subjective and objective comparison results demonstrate that the proposed method produces trusted contents under different scenes and baseline distances. The proposed method still has some limitation in the classification of disocclusions. Due to the complexity of the scene, there would be a situation where two types of disocclusions are combined. The current classification based on pixel distribution may not achieve optimal results. Our future work will consider a more intelligent and robust classification method to achieve disocclusion segmentation and sequential processing. Reducing the computation cost is also an important concern. This paper does not explore the temporal correlation between the adjacent frames. In this case, there may be a common region in the disocclusions exposed in the adjacent two frames. Since our method is to process each frame separately, it may produce different filling results and cause flicker between frames. For the rendering of single-frame images, such as stereoscopic image pairs, this phenomenon can be ignored. But for the rendering of virtual view sequences, it is necessary to maintain temporal consistency. Compared to the disocclusion filling in each frame, considering the movement of FG object in the time domain and reusing the inpainted BG information can reduce computational complexity and maintain the temporal consistency [24]. In future work, we will consider using global optimization to recover the texture of disocclusions, so that the flicker between frames can be improved. Further, deep learning techniques have proved successful in target detection and image inpainting, which can provide some inspiration for FG object detection and real-time disocclusion filling [9, 27, 30].

Acknowledgments The authors would like to thank the Interactive Visual Media Group at Microsoft Research for making the MSR 3D Video Dataset publicly available.

Funding This work was funded by the National Major Project of Scientific and Technical Supporting Programs of China during the 13th Five-year Plan Period (grant numbers 2017YFC0109702, 2017YFC0109901, and 2018YFC0116202).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Ahn I, Kim C (2013) A novel depth-based virtual view synthesis method for free viewpoint video. IEEE Trans Broadcast 59(4):614–626. https://doi.org/10.1109/tbc.2013.2281658
2. Chen W-Y, Chang Y-L, Lin S-F, Ding L-F, Chen L-G (2005) Efficient depth image based rendering with edge dependent depth filter and interpolation. In: 2005 IEEE International Conference on Multimedia and Expo. IEEE, pp 1314–1317
3. Chen XD, Liang HT, Xu HY, Ren SY, Cai HY, Wang Y (2019) Artifact handling based on depth image for view synthesis. Appl Sci-Basel 9(9):19. https://doi.org/10.3390/app9091834
4. Chen XD, Liang HT, Xu HY, Ren SY, Cai HY, Wang Y (2020) Virtual view synthesis based on asymmetric bidirectional DIBR for 3D video and free viewpoint video. Applied Sciences-Basel 10(5):19. https://doi.org/10.3390/app10051562
5. Criminisi A, Perez P, Toyama K (2004) Region filling and object removal by exemplar-based image inpainting. IEEE Trans Image Process 13(9):1200–1212. https://doi.org/10.1109/tip.2004.833105
6. Daribo I, Saito H (2011) A novel Inpainting-based layered depth video for 3DTV. IEEE Trans Broadcast 57(2):533–541. https://doi.org/10.1109/tbc.2011.2125110
7. de Oliveira AQ, Walter M, Jung CR (2018) An artifact-type aware DIBR method for view synthesis. IEEE Signal Process Lett 25(11):1705–1709. https://doi.org/10.1109/lsp.2018.2870342

8. Deng ZM, Wang MJ (2018) Reliability-based view synthesis for free viewpoint video. Appl Sci-Basel 8(5): 15. https://doi.org/10.3390/app8050823
9. Dong GS, Huang WM, Smith WAP, Ren P (2020) A shadow constrained conditional generative adversarial net for SRTM data restoration. Remote Sens Environ 237:14. https://doi.org/10.1016/j.rse.2019.111602
