Muddala, S.; Sjöström, M.; Olsson, R., " Virtual view synthesis using layered depth image generation and depth-based inpainting for filling disocclusions and translucent disocclusions," *Journal of Visual Communication and Image Representation*, vol. 38, pp. 351-366, July 2016

http://dx.doi.org/10.1016/j.jvcir.2016.02.017

Virtual View Synthesis Using Layered Depth Image Generation and Depth-Based Inpainting for Filling Disocclusions and Translucent Disocclusions

Suryanarayana M. Muddala[a], Mårten Sjöström[a,*], Roger Olsson[a]

[a] *Dept. of Information and communication systems, Mid Sweden University, 85170, Sundsvall, Sweden*

Abstract

View synthesis is an efficient solution to produce content for 3DTV and FTV. However, proper handling of the disocclusions is a major challenge in the view synthesis. Inpainting methods offer solutions for handling disocclusions, though limitations in foreground-background classification causes the holes to be filled with inconsistent textures. Moreover, the state-of-the art methods fail to identify and fill disocclusions in intermediate distances between foreground and background through which background may be visible in the virtual view (translucent disocclusions). Aiming at improved rendering quality, we introduce a layered depth image (LDI) in the original camera view, in which we identify and fill occluded background so that when the LDI data is rendered to a virtual view, no disocclusions appear but views with consistent data are produced also handling translucent disocclusions. Moreover, the proposed foreground-background classification and inpainting fills the disocclusions with neighboring background texture consistently. Based on the objective and subjective evaluations, the proposed method outperforms the state-of-the art methods at the disocclusions.

Keywords: View synthesis, depth-image based rendering, image inpainting, texture synthesis, hole filling, disocclusions, translucent disocclusions, layered depth image

1. Introduction

Three Dimensional Video (3DV) technologies offer an immersive user experience. In principle, 3DV in the stereo format includes two videos of the same scene but from slightly different viewpoints. The two views are then presented to the left and right eyes through separate channels. With the development of 3D display technology, auto stereoscopic displays are now available in the market [1]. In contrast to the stereo displays and current cinema technology, auto stereoscopic displays create depth impression without any additional eye wear. Free viewpoint television (FTV) is among the emerging applications of 3D video,

[*] Corresponding author
Email address: marten.sjostrom@miun.se (Mårten Sjöström)

which has even surpassed popularity of 3DTV in the topic of 3D video and content generation [2]. 3DTV provides the depth-impression using stereo, whereas FTV offers the viewer to look around the scene. In order to provide such an experience, these technologies require a number of camera views captured from different viewpoints. Usually the capturing and transmission of a big number of views is not a feasible solution, and so view synthesis is employed as an alternative solution.

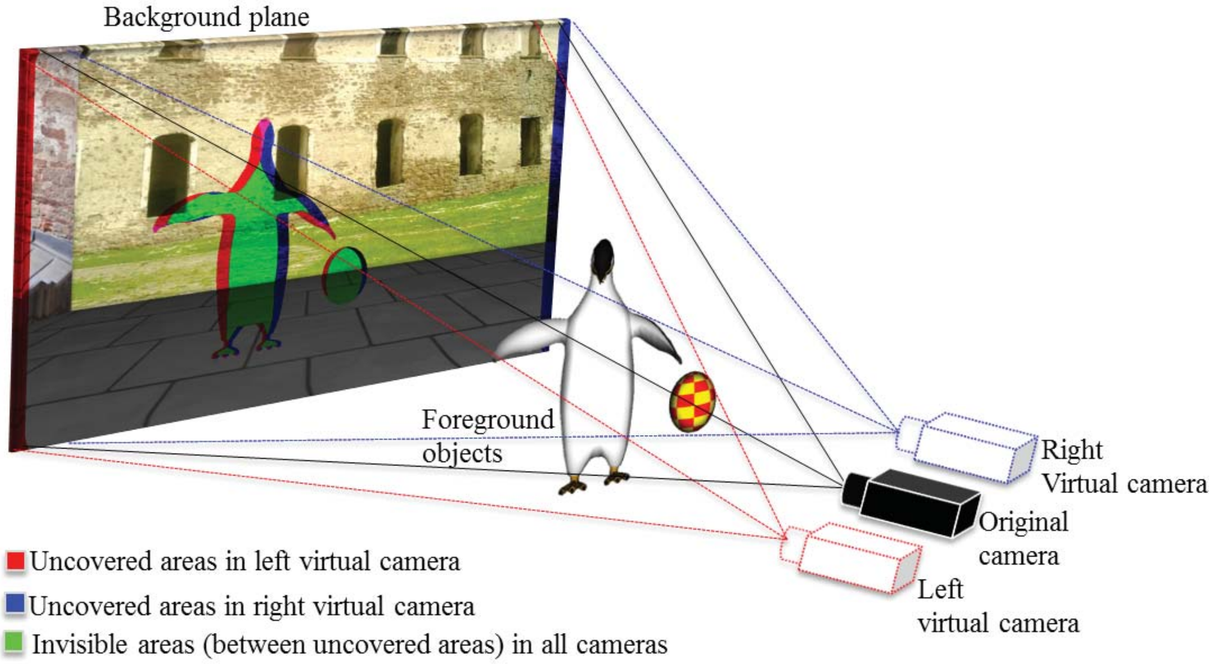

Background plane
Foreground objects
Right Virtual camera
Original camera
Left virtual camera
🟥 Uncovered areas in left virtual camera
🟦 Uncovered areas in right virtual camera
🟩 Invisible areas (between uncovered areas) in all cameras

Figure 1: Illustration of view-extrapolation of virtual cameras from original camera. Uncovered areas are invisible in original camera but visible in the virtual cameras, so they should be filled.

Technically, virtual views are generated by using the principle of perspective geometry. An illustration of virtual view generation is shown in Fig. 1. In this respect, video-plus-depth (V+D) and multiview video-plus-depth (MVD) are common 3DV formats to efficiently transfer 3D video to the end user [3, 4]. Later, additional views can be produced at the end user side by considering the requirement of the display. Depth-Image-Based Rendering (DIBR) is a widely used method to render a new viewpoint, using a texture and depth information [5]. We define the inputs to the DIBR as the original views and outputs as the virtual views (aka rendered views). The virtual view consists of the warped texture and depth images. True depth at virtual view is the depth obtained by some capture method.

As a consequence of DIBR, the warped images exhibit artifacts, namely *ghosting* and *uncovered areas* (aka *holes*), which affect the visual quality severely (see Fig. 2(a)). Ghosting artifacts are a mixture of colors at the edges that are projected into the neighboring objects due to the depth and texture misalignment at

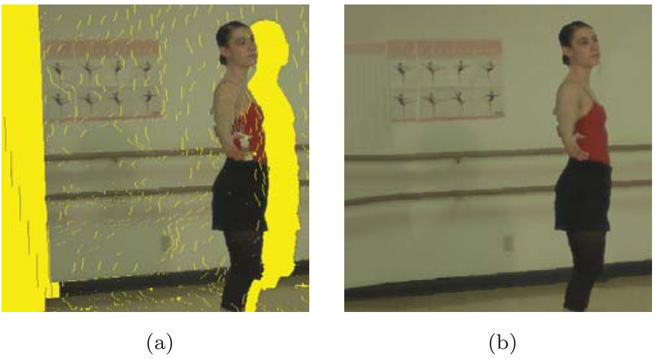


Figure 2: View synthesis results: (a) Synthesized image with holes (in yellow color); (b) Proposed synthesized image.

the depth discontinuities. *Holes* can be classified into *cracks*, *disocclusions* and *out-of-field areas*. Cracks are usually 1 to 2 pixel-wide missing information, which appear due to rounding the projected pixel position to the nearest integer. In the case when cracks appear on an object and that object causes an occlusion, texture from occluded objects seeps through the cracks and so we call them *translucent cracks*. Image processing techniques with different complexity levels have been developed to handle these artifacts, yet handling disocclusions and out-of-field areas are still challenging. Disocclusions are the result of baseline (distance between the cameras) and a distinct change in the depth between neighboring pixels, normally occurring at object borders. In this context, we define foreground (FG) to be the part of the scene closer to the camera that occludes other objects and background (BG) to be the part of the scene farther from the camera which is partially occluded by foreground in the original image. Out-of-field areas are caused in the virtual camera views due to the lack of information at the image boundaries. Generally holes get larger when the baseline increases. But large holes are not only involved in a large-baseline setup like FTV. They also appear when there is a large depth discontinuity between FG and BG, which increases the importance of efficient handling of the holes.

Besides the above mentioned artifacts, there exists another type of artifact, which we define as the *translucent disocclusion*. Translucent disocclusions differ from common disocclusions by exposing texture information that is present behind the BG. Translucent disocclusions are only visible when the depth has three or more layers, and where there are occlusions present between layers. An example of this case is shown in Fig. 3(a) in the area near the woman's hand. Note in Fig. 3(c) that the wall texture is seeping through the woman's body. The magnitude of the disturbance created by translucent disocclusion artifacts depends on the placement of occlusions in the scene and the occlusion area. These artifacts should be addressed especially if they have a large size, since that severely affects the perceived visual quality. Conventional *hole-filling* methods aim at the disocclusion artifacts and leave translucent disocclusion unattended, which consequently degrades the virtual view quality (see Fig. 4(b) to (f)).

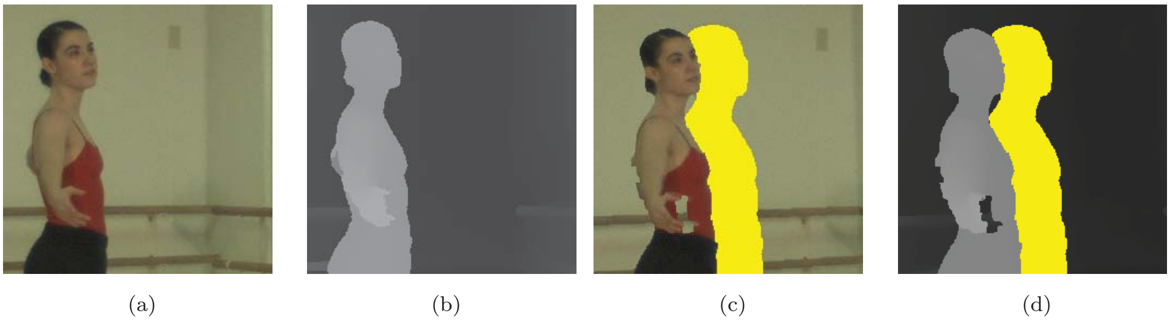


Figure 3: Illustration of translucent disocclusion: (a) Texture image; (b) Depth image; (c) Warped image; (d) Warped depth image (disocclusions in yellow color).

