# A Novel Deep Learning-based Disocclusion Hole-Filling Approach for Stereoscopic View Synthesis

Wei Liu, Mingyue Cui, and Liyan Ma

***Abstract*—An important technique for converting 2D videos into 3D is depth image-based rendering (DIBR), which creates virtualized perspectives with textured images and corresponding depth maps. Yet, the majority of the currently used methods struggle in handling the disocclusion holes in warped virtual images. This research presents a unique deep learning-based disocclusion hole-filling approach in stereoscopic vision synthesis as a solution to this issue. Firstly, we explicitly take into account some particular limitations of the synthesized virtual views and designate them as scene influence maps in the network, which might offer some significant extra scene cues to lessen hallucinated content mixing among various layers. Then, an enhanced directional scene influence map, which diffuses using a novel anisotropic diffusion equation under consistent stereoscopic constraints, is further investigated for efficient disocclusion hole filling. Empirical analyses and comparisons on the *Middlebury* and *KITTI* datasets confirmed that our approach outperforms previous deep learning-based generative inpainting algorithms for disocclusion hole filling in the warped views.**

***Index Terms*—disocclusion hole filling, deep learning, scene influence map, stereoscopic synthesis.**

## I. Introduction

RECENT developments in three-dimensional (3D) videos provide more immersive visual experiences to viewers than traditional two-dimensional (2D) videos. A rising demand for 3D content is a result of the subsequent 3D industry growth. Currently, most material available for 3D TV broadcasting is produced by capturing a few image streams and transmitting them towards a receiver for viewing. Meanwhile, the procedure is time-consuming and expensive due to shot planning, camera rig management, costly hardware, and the substantial post-processing needed to correct stereographic inaccuracies.

Depth image-based rendering (DIBR) can confidently be considered as an alternative approach that allows the creation of virtual views by utilizing only one referencing texture and its coordinate depth map [1]. DIBR is a practical way to convert 2D to 3D and it requires only a single viewpoint. The corresponding depth information can be extracted from monocular image sequences using computer vision techniques [2]. This lowers the total cost of the system and, more crucially, makes it possible to utilize massive 2D multimedia libraries that already exist.

The key technique of DIBR is referred to as 3D warping [3]. This involves projecting every pixel in the reference picture into world coordinates by utilizing the depth data, then reprojecting all of the resultant points onto the object's viewing plane. However, a major drawback of DIBR is the potential for artifacts to display within the virtual image. Once a specific foreground available in the referenced view obscures the background, disocclusion may be seen in the virtual view [4]. The disocclusions look like voids since no pixels get warped into such regions, which significantly lowers the virtual view's visual quality. This DIBR drawback requires filling these holes, for which many algorithms have been developed.

There are two sorts of traditional techniques. For the first one, in order to minimize the size of the hole, the depth map must first be preprocessed by adding a low-pass filter [5]. The foreground-background depth difference is a factor in the creation of occlusions. Deep discontinuities often lead to significant disocclusion. These techniques consistently highlight the use of depth map refining prior to DIBR to prevent holes [6]; however, they may contribute additional geometric distortions or artifacts near the disocclusion regions. Several improved approaches have been proposed to mitigate this problem, such as asymmetric smoothing [7], [8], scene structure, and content-dependent adaptive filters [9], [10]. Such approaches avoid extra smoothing in areas that aren't holes by maintaining greater smoothing in a few designated limited regions rather than the full picture. However, these approaches may reduce the 3D effect with depth maps' smoothing. Furthermore, only tiny baseline conditions are appropriate for such depth preprocessing techniques. For larger baselines, a single smoothing is no longer sufficient due to the increased disocclusion area.

The second technique fills up the disocclusion through utilizing the textural correlation between neighboring pixels. The inpainting-based approach to filling holes is another available option [11]. The example-based inpainting algorithm proposed in [12] can determine the priority of the pixels at hole boundaries based on confidence and data terms. The background texture must be used to fill the revealed regions with uneven fills, though, as they are a part of the backdrop. Employing the technique of [12] may introduce some foreground textures into the hole regions, causing the foreground to blend. To address this flaw, various methods

Manuscript received May 24, 2022; revised March 21, 2023. This work was supported in part by Key Scientific and Technological Project of Henan Province under Grant 232102211047, Foundation of Excellent-Young-Backbone Teacher of Colleges and Universities in Henan Province under Grant 2019GGJS182 and Key Scientific Research Project of Henan Colleges and Universities under Grant 21B120001.

Wei Liu is an Associate Professor at the College of Electromechanic Engineering, Nanyang Normal University, Henan, Nanyang 473061, China (corresponding author, e-mail: lw3171796@163.com).