10. Fehn C (2004) Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV. In: Woods AJ, Merritt JO, Benton SA, Bolas MT (eds) Stereoscopic displays and virtual reality systems Xi, Proceedings of the Society of Photo-Optical Instrumentation Engineers (Spie), vol 5291. Spie-Int Soc Optical Engineering, Bellingham, pp 93–104. https://doi.org/10.1117/12.524762
11. Greene N, Kass M, Miller G (1993) Hierarchical Z-buffer visibility. In: Proceedings of the 20th annual conference on Computer graphics and interactive techniques. ACM, pp 231–238
12. Han DX, Chen H, Tu CH, Xu YY (2018) View synthesis using foreground object extraction for disparity control and image inpainting. J Vis Commun Image Represent 56:287–295. https://doi.org/10.1016/j.jvcir.2018.10.004
13. Kao CC (2017) Stereoscopic image generation with depth image based rendering. Multimed Tools Appl 76(11):12981–12999. https://doi.org/10.1007/s11042-016-3733-3
14. Kim S, Jang J, Lim J, Paik J, Lee S (2018) Disparity-selective stereo matching using correlation confidence measure. J Opt Soc Am A-Opt Image Sci Vis 35(9):1653–1662. https://doi.org/10.1364/josaa.35.001653
15. Lei JJ, Zhang CC, Wu M, You L, Fan KF, Hou CP (2017) A divide-and-conquer hole-filling method for handling disocclusion in single-view rendering. Multimed Tools Appl 76(6):7661–7676. https://doi.org/10.1007/s11042-016-3413-3
16. Li S, Zhu C, Sun MT (2018) Hole filling with multiple reference views in DIBR view synthesis. IEEE Trans Multimedia 20(8):1948–1959. https://doi.org/10.1109/tmm.2018.2791810
17. Liang ZY, Shen JB (2020) Local semantic Siamese networks for fast tracking. IEEE Trans Image Process 29:3351–3364. https://doi.org/10.1109/tip.2019.2959256
18. Lie WN, Hsieh CY, Lin GS (2018) Key-frame-based background sprite generation for hole filling in depth image-based rendering. IEEE Trans Multimedia 20(5):1075–1087. https://doi.org/10.1109/tmm.2017.2763319
19. Lim H, Kim YS, Lee S, Choi O, Kim JD, Kim C (2011) Bi-layer inpainting for novel view synthesis. In: 2011 18th IEEE International Conference on Image Processing. IEEE, pp 1089–1092
20. Liu SH, Paul A, Zhang GC, Jeon G (2015) A game theory-based block image compression method in encryption domain. J Supercomput 71(9):3353–3372. https://doi.org/10.1007/s11227-015-1413-0
21. Liu W, Ma LY, Qiu B, Cui MY, Ding JW (2017) An efficient depth map preprocessing method based on structure-aided domain transform smoothing for 3D view generation. PLoS One 12(4):20. https://doi.org/10.1371/journal.pone.0175910
22. Luo GB, Zhu YS (2017) Foreground removal approach for hole filling in 3D video and FVV synthesis. IEEE Trans Circuits Syst Video Technol 27(10):2118–2131. https://doi.org/10.1109/tcsvt.2016.2583978
23. Luo G, Zhu Y, Li Z (2016) Zhang L a hole filling approach based on background reconstruction for view synthesis in 3D video. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, In, pp 1781–1789
24. Luo GB, Zhu YS, Weng ZY, Li ZT (2020) A Disocclusion Inpainting framework for depth-based view synthesis. IEEE Trans Pattern Anal Mach Intell 42(6):1289–1302. https://doi.org/10.1109/tpami.2019.2899837
25. Mori Y, Fukushima N, Yendo T, Fujii T, Tanimoto M (2009) View generation with 3D warping using depth information for FTV. Signal Process-Image Commun 24(1–2):65–72. https://doi.org/10.1016/j.image.2008.10.013
26. Muddala SM, Sjostrom M, Olsson R (2016) Virtual view synthesis using layered depth image generation and depth-based inpainting for filling disocclusions and translucent disocclusions. J Vis Commun Image Represent 38:351–366. https://doi.org/10.1016/j.jvcir.2016.02.017
27. Nazeri K, Ng E, Joseph T, Qureshi F, Ebrahimi M (2019) EdgeConnect: generative image inpainting with adversarial edge learning. arXiv preprint arXiv:190100212