1.1. RelatedWork

Several methods have been proposed in the literature to reduce the artifacts in the virtual view. Ghosting artifacts are reduced by first detecting the depth discontinuities by estimating the FG and BG contribution and removing pixels in the vicinity of the discontinuities [11, 12]. We can classify the hole-filling methods into two categories: *Pre-processing* and *Post-processing*.

Pre-processing: Cracks are commonly reduced by using backward warping approach [13]. In this approach, first the depth maps are warped to the virtual view point using forward warping. Then the warped depth map is smoothed using bilateral filter or asymmetric filters to preserve the edges. Finally, the texture of the warped image pixel is located in the original image by using filtered depth and backward warping.

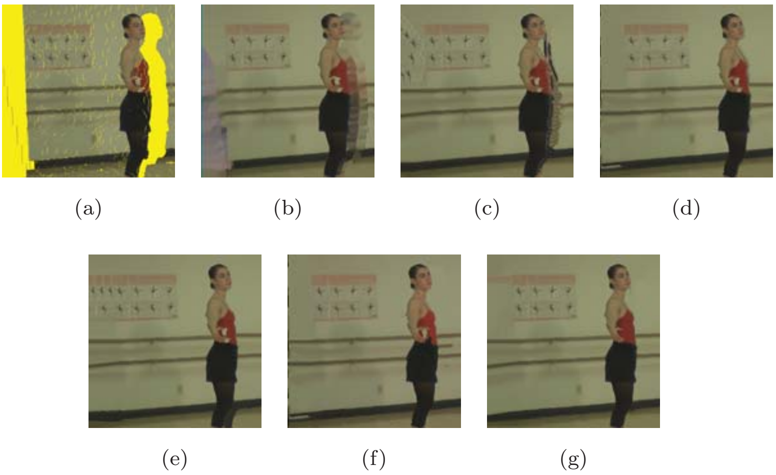


Figure 4: View synthesis results "Ballet" (frame28): (a) Warped image with holes (in yellow color); (b) VSRS method [6]; (c) Daribo et al. method [7]; (d) Gautier et al. method [8]; (e) Ahn et al. method [9]; (f) Wolinski et al. method [10]; (g) Proposed method.

Pre-processing methods only reduce cracks and so post-processing steps are necessary to handle the rest of the holes [14].

Post-processing: the original view is projected into the virtual view and then hole-filling techniques are applied to reduce the artifacts using *interpolation*, *extrapolation* and *image inpainting* methods. Cracks are commonly avoided by interpolating the nearest pixel values [12]. Disocclusions are often filled with neighboring warped images in the case of view-interpolation (i.e., producing an intermediate view using its neighbors), whereas in the view-extrapolation case (i.e., using single view), the disocclusions problems need to be handled with only the information available in the original view. Interpolation and extrapolation are simple approaches that fill the holes with the texture from neighboring pixels around the holes [6, 15]. These approaches give acceptable results when the holes are small and less noticeable. However, when the holes are large, the result looks unnatural due to stretching of the border pixels into the holes.

Image inpainting has been an active research topic in computer vision for a decade. Recently, image inpainting techniques have been applied to disocclusion problems. The methods used in the virtual view synthesis are classified into *diffusion* and *exemplar-based* methods. Diffusion methods are basically structural inpainting methods, which identify the linear structures in the neighborhood and propagate them into the missing regions [16, 17]. View synthesis reference software (VSRS) also applied a diffusion method to fill the holes [6]. Undesirably, holes are partially filled with the FG due to diffusing the boundary information into the hole. To avoid FG propagation in the diffusion methods, Oh et al. [18] changed the hole pixels values near the FG boundary with the BG values on the hole boundary. Although diffusion methods propagate linear structures into missing regions, blurring and structure discontinuities arise for large holes due to not considering the edge information during the process.

The exemplar-based methods use both structural and texture properties to fill the holes [19, 20]. Criminisi et al. proposed an efficient filling order to fill the holes [20]. The filling order is computed by assigning a priority value to each possible target patch along the hole boundary. This priority value is determined by using a confidence and data term. The confidence term is defined as percentage of non-hole pixels in a patch and the data term is a measure of structure details, which gives isophote direction. In this context, the highest priority value patch is named target patch and the patch used to fill the target patch is called source or best patch. Once the target patch is selected, pixels belonging to the hole in the patch are filled by searching for a source patch in the neighborhood; this step is referred to as patch matching and filling. However, applying the Criminisi et al. method to fill disocclusions can cause FG to propagate into the disocclusion regions, since the filling order is computed on the entire hole boundary irrespective of the FG and BG.

The exemplar-based methods have the potential to reconstruct the missing structure in the image, and so have an advantage over the diffusion methods. Hence the following inpainting methods have used the depth, various priority choices and patch matching strategies in the inpainting process. We further classify the depth dependent inpainting methods into three categories, which use true depth at virtual view position, pre-processed depth and warped depth.

Certain inpainting methods [7, 8] have used the true depth at the virtual view point (measured or separately estimated) in the computation of priority to select the target patch and in the patch matching. A structure tensor based data term is used in the priority computation in order to propagate the linear structures into the missing regions [8]. However, these methods still result noticeable artifacts around FG objects in the virtual view. This is because the method by Daribo et al. uses a depth regularity term in the computation of priority, which could select the patches that contains FG. The other reason is that neither method considers a FG-BG classification in the patch matching procedure. Therefore, the target patch could contain FG data when the inpainting process is close to FG objects [7, 8]. To avoid such FG propagation, the source region is classified into FG and BG to find the best patch in [21]. Since the FG and BG classification method [22] works well only when there are two distinguishable regions, resulting inpainted images suffer from background leaking and jaggedness artifacts. *Background leaking* is a propagation of inconsistent BG with neighborhood, and *jaggedness* is inconsistent textures along the boundary of FG objects [23]. Further, filling order is computed using Markov random fields (MRF) and belief propagation in [24]. Moreover, the assumption of the true depth is not a feasible in current 3DTV settings because the depth at the virtual view point needs to be estimated.

As the true depth at the virtual point is not available, warped depth map is filled using pre-processing and inpainting techniques [25]. However, the hole-filling with FG texture is possible due to the filling priority. Further, to provide interview consistency, a view synthesis method is presented in [10]. In this method, first they classified the depth into several FG-BG layers and then located the disocclusions and inpainted them using the filled depth map. Finally, inpainted layers are projected into the virtual views. Although, this method creates consistency between the views, the filled regions are inconsistent with neighboring BG as a consequence of the depth classification used in the inpainting and the priority computations. I.e. FG and BG layers are classified as one layer instead of two when there are depth discontinuities whose depth values are less than a given tolerance value. And as a result, holes are filled with FG. Moreover, the virtual view possesses translucent artifacts and FG propagation.

To reduce the FG propagation into the hole, a priority is computed on the BG boundary using edge detection. Furthermore, the source region is classified into FG-BG using MRF and random walk segmentation [26], [27]. However, artifacts still exist around FG objects. The hole-filling is further improved by using a priority and FG-BG classification derived from the hole boundary depth values [9]. However, holes are filled with inconsistent textures due to the following reasons: (i) filling from one specific direction (right-to-left for disocclusions on the right of an object, or vice versa), which result in a structure propagation restricted to that direction, (ii) the FG-BG classification, which causes background leaking and jaggedness as mentioned earlier in this section. In another attempt, the curvature data term, depth guided priority and the

FG-BG classification in the patch matching are used in the inpainting [28]. This method also suffers from the background leaking and jaggedness. These problems are partially addressed later by removing the FG in the patch matching [23].

On the other hand, disocclusions are reduced by capturing the information related to the occluded areas in the original camera. *Occlusion* is defined as a region where the BG data is covered by the FG in the original view. If a FG is occluded by yet another FG, we define them as *overlaid occlusions*. From the perspective of the virtual camera, the occlusions and overlaid occlusions in the original view correspond to the disocclusions and the translucent disocclusions in the virtual view. A method to represent the V+D format with the occlusion data is Layered Depth Images (LDI) [29]. The advantage of LDV is that rendering of a virtual view can be efficiently executed without additional handling of occlusion problems, since the occlusion data is already captured using multiple cameras. In a simplest form of LDV i.e. when a single occlusion layer exists, the virtual view quality reduces when the size of the overlaid occlusion is large. Moreover, in contrast to MVD, the rendered virtual view quality reduces from the reference view point due to resampling issues. However, by knowing the number of viewpoints to be rendered, a multiple LDV or MVD data are sufficient to get desired quality without occlusion problems. Due to the fact that more sources are required to capture the occlusion data, a method is presented to generate the occlusion data from a stereo pair [30]. Various methods to acquire the occlusion data from different camera arrangements are presented in [31]. However, occlusion layers generations from a single V+D is not explicitly presented. Moreover, the rendering methods could cause the following problems in the virtual views: i) translucent disocclusions, since the rendering methods use a single occlusion layer, ii) filling holes with inconsistent textures because the structure details are not considered while inpainting.

The above methods show some improvements with new advances in handling disocclusion problems, however the results of the mentioned methods exhibit the translucent disocclusions problems (see Fig. 4 near woman's hand). Moreover, the above approaches have a problem in filling the holes with consistent BG from neighboring regions, which consequently affects the visual quality of the virtual view.

1.2. Overview of proposed view synthesis method

In this paper, we propose a view synthesis method for extrapolating virtual views. The two major novelties of the method are in the layered depth image (LDI) generation and FG-BG classification for the depth-based inpainting. In contrast to the traditional view synthesis methods, we introduce the LDI using V+D in the original view, in which we inpaint the introduced layers with data from the original view. The layers are created such that LDI data belongs to the occluded BG, so that when the LDI data is rendered to a new virtual view, no disocclusions appear, but instead views with consistent data are produced. In similar to [10], inpainting is performed in the original view, since the data in the virtual view consists of other artifacts, which affects the inpainting process. Unlike [10], the layers formed in the proposed method

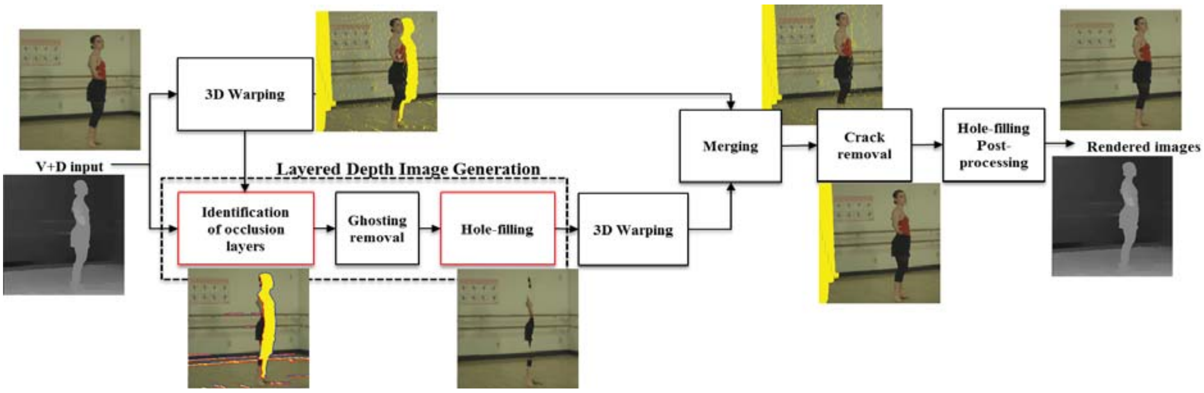

V+D input
3D Warping
Layered Depth Image Generation
Identification of occlusion layers
Ghosting removal
Hole-filling
3D Warping
Merging
Crack removal
Hole-filling Post-processing
Rendered images

Figure 5: Proposed view synthesis method using layered depth image generation and depth-based inpainting.

are depending on the occlusions and inpainted with the proposed FG-BG classification and inpainting.