Mingyue Cui is an Associate Professor at the College of Electromechanic Engineering, Nanyang Normal University, Henan, Nanyang 473061, China (e-mail: cuiminyue@sina.com).

Liyan Ma is an Associate Professor at the College of Computer Engineering and Sciences, Shanghai University, Shanghai 200444, China (e-mail: liyanma@shu.edu.cn).

employ depth [13], [14], [15] or background-foreground knowledge [16], [17] as viable limitations to eliminate foreground textures during filling in depth-based view synthesis. However, in intricate sceneries with non-repeating patterns, they would not create convincing contents. Furthermore, they are computationally expensive due to their iterative nature.

To solve these issues and outperform prior conventional approaches, deep learning-based algorithms were subsequently investigated. These deep learning-based techniques are motivated by deep learning's effectiveness in a variety of uses, including picture denoising [18], super-resolution [19], and target recognition [20], [21], [22]. Several studies have built networks to deal with disocclusion holes that arise after DIBR procedures. To restore the obstructed parts of the warped images, researchers specifically approach the disocclusion topic in the form of a creative picture inpainting task, in which they apply deep learning-based inpainting algorithms [23], [24]. Recent work includes designing CNN architectures to handle irregularly shaped holes better [25], [26] and two-stage approaches that employ structural-content separation, such as predicting structures (e.g., contours/edges of missing sections), then completing features according to the forecasted structures [27], [28], [29]. Our disocclusion hole-filling model is inspired by these recent two-stage methods with two key differences, which are our main contributions in this study.

- First, the majority of existing deep learning-based inpainting techniques make the assumption that CNNs may implicitly learn the scene structure and layer information with no additional guidance. As a result, they would not offer any other details regarding this model. Although some two-stage approaches have used additional cues as predicted structures to improve the results, important implicit priors in the DIBR process still have not been fully considered, which may cause quality degradation in the warped views as a result of low foreground-background spatial correlation near the boundary areas. To overcome these limitations, in this study, we explore various unique constraints of the synthesized virtual views and define them as scene influence maps to reduce hallucinatory structural combinations in the warped views in the network.
- Second, the goal of picture inpainting is to provide convincing substance to the empty spaces in images. Traditionally, holes are supposed to be filled with available texture information in a visually plausible manner without other constraints, so many approaches fill the holes using all valid image content. However, this greedy way can cause undesired visual inconsistency in the warped view by failing to consider the consistency between the hole and surrounding areas. Usually, disocclusion holes are localized within transitional regions of different layers, where neighboring information from the background is more reliable for recovering such information. Considering these, our proposed scene influence maps in this study are generated by diffusing from anchor points in the hole regions. Furthermore, an enhanced directional scene influence map, which diffuses using a novel anisotropic diffusion equation under consistent stereoscopic constraints, is explored for efficient disocclusion hole filling.

The other sections covered in this article have been structured as follows. The technological framework related to these suggested approaches is thoroughly presented in Section II. Section III examines and highlights the main experimental findings. Section IV includes a few closing notes.

## II. Proposed approach

The assumption made by the GAN-based frameworks included in early image inpainting approaches [27], [28], [29] projected the network's ability to implicitly acquire some of the required information about the layout and quality of the image. This was one of the most challenging optimization issues for neural networks. Modern techniques have provided more specific information, like edges [27], structure constraints [28], [29], and semantic labels [30], to provide realistically textured content with a credible structural foundation. Though, disocclusion hole filling in the stereoscopic synthesis cannot be regarded simply as a generative inpainting problem, so adding the aforementioned additional information to the network is still insufficient to address these issues effectively.

Fig.1 demonstrated a scenario covering the hole filling in warped images. Clear disocclusion hole region is marked with blue in Fig.1(a). The hole region is dispersed along the areas that separate the contents of the foreground-background. If we are only assessing the visual quality of the image, a state-of-the-art generative inpainting method, like EdgeConnect (EC), can obtain visually realistic results, as shown in Fig.1(b). However, the zoomed rectangular region of Fig.1(b) shows that the newly reconstructed boundary does not along the hole edges. The underlying reason for this problem is that the restored contents in the hole mixed the texture information from both the foreground and background contents. The background content and the disocclusion zone are frequently thought to be part of the same physical surface, hence their texture patterns must be identical. Therefore, although the restored region of Fig.1(b) is semantically meaningful, it did not meet the implicit constraints in the 3D warped scene, which may cause a serious stereo-mismatching problem when displayed in a stereo projection system. In our study, some special scene cues in the warped view are extracted and then represented by a scene influence map. Thus, our approach can further optimize the final result under the stereoscopic constraints more naturally, as shown in the zoomed rectangular region of Fig.1(c), where the reconstructed layer boundaries match the original texture image well. Fig.1(d) is the corresponding ground-truth image. We will present the suggested approach in the following subsections.