28. Ndjiki-Nya P, Koppel M, Doshkov D, Lakshman H, Merkle P, Muller K, Wiegand T (2011) Depth image-based rendering with advanced texture synthesis for 3-D video. IEEE Trans Multimedia 13(3):453–465. https://doi.org/10.1109/tmm.2011.2128862
29. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics 9(1):62–66
30. Rahmatov N, Paul A, Saeed F, Hong WH, Seo H, Kim J (2019) Machine learning-based automated image processing for quality management in industrial internet of things. Int J Distrib Sens Netw 15(10):11. https://doi.org/10.1177/1550147719883551

31. Scharstein D, Pal C (2007) Learning conditional random fields for stereo. In: 2007 Ieee Conference on Computer Vision and Pattern Recognition, Vols 1–8, New York. IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp 1688−+
32. Shao F, Lin WC, Fu RD, Yu M, Jiang GY (2017) Optimizing multiview video plus depth retargeting technique for stereoscopic 3D displays. Opt Express 25(11):12478–12492. https://doi.org/10.1364/oe.25.012478
33. Shen JB, Jin XG, Mao XY, Feng JQ (2006) Completion-based texture design using deformation. Vis Comput 22(9–11):936–945. https://doi.org/10.1007/s00371-006-0079-2
34. Shen JB, Jin XF, Zhou CA, Wang CCL (2007) Gradient based image completion by solving the Poisson equation. Comput Graph-UK 31(1):119–126. https://doi.org/10.1016/j.cag.2006.10.004
35. Shen JB, Wang DP, Li XL (2013) Depth-aware image seam carving. IEEE T Cybern 43(5):1453–1461. https://doi.org/10.1109/tcyb.2013.2273270
36. Sun W, Au OC, Xu L, Li Y, Hu W (2012) Novel temporal domain hole filling based on background modeling for view synthesis. In: 2012 19th IEEE International Conference on Image Processing. IEEE, pp 2721–2724
37. Tam WJ, Alain G, Zhang L, Martin T, Renaud R (2004) Smoothing depth maps for improved steroscopic image quality. In: Three-Dimensional TV, Video, and Display III, 2004. International Society for Optics and Photonics, pp 162–173
38. Tech G, Chen Y, Muller K, Ohm JR, Vetro A, Wang YK (2016) Overview of the multiview and 3D extensions of high efficiency video coding. IEEE Trans Circuits Syst Video Technol 26(1):35–49. https://doi.org/10.1109/tcsvt.2015.2477935
39. Wang WG, Shen JB (2018) Deep visual attention prediction. IEEE Trans Image Process 27(5):2368–2378. https://doi.org/10.1109/tip.2017.2787612
40. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612. https://doi.org/10.1109/tip.2003.819861
41. Wang LH, Huang XJ, Xi M, Li DX, Zhang M (2010) An asymmetric edge adaptive filter for depth generation and hole filling in 3DTV. IEEE Trans Broadcast 56(3):425–431. https://doi.org/10.1109/tbc.2010.2053971
42. Wang WG, Shen JB, Yu YZ, Ma KL (2017) Stereoscopic thumbnail creation via efficient stereo saliency detection. IEEE Trans Vis Comput Graph 23(8):2014–2027. https://doi.org/10.1109/tvcg.2016.2600594
43. Wang WG, Shen JB, Ling HB (2019) A deep network solution for attention and aesthetics aware photo cropping. IEEE Trans Pattern Anal Mach Intell 41(7):1531–1544. https://doi.org/10.1109/tpami.2018.2840724
44. Zhang L, Tam WJ (2005) Stereoscopic image generation based on depth images for 3D TV. IEEE Trans Broadcast 51(2):191–199. https://doi.org/10.1109/tbc.2005.846190
45. Zhang L, Zhang L, Mou XQ, Zhang D (2011) FSIM: a feature similarity index for image quality assessment. IEEE Trans Image Process 20(8):2378–2386. https://doi.org/10.1109/tip.2011.2109730
46. Zhang L, Shen Y, Li HY (2014) VSI: a visual saliency-induced index for perceptual image quality assessment. IEEE Trans Image Process 23(10):4270–4281. https://doi.org/10.1109/tip.2014.2346028
47. Zhu C, Li S (2016) Depth image based view synthesis: new insights and perspectives on hole generation and filling. IEEE Trans Broadcast 62(1):82–93. https://doi.org/10.1109/tbc.2015.2475697
48. Zitnick CL, Kang SB, Uyttendaele M, Winder S, Szeliski R (2004) High-quality video view interpolation using a layered representation. ACM Trans Graph 23(3):600–608. https://doi.org/10.1145/1015706.1015766