This paper is organized as follows: Section 2 explains the proposed view synthesis method and depth-based inpainting. The experimental set-up and the evaluation criteria are described in Section 3. The results and analysis of the proposed method are presented in Section 4. Finally, Section 5 discusses the limitations and concludes this paper.

2. Proposed view synthesis method

The proposed view synthesis method for extrapolating the virtual views is shown in Fig. 5. At first, the LDI data is generated using the 3D warping information and the depth values at the depth discontinuities. In addition to LDI data, depth thresholds for the FG-BG classification are determined during the layer formation step. Then, the introduced layers are inpainted using the thresholds. In the next step, the inpainted LDI data is warped to the virtual view positions. Subsequently the LDI data is merged and the cracks are filled in the warped image. Finally, proposed inpainting method is applied without depth classification to fill the remaining holes, such as out-of-field areas.

2.1. Warping

A virtual view is generated by the geometrical 3D warping [5]. At first the original image pixels are projected into the 3D world points using depth and camera parameters. Then these 3D world points are projected into the virtual view image plane. The extrinsic and intrinsic camera parameters together define a projection matrix given by

$$\mathbf{P} = \mathbf{K} \cdot \mathbf{I}_{3\times 4} \cdot \left(\begin{array}{c|c} \mathbf{R} & \mathbf{t} \\ \hline \mathbf{0} & 1 \end{array}\right) \tag{1}$$

that projects the 3D world point of the scene into a position $\mathbf{m}$ on the image plane of a camera. $\mathbf{I}_{3\times 4}$ is the identity matrix $\mathbf{R}$ is a rotation matrix and $\mathbf{t}$ a translation vector, which together represent the extrinsic parameters. **K** is the intrinsic matrix, which describe the focal length, image center and camera pixels sizes. The projection matrix **P**, The projection matrix is invertible and is given by

$$\mathbf{P}^{-1} = \left(\begin{array}{c|c} \mathbf{R}^{-1}\mathbf{K}^{-1} & -\mathbf{R}^{-1}\mathbf{t} \\ \hline \mathbf{0} & 1 \end{array}\right), \tag{2}$$

Note that both 3D world points and 2D image points are represented in homogenous coordinates. The general mathematical equation for a pixel position in the virtual view image is given by:

$$z_{\mathrm{v}}\mathbf{m}_{\mathrm{v}} = z_{\mathrm{o}}\mathbf{P}_{\mathrm{v}}\mathbf{P}_{\mathrm{o}}^{-1}\left(\mathbf{m}_{\mathrm{o}}\right) \tag{3}$$

where, z is a depth value, subscripts v and o corresponds to virtual and original views, respectively.

2.2. Occlusion Layers Generation

The generation of the LDI is based on identifying the occlusion layers in the original view. As the occlusions belong to the BG, we identify the occlusions by using the depth values and the projected locations of pixels in the virtual view.

2.2.1. Identification of Occlusion Layers

The basic idea for locating the occlusions is based on the projected pixels positions difference in similar to the traditional DIBR methods [25, 30]. In addition, we use the location of the FG and a depth threshold that derived from pixels at the depth discontinuity to locate occlusion layers.

Given the properties of DIBR, disocclusions occur between pixels that are neighboring in the original image, have a distinct difference in depth and that become separated in the warped image. For example, the occlusion behind a FG-object in the original image becomes visible as a disocclusion in the virtual image. The disocclusion appear to the right of the FG-object (see Fig. 6), when the virtual image is rendered from a camera positioned to the right of the original camera.

Let I_{o} and I_{v} be original and virtual view images, respectively. Furthermore, let the warping operation between these be represented by $\Gamma : \mathbf{I}_{\mathrm{o}} \rightarrow \mathbf{I}_{\mathrm{v}}$. We define a depth discontinuity pixel pair (DDPP) as a neighboring pixel pair $\{\mathbf{f}_{\mathrm{o}}, \mathbf{p}_{\mathrm{o}}\}$ that satisfies the following condition:

$$\mathrm{DDDP} = \{\mathbf{f}_{\mathrm{o}}, \mathbf{p}_{\mathrm{o}}\}\,;\, \|\mathbf{d}_{\mathrm{v}}\|_2 > \eta, \tag{4}$$

where η is an occlusion threshold, and $\mathbf{d}_{\mathrm{v}}$ is a displacement vector given by

$$\mathbf{d}_{\mathrm{v}} = \begin{pmatrix} d_x \\ d_y \end{pmatrix} = \mathbf{f}_{\mathrm{v}} - \mathbf{p}_{\mathrm{v}}, \tag{5}$$

where $\mathbf{f}_{\mathrm{v}}$, $\mathbf{p}_{\mathrm{v}}$ are projected pixels in I_{v}:

$$\begin{aligned} \Gamma :&\mathbf{f}_{\mathrm{o}} \rightarrow \mathbf{f}_{\mathrm{v}}, \\ \Gamma :&\mathbf{p}_{\mathrm{o}} \rightarrow \mathbf{p}_{\mathrm{v}}. \end{aligned} \tag{6}$$

Note that the above definition is presented within the context of a right warped image. It is straightforward to change Eq. (4) for a left warped image. A summary of the notation used in this paper is presented in Table 1.

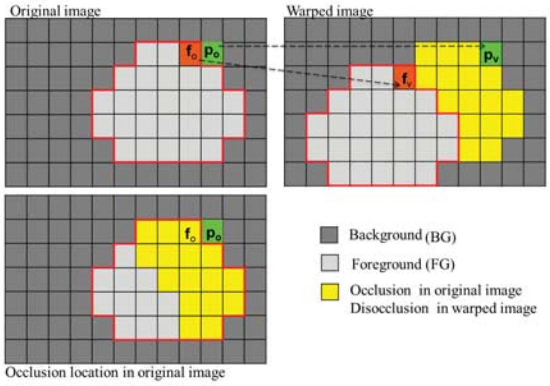

Original image
Warped image
f_o
p_o
p_v
f_v
f_o
p_o
Background (BG)
Foreground (FG)
Occlusion in original image
Disocclusion in warped image
Occlusion location in original image

Figure 6: Identification of occlusion location

Next, the pixels in the DDPP are labeled as FG and BG using their respective depth values. When there are multiple depth layers, DDPPs might overlap and produce inconsistent FG-BG labelling. Meaning that a pixel labeled BG in one DDPP would be labeled FG in another DDPP. To circumvent this we apply a labelling refinement step that selects either BG or FG based on the projected pixels' positions difference.

Once the DDPPs are identified using Eq. (4) and labeled, the occlusion in the original view is located by using horizontal and vertical displacements and a depth value. The pixels that belong to the occlusions are identified at DDPP $\{\mathbf{p}_{\mathrm{o}}, \mathbf{f}_{\mathrm{o}}\}$ using a depth threshold as follows:

Table 1: Notation

Variable	Description
$\mathbf{f}$, $\mathbf{p}$, $\mathbf{q}$	Pixels in an image
I	Texture image
Z	Depth map
Ω	Hole region
$\delta\Omega$	Hole boundary
Φ	Source region $(\mathrm{I} - \Omega)$
Ψ	2D patch
$\Psi_{\mathbf{p}}$	2D patch centered at $\mathbf{p}$

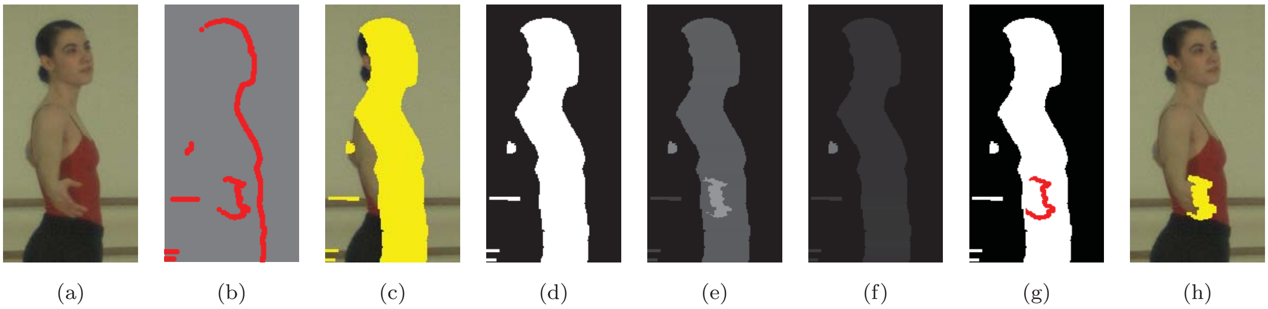


Figure 7: “Ballet” (frame28); (a) Original image; (b) Depth discontinuity pixels (in red color); (c) Occlusion in original image (in yellow color); (d) Occlusion mask; (e) Average depth threshold image; (f) BG depth threshold image; (g) Depth discontinuity pixels in former occlusion mask; (h) Overlaid occlusion.

$$\mathrm{H}\left(\mathbf{q}\right)=\begin{cases}1 & \text{if } (\mathrm{Z}(\mathbf{q})>T_1), \forall \mathbf{q}\in l;\\ 0 & \text{otherwise.}\end{cases} \tag{7}$$

where l is a line between pixels $\mathbf{p}_\mathrm{o}(x,y)$ and $\mathbf{p}_\mathrm{o}(x-dx,y-dy)$, H is an occlusion mask, T_1 is an occlusion depth threshold, which is an average of the depth discontinuity pixels depth values and is given as:

$$T_1=\frac{\mathrm{Z}(\mathbf{f}_\mathrm{o})+\mathrm{Z}(\mathbf{p}_\mathrm{o})}{2}, \tag{8}$$

From the definition of overlaid occlusions, we identify these in the original view by finding the presence of depth discontinuity pixel pairs (DDPP) in the former identified occlusion layer (see Fig. 7(g) and (h)). In this way, occlusion layers are formed in the LDI.