### A. Framework of the proposed approach

Fig.2 demonstrates that to restore the hole areas in the warped view step by step and to enable the model to be aware of the scene contextual information, the framework of our deep learning-based hole-filling approach comprises two stages:

(1) Stage I: scene cue extraction

In the first stage, the underlying stereoscopic constraints, which are more useful in guiding hole inpainting in the

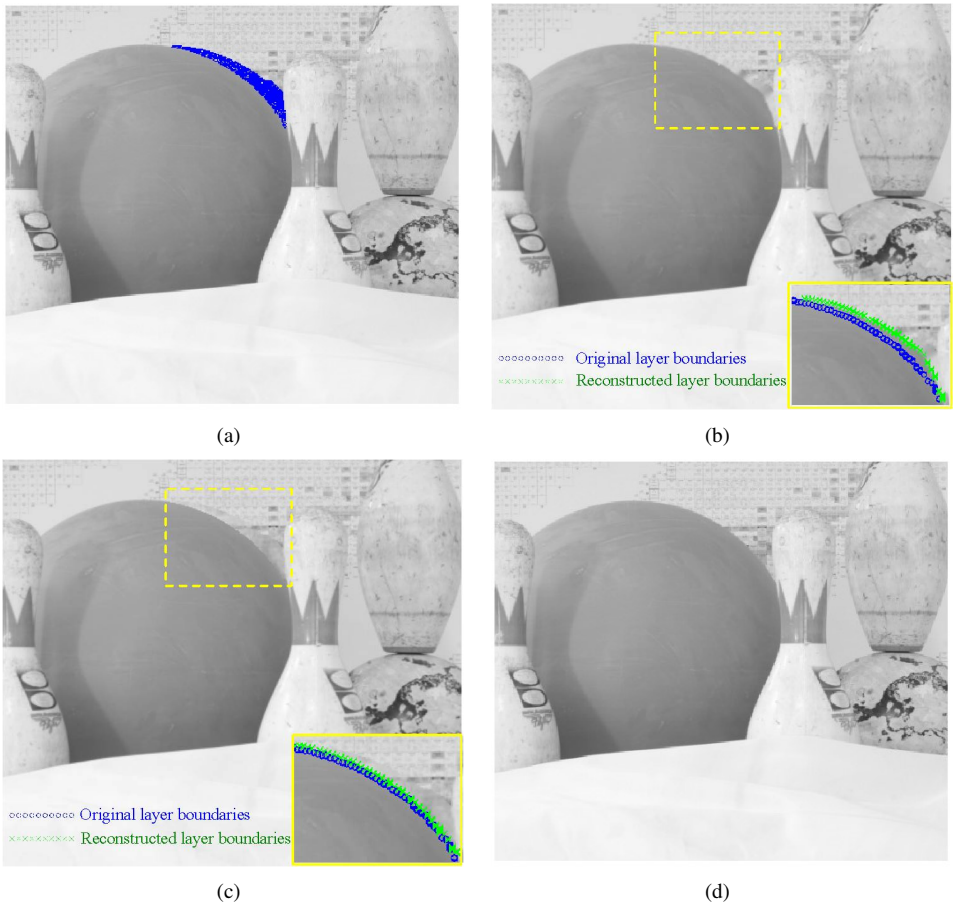


Fig. 1. Example of hole filling in the warped view. (a) warped image with a disocclusion hole, (b) hole-filling results with EdgeConnect (EC) [27], (c) results of the suggested approach, and (d) ground-truth image