2.2.2. Occlusion depth threshold

As the occlusion layers (see example Fig. 7(c)) are part of the BG, they should be filled with consistent BG information. A single threshold is sufficient to classify FG and BG when there is an occlusion between two layers. In the overlaid occlusion case, a second threshold is required in order to distinguish the layers apart; if only one threshold is used, any information farther away than the FG would be acceptable, which may cause translucent disocclusion in the virtual views when the occlusions in the original views are filled with information from an inconsistent BG layer. The two depth thresholds are derived from the DDPP values in order to distinguish between FG and BG information. The disocclusions appear between the DDPPs, so selecting the threshold between DDPP allows differentiating FG and BG locally in an effective way. Another threshold that is based on BG depth values in DDPP allows identifying the BG in the overlaid occlusions. This threshold information is created in the form of images. We name the threshold images *average threshold image* Z_A and *BG threshold image* Z_B, respectively. The two threshold images are defined by extrapolating the depth values of the DDPPs at the occlusion. Meaning that Z_A and Z_B contain average DDPP depth values and BG depth values of DDPP, respectively. Thus the two depth thresholds are formed to facilitate the occlusion inpainting with consistent BG information. All pixels within the occlusion are set as functions of the DDPP depth values at $\{\mathbf{p}_\mathrm{o}, \mathbf{f}_\mathrm{o}\}$ as:

$$\left.\begin{aligned} \mathrm{Z}_A(\mathbf{q}) &= T_1, \\ \mathrm{Z}_B(\mathbf{q}) &= \mathrm{Z}(\mathbf{p}_\mathrm{o}), \end{aligned}\right\} \quad \forall \mathbf{q} \in l, \tag{9}$$

Note that pixels on different lines l a calculated from different DDPPs. Examples of Z_A and Z_B are shown in Fig. 7(e) and (f). The brightest region in Fig. 7(e) corresponds to the overlaid occlusion.

2.3. Ghosting removal

Ghosting artifact creates an annoying experience; it also affects the hole-filling process due to the mixed texture of FG and BG on the boundary of the hole. Consequently, holes are filled with inconsistent texture with BG. As the proposed method has already identified the FG and BG pixels, artifacts are simply reduced by removing two pixels along the BG side.

2.4. Proposed hole-filling method

The proposed hole-filling method is adopted from [28] and extended with the proposed depth constraints in the following steps: FG-BG boundary extraction, filling priority and patch matching and filling (see Fig. 8). Unlike the reference methods presented in the literature, the BG data for computing the filling priority and patch matching is derived from the proposed classification introduced in section 2.2.2. A depth threshold for classification is derived from the depth discontinuities values, so the proposed method ensures that the holes are filled from the BG.

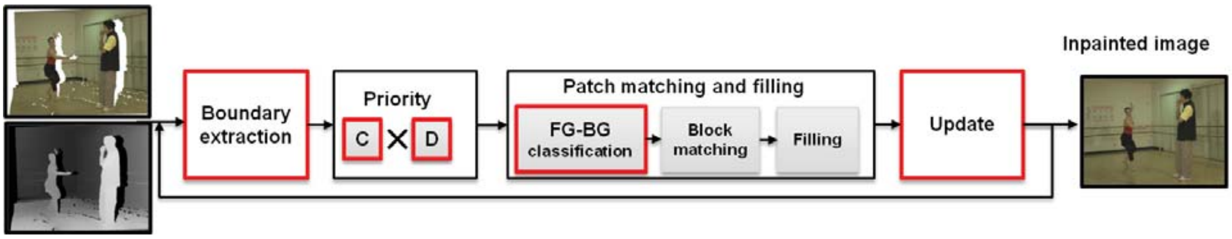
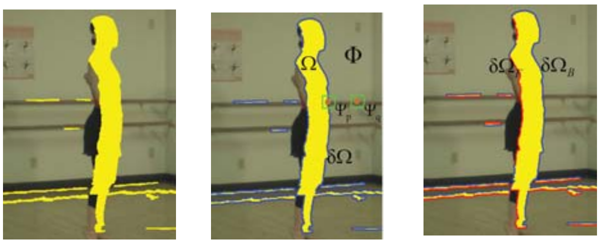


Figure 8: Proposed depth-based inpainting method.

2.4.1. FG-BG *boundary extraction*

Inpainted image quality is highly dependent on the filling order [20]. In the view synthesis context, the filling order should also take the FG information into consideration to avoid the filling from FG side. We classify the hole boundary into FG and BG using the depth threshold image Z_A. The BG boundary classification is as follows: first the hole boundary $\delta\Omega$ is obtained by convolving the hole with Laplacian kernel (see Fig. 9(b)). Then by utilizing a depth value from Z_A, the hole boundary is classified into the FG boundary $\delta\Omega_F$ and the BG boundary $\delta\Omega_B$ i.e., $\delta\Omega = \delta\Omega_F \cup \delta\Omega_B$ (see Fig. 9(c)).

$$\delta\Omega\left(\mathbf{p}\right) = \begin{cases} \delta\Omega_F & \text{if } \left(\mathrm{Z}\left(\mathbf{p}\right) > \max\left(\mathrm{Z}_A\left(\mathbf{q}\right)|_{\mathbf{q}\in\Psi_{\mathbf{p}}\cap\Omega}\right)\right); \\ \delta\Omega_B & \text{otherwise.} \end{cases} \tag{10}$$



(a) (b) (c)

Figure 9: Hole boundary extraction: (a) Texture image with holes; (b) Hole boundary with notations (boundary in blue color); (c) Depth guided boundary (FG and BG boundaries are in red and blue colors).

2.4.2. Filling Priority

The filling priority on the BG boundary $\delta\Omega_B$ is computed as the product of confidence and data terms:

$$P\left(\mathbf{p}\right) = C\left(\mathbf{p}\right) \cdot D\left(\mathbf{p}\right), \tag{11}$$

where, $C\left(\mathbf{p}\right)$ is the confidence term and $D\left(\mathbf{p}\right)$ is the data term at $\mathbf{p}$.

Depth-based confidence term: Confidence is one of the important terms that drives the filling inwards. Since the term computes the confidence irrespective of the FG and BG, it leads to inconsistent filling. To avoid this problem, unlike [9], the confidence term is only computed for the patches which have the BG information.

$$C\left(\mathbf{p}\right) = \frac{1}{|\Psi_{\mathbf{p}}|} \sum_{\mathbf{q}\in\Psi_{\mathbf{p}\cap\Phi}} C\left(\mathbf{q}\right), \tag{12}$$

$$C\left(\mathbf{q}\right) = \begin{cases} 0 & \text{if } \left(\mathrm{Z}\left(\mathbf{q}\right) > \max\left(\mathrm{Z}_A\left(\mathbf{q}\right)|_{\mathbf{q}\in\Psi_{\mathbf{p}}\cap\Omega}\right)\right); \\ 1 & \text{otherwise.} \end{cases} \tag{13}$$

Curvature data term: Although the isophote and structure tensor data terms aim at propagating the linear structure into the missing regions, holes are filled with inconsistent structures due to the selection of

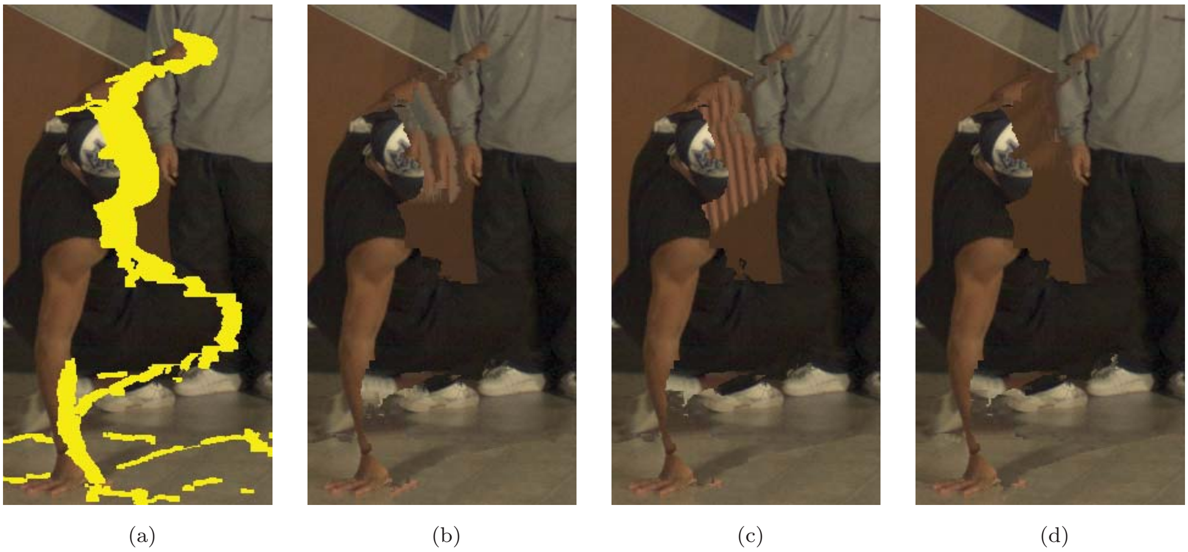


Figure 10: Data terms: (a) Texture image with holes; (b) Filled with Isophote data term [20]; (c) Filled with structure tensor data term [9]; (d) Filled with proposed curvature data term.

strong gradients (see Fig. 10 (b) and (c)). We have adopted the curvature data term to find the structure details in the neighborhood [28]. This is because curvature consists of the geometry of the isophote. According to [32], the depth in data term does not influence the image quality significantly; therefore the data term is only calculated on the texture image.

$$k_{\mathbf{p}} = \nabla \cdot \left(\frac{\nabla \mathrm{I}_{\mathbf{p}}}{|\nabla \mathrm{I}_{\mathbf{p}}|} \right), \tag{14}$$

$$D(\mathbf{p}) = 1 - |k_{\mathbf{p}}|\,. \tag{15}$$

where $k_{\mathbf{p}}$ is the curvature of the isophote through a pixel $\mathbf{p}$, $\nabla\cdot$ is the divergence at $\mathbf{p}$.

2.4.3. Patch matching and filling

The source region is classified into FG and BG, as the FG-BG classification is crucial to fill the target patch with BG information. The threshold for classification is selected from the threshold images Z_A and Z_B. Also the FG pixels are removed from the target patch during block matching to avoid jaggedness [23].

$$\Phi_B = \Phi - \Phi_F, \tag{16}$$

$$T_U = \max \left(\mathrm{Z}_A(\mathbf{q})\,|_{\mathbf{q} \in \Psi_{\hat{\mathbf{p}}} \cap \Omega} \right), \tag{17}$$

$$T_L = \min \left(\mathrm{Z}_B \left(\mathbf{q} \right) |_{\mathbf{q} \in \Psi_{\hat{\mathbf{p}}} \cap \Omega} \right) - \chi, \tag{18}$$

where Φ_F is the FG source region, where the depth values are higher than the depth threshold T_U. In the case of the overlaid occlusions, as they occur between two different depth regions, the BG source region Φ_B is selected between the depth thresholds T_U and T_L. This is to avoid the patch selection from inconsistent BG regions. χ is a depth tolerance threshold, which allows to find a best match according to depth.

$$\Psi_{\hat{\mathbf{q}}} = \arg \min_{\Psi_{\mathbf{q}} \in \Phi_B} \left\{ SSD(\Psi_{\hat{\mathbf{p}}}, \Psi_{\mathbf{q}}) + \beta \cdot SSD(\mathrm{Z}_{\hat{\mathbf{p}}}, \mathrm{Z}_{\mathbf{q}}) \right\}, \tag{19}$$

where, $\Psi_{\hat{\mathbf{p}}}$, $Z_{\hat{\mathbf{p}}}$ are target texture and depth patches and $\Psi_{\hat{\mathbf{q}}}$ is source patch. SSD is sum of squared differences and β is a weighting coefficient to equalize the effect of the depth and texture. Similar to [28], holes in the depth map are also filled simultaneously along with the texture image with a weighted average of N patches.

2.4.4. Update

As the proposed filling process is iterative, the data term is filled with the best patch data. The filling region is updated such that the filled region becomes the source region for the next iteration and the confidence term is updated as follows:

$$C\left(\mathbf{q}\right) = C\left(\hat{\mathbf{p}}\right), \forall \mathbf{q} \in \Psi_{\hat{\mathbf{p}}} \cap \Omega. \tag{20}$$

Unlike the reference methods presented in the literature, each hole is filled separately using the proposed method when neighboring holes exist in the vicinity of a patch size. This is to allow the method to fill the holes consistently with its neighboring data.

2.5. Warping Layered Depth Image and Merging

After the occlusion layers in the original texture and depth images are filled, they are warped to the virtual view position. Then, warped layers are merged using Z-buffering (when two pixels are projected at the same location, the pixel closer to the camera is selected). Note that the overlaid occlusions in the depth images are filled using BG depth values at the depth discontinuity to ensure that the filled depth map does not create artifacts after projection.

2.6. Crack removal and Hole-filling post-processing

Both conventional and translucent disocclusions are efficiently handled by the proposed method. However, certain types of holes might still be present after a virtual view is produced by warping, making a post-processing hole-filling operation necessary. Such holes are mainly caused by the forward warping calculations and are manifested as the well-known cracks and out-of-field areas. Another type of holes may appear when a depth map with certain properties is being processed by the LDI hole-filling. Depth maps that have significantly varying values surrounding the occlusion area will exhibit minor depth discontinuities in the filled area, where inpainting results stemming from different surroundings coincide. These discontinuities also manifest, after warped view synthesis, as disocclusions with an extent that is slightly larger than cracks but significantly smaller than conventional disocclusions. To remove these artifacts in the virtual view, we apply the following steps in succession:

1. Crack filling: Holes which are smaller than patch areas are filled by simple BG propagation. Such small holes have no priority during inpainting, since they contain FG data, which implies that patches with no-priority are filled after all the remaining holes in the image are filled. This leads to inconsistent hole-filling, since there is a less possibility to find consistent texture to fill the holes.
2. Hole-filling post-processing: The remaining holes such as out-of-field areas and any other holes are filled with the proposed inpainting method operating without depth classification. Since there is no information about the FG at the borders of these holes and depth classification is redundant.

3. Test Arrangement and Evaluation Criteria

We used five MVD test sequences with various texture and depth characteristics to validate our proposed view synthesis method in different conditions and to compare the resulting visual quality with state-of-the-art methods. The sequences are “Ballet”, “Breakdancers”, “Lovebird1”, “Newspaper” from the 3DVC reference set and “Poznan Street” [33, 34, 35, 36]. Details of the sequence characteristics are presented in Table 2. The hole size in the table is given for the nearest rendered view and is computed as follows: first the percentage of pixels belonging to holes in each rendered image is computed; then the average percentage of the pixels belonging to holes over the sequence is calculated. All the test sequences are used in the V+D scenario, which means the V+D input data is used to produce extrapolated virtual views. The virtual views are rendered to virtual camera positions such that they match already available real camera views. This allows for the extrapolated virtual views to be compared to the ground truth depth and texture, which is a pre-requisite for full-reference objective metric evaluation.

Measuring inpainting visual quality in the rendered views is as complex as disocclusion problem, since it is an ill posed problem that has no unique solution and it require detailed investigation. Therefore, we adopted the commonly used evaluation methodology in the view synthesis context. We used both objective measurements and visual inspection to assess the quality of the proposed method. Peak Signal-to-Noise Ratio of the luminance component (YPSNR) and Mean Structural Similarity Index (MSSIM) are used as full-reference objective metrics. YPSNR measures the absolute difference whereas MSSIM corresponds to

Table 2: Test input data characteristics

Sequence Name	**Resolution**	**Frames**	**Camera arrangement**	**Texture Background**	**Depth Properties**	**Hole size**(Avg)
Ballet	1024x768	1-100	8 cameras with 30 cm spacing, 1D arc arrangement	Low structured	Large depth discontinuities and many layers	13.5%
Break dancers	1024x768	1-100	8 cameras with 20 cm spacing, 1D arc arrangement	Low structured	Small depth discontinuities and few layers	6.3%
Lovebird	1024x768	120-220	12 cameras with 3.5 cm spacing, 1D arrangement	High structured	Large depth discontinuities and few layers	4.3%
Newspaper	1024x768	1-100	9 cameras with 5 cm spacing, 1D arrangement	Medium structured	Small depth discontinuities and many layers	10%
Poznan street	1920x1088	150-250	9 cameras with 13.5 cm spacing, 1D arrangement	High structured	Small depth discontinuities and many layers	5%

the perceptual visual quality [37]. Both metrics are applied on the full image and not only on the pixels produced by the proposed inpainting method. Synthesized views are also presented, in addition to the objective metrics, to provide a means for qualitative visual comparison. Conclusions relative to those views drawn by expert viewers are presented as well.

We compare our results with the results from the state-of-the-art methods presented in Section 1.1 [6, 7, 8, 9, 10, 27]. Method VSRS used diffusion to fill holes in the warped image, whereas remaining methods used exemplar method to fill holes. However, not all functions from the reference methods are available, so that comparisons are performed using the available functions from the given reference methods. Methods Daribo et al., Gautier et al. and Ahn et al. are used in the traditional scenario, where hole-filling is applied after warping. Method Wolinski et al. applied hole-filling in the original image. Moreover, Wolinski et al. does not provide a solution to fill out-of-field areas and only aims for the view consistency. Hence, to have a fair comparison, unfilled areas in the warped image are filled using method Gautier et al., since Wolinski et al. used the inpainting method from Gautier et al. The proposed method is implemented in MATLAB, which is not the case for all of the reference methods. Thus the complete evaluation of the computational complexity is left for the future work.

There are several parameters involved in the classification and inpainting parts of the proposed method. We divide these into three experimental set-ups, each using a different set of key parameters for evaluation:

1. Translucent disocclusions
2. FG-BG classification
3. Depth-based inpainting

We will elaborate on the details of the above set-ups in the following three subsections.

3.1. Evaluation of translucent disocclusions

The occurrence of translucent disocclusions depends on the scene and the depth map characteristics. Thus we have selected a subset of the "Ballet" sequence, which suffers from translucent disocclusions. The chosen subset includes 5 frames from the camera view position 5, rendered to view 4 and view 7 positions. In the process of evaluation, we compare the virtual views that are filled with translucent disocclusion handling (TDH) and no translucent disocclusion handling (NTDH). TDH is defined as the process which identifies the overlaid occlusions and then fills them.

3.2. Evaluation of FG-BG classification

A subset of the "Ballet" sequence that possesses many depth layers is selected, in order to test the influence of FG-BG classification in the patch matching. The chosen subset includes 5 frames from the "Ballet" sequence view 5, rendered to view 3 and view 4 positions.

Although the test images are from the same sequence as in Section 3.1, effects from the depth classification and appearance of translucent disocclusions depend on the scene and characteristics of the depth image. Thus, different frames from various viewpoints are selected for the depth classification evaluation compared to the test images in Section 3.1. In this evaluation, we select the depth-classification techniques that are employed in the reference methods Ahn et al. and Choi et al., since the FG-BG classification is the control parameter. Depth classification evaluations are limited to methods Ahn et al. and Choi et al., since the methods VSRS, Daribo et al. and Gautier et al. have not used the depth classification in the inpainting process, and in the case of method Wolinski et al., functions are not available. Note that the FG-BG classification is the only control parameter in this comparison, so that the remaining parameters in the inpainting process are kept unchanged.

3.3. Evaluation of depth-based inpainting

The evaluation of depth-based inpainting in the context of view synthesis is divided into 1D and general modes since the reference method Wolinski et al. functions are not available in 1D case. The 1D mode is where original and virtual view cameras are arranged in x-axis, whereas in general mode the virtual view camera involves rotation. Worth noting that results from Gautier et al. method are not reported in 1D case for test sequence “Poznan Street” v3, since the reference method has limitation in filling the out-of-filed areas.

A set of 100 frames are selected from the test sequences for evaluation purposes. To make it a fair comparison, few frames from the sequence are excluded in the measurements, since the results from the reference method Gautier et al. show color artifacts and unfilled holes in the inpainted images. The discarded frames sequences are from “Ballet” v6, “Breakdancers” v6 and “Poznan Street” v3 respectively. The notation used for the rendered views e.g., v5→v4 indicates that the original camera view 5 is warped to the virtual camera view 4.

We have empirically selected the following parameter values for all sequences: In the occlusion identification step, we use $\eta = 5$ pixels in Eq. (4) for small base lines, because holes with the size less than a patch width affect the filling process in finding consistent textures. Moreover, selecting larger values of η requires an additional hole filling post-processing step after warping, since the holes less than the threshold are not located in the original view. In the patch matching step, a search region of 120x120 pixels and patch size of 11x11 pixels are selected. In the case of overlaid occlusion filling, a patch size of 9x9 pixels is used based on the assumption that small areas can be effectively filled with a small patch size. The influence of search region and patch size is presented in [9]. If the search region has enough number of best patches, we set $N = 5$, otherwise $N = 1$. The depth tolerance parameter controls the range of depth values in the block matching search. Selection of the depth tolerance is depending on the properties of the depth image. Larger values of χ allows holes to be filled with absolute background, that can cause translucent problems. On the other hand, smaller χ values lead to inconsistent hole-filling, when the search region has not enough depth values, thus the depth tolerance is set to $\chi = 5$ in Eq. (18). Weighting coefficient $\beta = 3$ is used in Eq. (19), similar to the previous depth-inpainting method that gives similar weight as texture in the block matching step. The number of occlusion layers depends on occlusions in the original view; however, the number of layers inpainted by the proposed method is set to two. A detailed analysis on these parts is left for the future work.

4. Results and Analysis

4.1. Evaluation of translucent disocclusions

The objective evaluation (average YPSNR and MSSIM over selected frames) of translucent disocclusions are presented in Table 3. YPSNR and MSSIM measurements consistently demonstrate that the objective quality is improved by the translucent disocclusion handling. The improvements are very small in the range of 1/100, since the disocclusions occupy only a very small portion of the image. Subjective results in Fig. 11(b) and (c) show the difference in the case of TDH and NTDH. Results look unnatural with NTDH because the BG is seeping through FG. Results highlight the importance of translucent disocclusions handling for improving the rendered image quality. As the occurrence of this problem depends on the scene and the depth map characteristics, any sequence that would have these properties will benefit by identifying and inpainting them.

Table 3: Translucent disocclusion objective quality evaluation

Ballet Seq	**YPSNR**		**MSSIM**	
	TDH	NTDH	TDH	NTDH
v5→v4	**32.13**	32.05	**0.8727**	0.8724
v5→v7	**27.84**	27.44	**0.8172**	0.8158

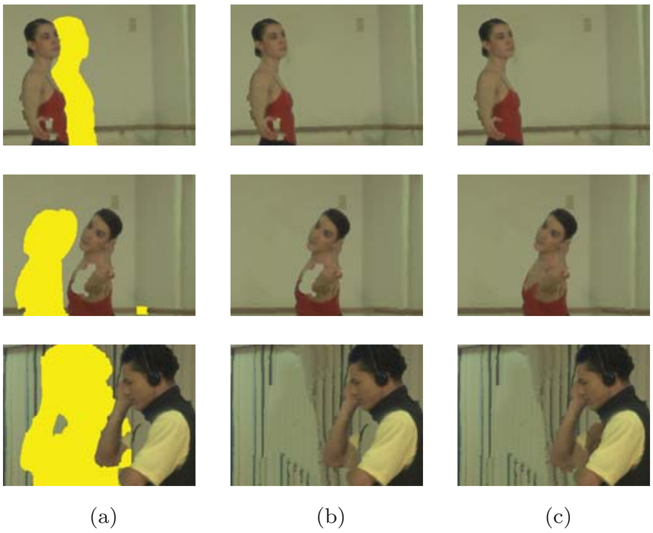


Figure 11: Translucent disocclusion visual quality evaluation "Ballet" (frame28 v5→v4) in the first row and (frame19 v5→v7) on last two rows: (a) Warped image with disocclusions; (b) NTDH; (c) TDH.

4.2. Evaluation of FG-BG classification

Results for the objective quality (average YPSNR and MSSIM over selected frames) assessment of the FG-BG classification are presented in Table 4. The objective measurements YPSNR and MSSIM consistently demonstrate that the proposed FG-BG classification improves the objective quality. Further, the subjective results for visual comparison are shown in Fig. 12. When filling with the reference methods, the structure is disconnected and inconsistent with the neighboring BG (see Fig. 12(b) and (c) at the woman's right leg and right side of the man at the wooden bar). Fig. 12(d) shows the proposed classification performs better than the reference classification methods by propagating the consistent BG. Understandably, to make the disocclusions removal plausible; a spatially consistent FG-BG is required.

Table 4: Depth classification objective quality evaluation (Note that classification from Ahn et al., Choi et al. only compared not the entire method).

Ballet Seq	YPSNR			MSSIM		
	Proposed	Ahn	Choi	Proposed	Ahn	Choi
v5→v4	**32.09**	31.59	31.58	**0.8731**	0.8719	0.8689
v5→v3	**27.99**	27.46	27.35	**0.8402**	0.8375	0.8308

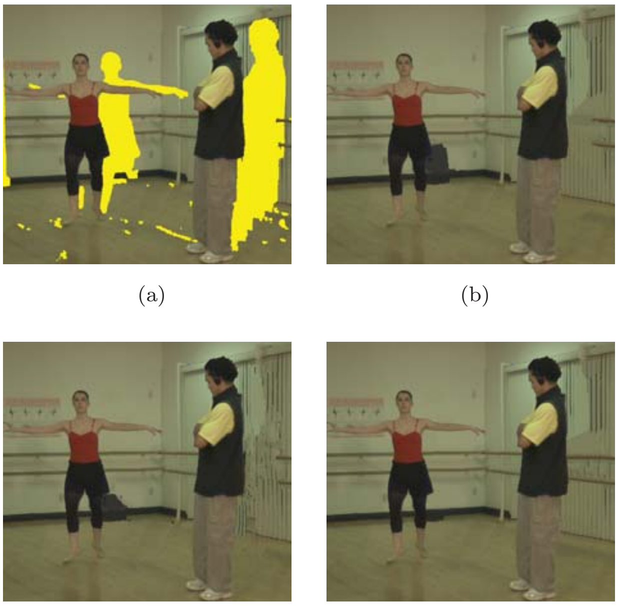

(a) (b) (c) (d)

Figure 12: Depth classification visual quality evaluation: (a) Texture image with holes ("Ballet" frame37 v5→v3); (b) Ahn et al. classification; (c) Choi et al. classification; (d) Proposed method.

4.3. Evaluation of depth-based inpainting

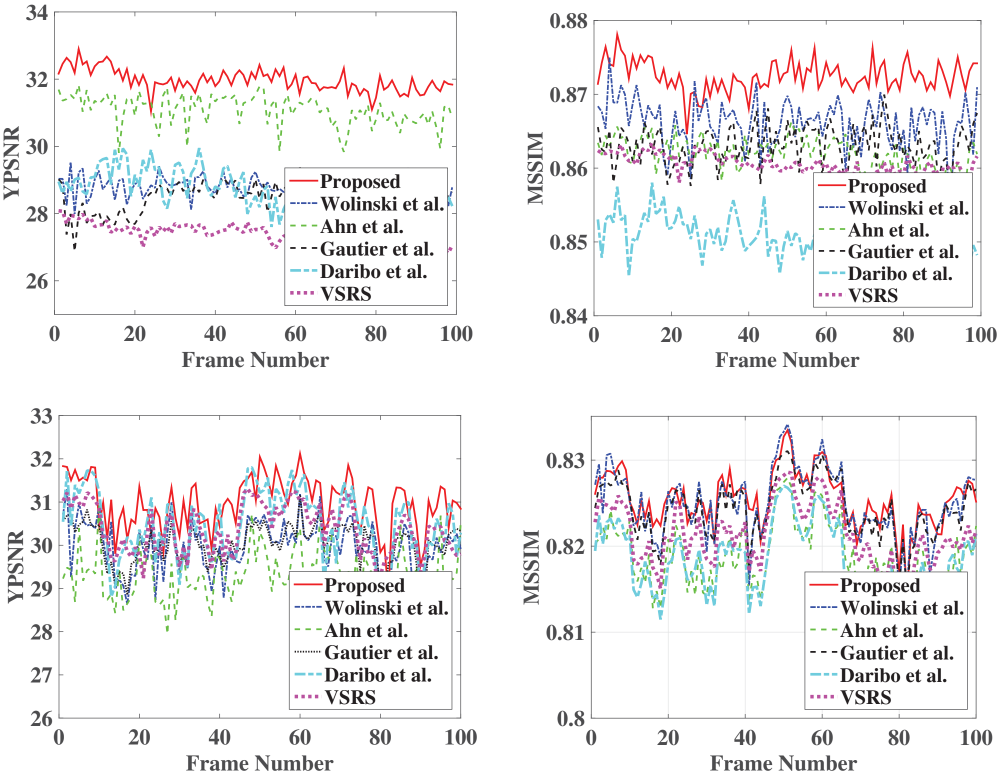
The objective evaluation results of depth-based inpainting quality in the general mode and 1D mode are presented in Tables 5 and 6 respectively. In addition to the average YPSNR and MSSIM over selected frames in Tables 5 and 6, the objective metrics of 100 frames are presented to show the consistency over the frames (see Fig. 13). Results in Table 5 demonstrate that the proposed method outperforms the state-of-the

Table 5: Objective Quality Evaluation in General mode

Sequence	YPSNR						MSSIM					
	Proposed	Wolinski	Ahn	Gautier	Daribo	VSRS	Proposed	Wolinski	Ahn	Gautier	Daribo	VSRS
Ballet v5→v4	**31.95**	28.75	31.08	28.43	28.68	27.34	**0.8724**	0.8631	0.8662	0.8620	0.8504	0.8605
Ballet v5→v6	**29.01**	26.80	27.71	27.63	26.52	26.37	**0.8495**	0.8439	0.8395	0.8335	0.8274	0.8493
Break dancers v5→v4	**30.97**	29.96	29.28	30.03	30.46	30.76	**0.8255**	0.8244	0.8250	0.8188	0.8182	0.8213
Break dancers v5→v6	**31.38**	30.90	30.98	31.08	30.54	31.26	0.8214	0.8220	0.8184	0.8211	0.8154	**0.8238**

art reference methods. In 1D mode, the proposed method performed better than depth-based inpainting methods Daribo et al. and Ahn et al., but the objective quality was slightly reduced compared to VSRS and Gautier et al. methods. Two reasons can explain the results from the depth-based inpainting methods comparison: i) the depth-based inpainting method Gautier et al. uses the true depth during the inpainting process, whereas the proposed method operates on general settings that the warped depth is estimated along with the texture. Despite that, holes were filled with an inconsistent texture in our proposed method due to insufficient BG information in the neighborhood. This problem usually occurs when the depth map quality is poor. Especially the "Newspaper" sequence has that characteristic. The virtual view has less occlusions and insufficient BG data in the patch matching step. ii) Major portion of the holes in the sequences "Newspaper" and "Poznan Street" are out-of-field-areas. As our approach does not apply the depth classification in filling out-of-field-areas, achieved results are expected. The same reasoning is valid for method VSRS. Although the holes are not filled with consistent textures using VSRS, it shows slightly better objective results in 1D case because the holes are small and exist between similar textures. This result implies that the diffusion can be a valid choice for filling small holes and homogenous regions. It is worth noting that objective measurements for ill posed problems such as the inpainting quality are still a challenging problem with no particular existing metric. Thus, subjective results are presented for the visual comparison.

The subjective results for visual inspection are shown in Fig. 14, 15 and 16. The results in Fig. 14(f) show the proper propagation of the structure into the holes and ensures spatial consistency with neighboring textures compared to the stat-of-the art methods in Fig. 14(b) to (e). As mentioned in the literature study,



(a) (b)

Figure 13: View synthesis with depth-based inpainting objective quality evaluation for 'Ballet" (v5→v4) in first row and "Breakdancers" (v5→v4) in second row: (a) YPSNR; (b) MSSIM.

VSRS method results show blurring artifacts and FG propagation in the filled region, Daribo et al. and Gautier et al. methods shows inconsistent structure propagation and jaggedness problem along FG objects and Ahn et al. method shows inconsistent structure propagation (see Fig. 14). The results in Fig. 15 further demonstrate the reconstruction of the consistent BG texture even when the data is missing between the two FG objects (see missing data between the small poles in Fig. 15(e)), whereas reference methods fill those regions with inconsistent BG information (see Fig. 15(b) to (d)).

In the general mode and even for the large disocclusions, subjective test results of the proposed method clearly demonstrate superior performance over the reference methods, especially at the translucent disocclusions areas in "Ballet" sequence (see column Fig. 16(a)). Although Wolinski et al. has slightly filled the translucent disocclusions, still uncovered parts are left unfilled (see row 6 in Fig. 16(a)). Moreover, when the disocclusions occur between different FG objects, the reference methods fail to reconstruct the missing data while the proposed method using LDI ensures that there are no disocclusions (see row 7 in Fig. 16(b) at man's hand and at woman's legs). Furthermore, holes are filled with FG for small holes when

Table 6: Objective Quality Evaluation in 1D mode

Sequence	YPSNR					MSSIM				
	Proposed	Ahn	Gautier	Daribo	VSRS	Proposed	Ahn	Gautier	Daribo	VSRS
Poznan street v5→v3	28.07	27.77	-	27.67	**28.20**	0.8341	0.8287	-	0.8279	**0.8393**
Lovebird1 v6→v4	**25.14**	25.04	25.13	24.94	24.94	0.8606	0.8572	0.8595	0.8551	**0.8655**
Newspaper v4→v6	23.24	20.55	**23.60**	23.13	23.32	0.8462	0.8330	0.8426	0.8367	**0.8525**

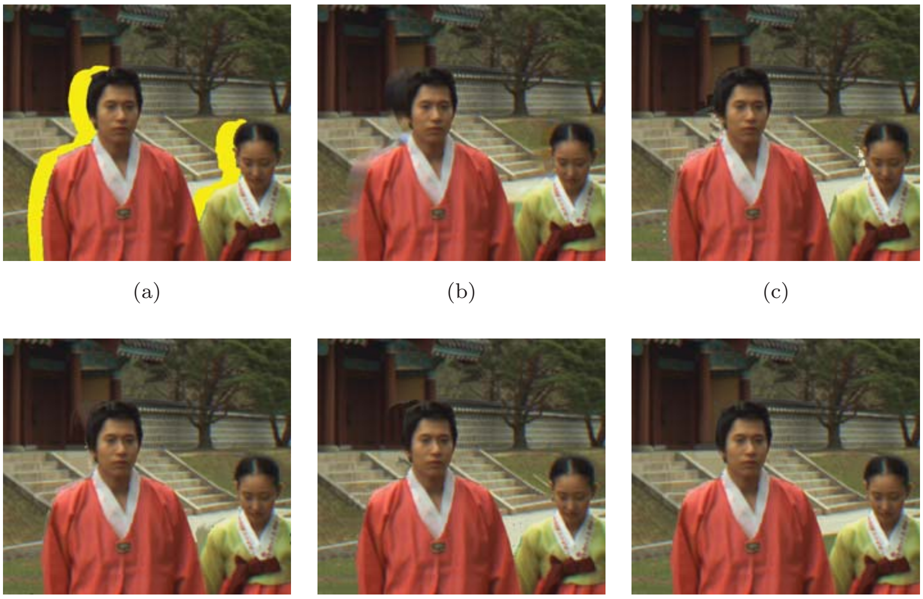

(a) (b) (c) (d) (e) (f)

Figure 14: Depth-based inpainting visual quality evaluation: "Lovebird" (frame192): (a) warped image with holes (in yellow color); (b) VSRS; (c) Daribo et al.; (d) Gautier et al.; (e) Ahn et al.; (f) Proposed method.

the depth difference between two layers is small (see example in Fig. 16(c) the holes near man with hat in the BG). This signifies that the importance of the LDI in filling the holes with consistent neighboring BG. The proposed method results in Fig. 16(c) and (d) show the consistent BG structures filling when there are strong gradients around holes. In the same scenario, the reference methods fill BG structure with strongest gradients, which looks unnatural and inconsistent with the neighboring BG (see Fig. 16(c) and (d) columns

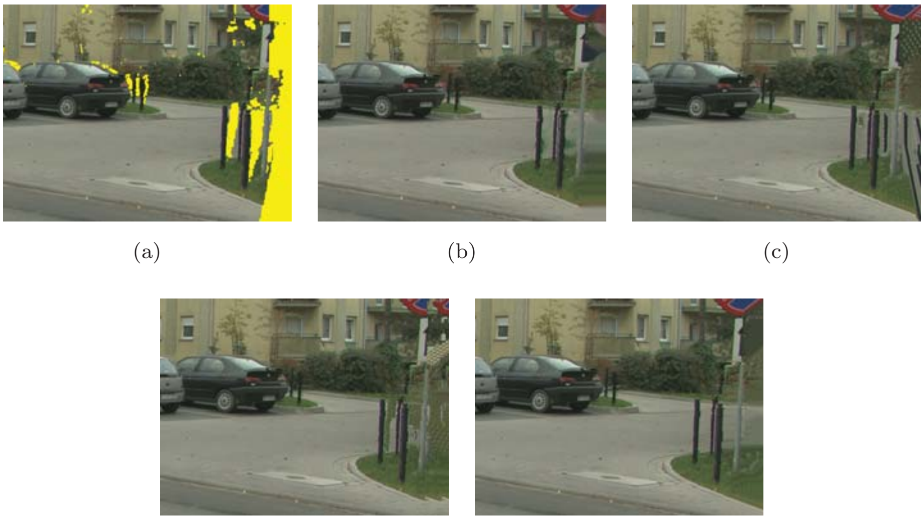


Figure 15: Depth-based inpainting visual quality evaluation: "Poznan street" (frame192): (a) warped image with holes (in yellow color); (b) VSRS; (c) Daribo et al.; (d) Ahn et al.; (e) Proposed method.

from row 2 to 5). In addition to the proposed method quality results, an insight about performance of the proposed method is given. The proposed method computation cost is high, it takes about 35 minutes to render a single frame "Ballet" v5→v4 with an unoptimized code in MATLAB. However, implementing the proposed method in C++ with optimization can greatly reduce the run time cost. Further, supplementary results can be found at paper web page [1].

5. Conclusions

We proposed a new view synthesis method with depth-based inpainting in order to improve the rendered view quality. The proposed layered depth image allows producing virtual views without disocclusion problems. Further, the hole-filling based on the proposed foreground-background classification in the original views, ensures the filling of occlusions consistent with the background. The proposed depth-based inpainting method computes the priority on the background with curvature data information, in order to propagate consistent structure details with respect to its neighboring background. The objective and subjective test results demonstrate that the proposed view synthesis method produces high quality virtual views and also consistently fills the translucent disocclusions for the tested sequences. The results signified the importance of the foreground-background classification in improving the virtual view image quality. The proposed

[1] http://www.miun.se/en/research/centers-and-institutes/stc/research-groups/realistic-3d/publications/virtual-view-synthesis-using-layered-depth-image-generation-and-depth-based-inpainting

method has some limitations in the translucent disocclusion when the background data is limited due to the poor quality of the depth map. Handling of out-of-field areas is another challenge when foreground objects are presented at the image boundaries because the proposed method is not applying any classification while filling out-of-field areas. Introducing foreground-background and dynamic classification during this step could be a way to reach consistent information to be filled in these areas. Moreover, inpainting holes in spatial domain can cause temporal consistency. Our future work will consider temporal consistency in the virtual views by using temporal information to build a sprite and inpainting the holes. Reducing the computational complexity is also a major concern. Reusing the inpainted information in the temporal domain instead of inpainting every frame can reduce the complexity and create temporal consistency as well.

Acknowledgment

The authors would like to thank the reviewers for their comments and suggestions, which has enabled us to improve the presentation of this research. This work has in part been supported by grants 20120328 and 20140200 of the KK Foundation, Sweden.

References

[1] O. Schreer, P. Kauff, T. Sikora, 3D Video communication Algorithms, concepts and real-time systems in human centered communication, John & Wiley Sons Ltd., 2005.

[2] M. Tanimoto, Ftv (free viewpoint television) for 3d scene reproduction and creation, in: Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on, 2006, pp. 172–172. doi:10.1109/CVPRW.2006.84.

[3] A. Smolic, K. Muller, P. Merkle, P. Kauff, T. Wiegand, An overview of available and emerging 3d video formats and depth enhanced stereo as efficient generic solution, in: Picture Coding Symposium, 2009. PCS 2009, 2009, pp. 1–4. doi:10.1109/PCS.2009.5167358.

[4] F. Zilly, C. Riechert, M. Ḿuller, P. Eisert, T. Sikora, P. Kauff, Real-time generation of multi-view video plus depth content using mixed narrow and wide baseline, Journal of Visual Communication and Image Representation 25 (4) (2013) 632 – 648. doi:10.1016/j.jvcir.2013.07.002.

[5] C. Fehn, Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv, Proc. SPIE Stereoscopic Displays and Virtual Reality Systems XI (2004) 93–104doi:10.1117/12.524762.

[6] Report on experimental framework for 3d video coding, ISO/IEC JTC1/SC29/WG11 MPEG2010/N11631, guangzhou, China (Oct. 2010).

[7] I. Daribo, H. Saito, A novel inpainting-based layered depth video for 3dtv, IEEE Transactions on Broadcasting 57 (2) (2011) 533–541. doi:10.1109/TBC.2011.2125110.

[8] J. Gautier, O. L. Meur, C. Guillemot, Depth-based image completion for view synthesis, in: 3DTV conference, 2011, pp. 1–4. doi:10.1109/3DTV.2011.5877193.

[9] I. Ahn, C. Kim, A novel depth-based virtual view synthesis method for free viewpoint video, IEEE Transactions on Broadcasting 59 (4) (2013) 614–626. doi:10.1109/TBC.2013.2281658.

[10] D. Wolinski, O. L. Meur, J. Gautier, 3d view synthesis with inter-view consistency, in: ACM Multimedia, 2013. doi:10.1145/2502081.2502175.

[11] K. Muller, A. Smolic, K. Dix, P. Merkle, P. Kauff, T. Wiegand, View synthesis for advanced 3d video systems, EURASIP Journal on Image and Video processingdoi:10.1155/2008/438148.

[12] M. Sjöström, P. Härdling, L. S. Karlsson, R. Olsson, Improved depth-image-based rendering algorithm, in: 3DTV Conference: The True Vision Capture, Transmission and Display of 3D Video (3DTV-CON), 2011, pp. 1–4. doi:10.1109/3DTV.2011.5877183.

[13] D. Tian, P. L. Lai, P. Lopez, C. Gomila, View synthesis techniques for 3d video, Proc. SPIE 7443 (2009) 10.1117/12.829372.

[14] L. Zhang, W. J. Tam, Stereoscopic image generation based on depth images for 3d tv, in: IEEE Transactions on Broadcasting, 2005, pp. 191–199. doi:10.1109/TBC.2005.846190.

[15] S. M. Muddala, M. Sjöström, R. Olsson, Edge-preserving depth-image-based rendering method, in: International Conference on 3D Imaging 2012 (IC3D), 2012. doi:10.1109/IC3D.2012.6615113.

[16] M. Bertalmio, G. Sapiro, V. Caselles, C. Ballester, Image inpainting, in: Proceedings of ACM Conf. Comp. Graphics (SIGGRAPH), 2000, pp. 417–424. doi:10.1145/344779.344972.

[17] A. Telea, An image inpainting technique based on the fast marching method, J. Graphics, GPU, Game Tools 9 (1) (2004) 23–34. doi:10.1145/344779.344972.

[18] K. J. Oh, S. Yea, Y. S. Ho, Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3-d video, in: Proceedings of the 27th Conference on Picture Coding Symposium, 2009, pp. 1–4.

[19] A. Efros, T. Leung, Texture synthesis by non-parametric sampling, in: International Conference on Computer Vision, 1999, pp. 1033–1038. doi:10.1145/344779.344972.

[20] A. Criminisi, P. Pérez, K. Toyama, Region filling and object removal by exemplar-based image inpainting, IEEE Transactions on Image Processing 13 (2004) 1200–1212. doi:10.1109/TIP.2004.833105.

[21] L. Ma, L. Do, P. H. N. de With, Depth-guided inpainting algorithm for free-viewpoint video, in: Image Processing (ICIP), 2012 19th IEEE International Conference on, 2012, pp. 1721–1724. doi:10.1109/ICIP.2012.6467211.

[22] N. Otsu, A threshold selection method from gray level histograms, IEEE Trans. Systems, Man and Cybernetics 9 (1979) 62–66. doi:10.1109/TSMC.1979.4310076.

[23] S. M. Muddala, M. Sjöström, R. Olsson, Depth-based inpainting for disocclusion filling, in: 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2014, pp. 1–4. doi:10.1109/3DTV.2014.6874752.

[24] J. Habigt, K. Diepold, Image completion for view synthesis using markov random fields and efficient belief propagation, in: Image Processing (ICIP), 2013 20th IEEE International Conference on, 2013, pp. 2131–2134. doi:10.1109/ICIP.2013.6738439.

[25] V. Jantet, C. Guillemot, L. Morin, Joint projection filling method for occlusion handling in depth-image-based rendering, in: 3D Research, 2011. doi:10.1007/3DRes.04(2011)4.

[26] H. Lim, Y. S. Kim, S. Lee, O. C., J. Kim, C. Kim, Bi-layer inpainting for novel view synthesis, in: Image Processing (ICIP), 2011 18th IEEE International Conference on, 2011, pp. 1089–1092. doi:10.1109/ICIP.2011.6115615.

[27] S. Choi, B. Ham, K. Sohn, Space-time hole filling with random walks in view extrapolation for 3d video, IEEE Transactions on Image Processing 22 (6) (2013) 2429–2441. doi:10.1109/TIP.2013.2251646.

[28] S. M. Muddala, R. Olsson, M. Sjöström, Disocclusion handling using depth-based inpainting, in: The Fifth International Conferences on Advances in Multimedia(MMEDIA), 2013.

[29] J. W. Shade, S. J. Gortler, L. W. He, R. Szelisk, Layered depth images, in: Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '98, 1998, pp. 231–242. doi:10.1145/280814.280882.

[30] B. Barenbrug, R.-P. M. Berretty, R. K. Gunnewiek, Robust image, depth, and occlusion generation from uncalibrated stereo, Proc. SPIE 6803 (2008) 68031J–68031J–8doi:10.1117/12.765508.

[31] B. Bartczak, P. Vandewalle, O. Grau, G. Briand, J. Fournier, P. Kerbiriou, M.Murdoch, M. Mller, R. Goris, R. Koch, R. van der Vleuten, Display-independent 3d-tv production and delivery using the layered depth video format, IEEE

Transactions on Broadcasting 57 (2) (2011) 477–490. doi:10.1109/TBC.2011.2120790.

[32] S. M. Muddala, R. Olsson, M. Sjöström, Depth-included curvature inpainting for disocclusion filling in view synthesis, International Journal On Advances in Telecommunications 6 (3 & 4) (2013) 132–142.

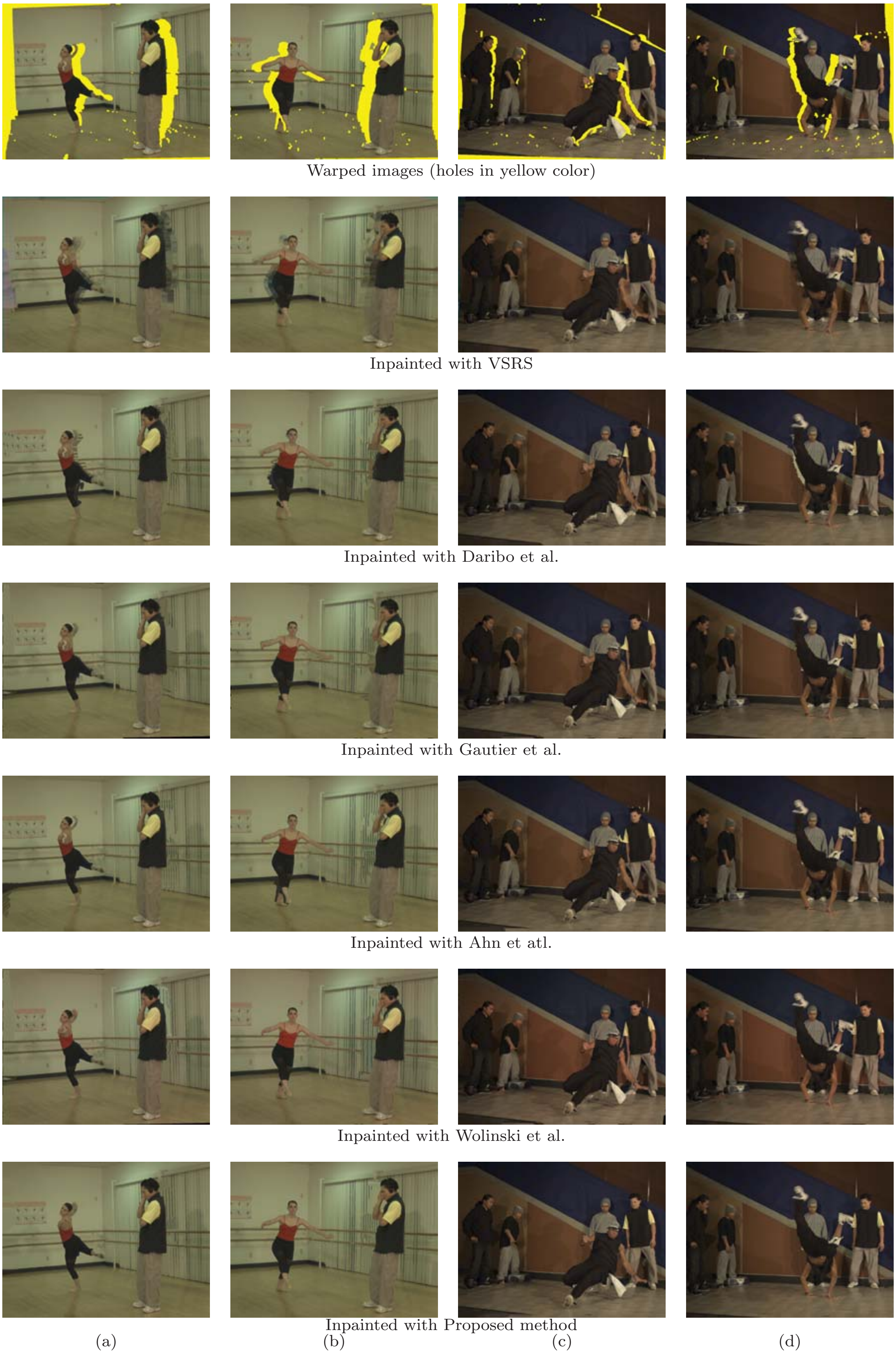
[33] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, R. Szeliski, High-quality video view interpolation using a layered representation, ACM Trans. Graph. 23 (3) (2004) 600–608.

[34] G. M. Um, G. Bang, N. Hur, J. Kim, Y. S. Ho, 3d video test material of outdoor scene, ISO/IEC JTC1/SC29/WG11/M15371 (Apr. 2008).

[35] Multiview video test sequence and camera parameters, ISO/IEC JTC1/SC29/WG11 MPEG2008/M15419, archamps, France (2008).

[36] M. Domanski, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, K. Wegner, Poznan multiview video test sequences and camera parameters, ISO/IEC JTC1/SC29/WG11 MPEG 2009/M17050, xian, China (2009).

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



Warped images (holes in yellow color)

Inpainted with VSRS

Inpainted with Daribo et al.

Inpainted with Gautier et al.

Inpainted with Ahn et atl.

Inpainted with Wolinski et al.

Inpainted with Proposed method

Figure 16: Depth-based inpainting visual quality evaluation: (a) “Ballet” frame19 v5→v4); (b) “Ballet” (frame4 v5→v6); (c) “Breakdancers” (frame31 v5→v4); (d) “Breakdancers” (frame75 v5→v6).