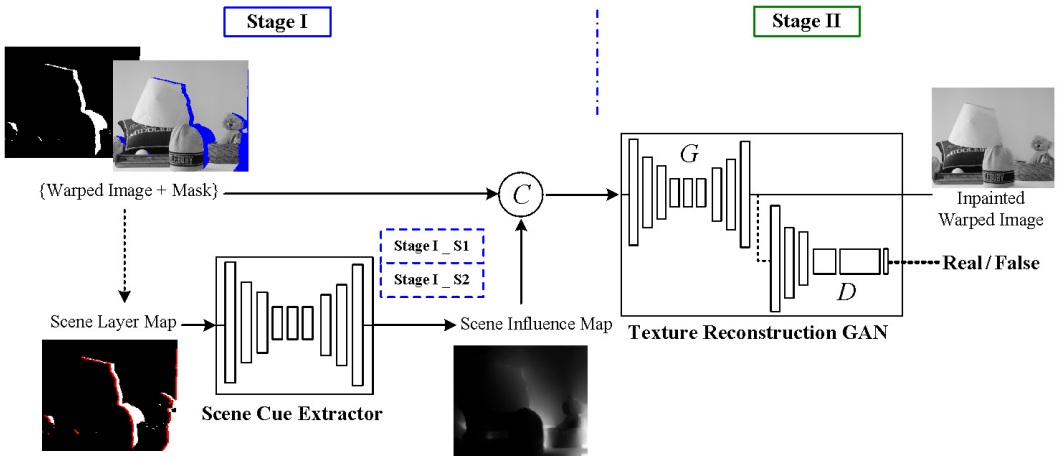


Fig. 2. Framework of the suggested deep learning-based disocclusion hole-filling method

warped view, are extracted in the form of a scene influence map, that is defined as:

$$S_c = \mathscr{F}_1(I_w, M_w) \tag{1}$$

where $\mathscr{F}_1$ represents a subnetwork, which acts as a scene cue extractor and produces scene influence maps $S_c$ in the warped view. $M_w$ represents the mask that differentiates different regions. It is in the form of a binary matrix in accordance with the hole locations within the distorted picture/warped image $I_w$, where 1 represents the hole region and 0 represents the background.

Two subnetwork learning schemes with different constraints, named *Stage I_S1* and *Stage I_S2* as displayed in Fig.2, were investigated in this stage to extract the underlying implicit scene stereoscopic cues of different levels, which will be discussed respectively and thoroughly in the below subsections.

(2) Stage II: texture reconstruction

In the second stage, a generative image inpainting-based network, $\mathscr{F}_2$, is used with the extracted $S_c$ as the priori information to enhance texture prediction $I'_w$ in the hole regions:

$$I'_w = \mathscr{F}_2(I_w, M_w, S_c) \tag{2}$$

The latent characteristics in the stereoscopic synthesis may be organically merged into the network with the aid of a scene influence map, preventing unnecessary foreground pixels from becoming distorted in these holes. As a result, disocclusion hole filling in this virtual viewpoint outperforms current generative inpainting algorithms in terms of performance.

### *B. Stage I_S1: Hole filling with scene stereoscopic constraints in the warped view*

Generating meaningful structures from the surrounding background contents for the missing regions is a challenge of hole-filling tasks in the warped view. Therefore, we first design a scene cue extractor, $\mathscr{F}_1$, to achieve this goal. It aims to learn the implicit constraints in the warped scene to the network and is realized by training with a scene influence map. In order to create the scene influence map, the initial anchor points are diffused. They are specified as:

$$S_{anchor}(p) = \begin{cases} 1 - exp(-D_l(p)), & if\ p \in R_H \\ 0, & if\ p \notin R_H \end{cases} \tag{3}$$

where $R_H$ represents the warped hole regions, $D_l(p)$ represents the distance from the point $p$ to the boundaries between the foreground and background in the layer boundary map $S_l$, which is represented by:

$$S_l(p) = \begin{cases} sgn(p), & if \quad M_w(p_l) > M_w(p_r) \\ 1 - sgn(p), & otherwise \end{cases} \cdots \\ \cdots, for \quad p \in \Delta(M_w) \tag{4}$$

where $\Delta(.)$ represents the *Laplacian* operator, $p_l$ and $p_r$ represent the left and right pixels of $p$, respectively, and $sgn$ represents an indicator function. For the left virtual view in our work, $sgn(p) = 1$, and for the right, $sgn(p) = 0$.

For stereoscopic synthesis, the restored textures in the warped view should have similar patterns to the background. Therefore, the anchor points adjacent to background regions are assigned large values, meaning they have a high influence-degree coefficient. Thus, when a domain transform-based filter [31] is used to diffuse the initial anchor points of the hole regions into the inner background regions and generate a scene influence map, the heuristics from the backdrop may be used to better guide hole filling in the virtual perspective. This processing can be expressed as:

$$ct(u) = \int_0^u \frac{\sigma_H}{\sigma_s} + \frac{\sigma_H}{\sigma_r} |I_{gt}(x)| dx \tag{5}$$

where $ct(u)$ indicates the transformed domain, and $\sigma_H, \sigma_s, \sigma_r$ represent all the terms utilized in controlling the diffusion impact of the initial anchor point distribution map. $I_{gt}$ denotes the ground-truth picture of the warped view.

In the transformed domain, the final scene impact map may be stated with a recursive form as follows:

$$\begin{cases} S_c[n] = (1 - a^d) S_{anchor}[n] + a^d S_c[n-1] \\ d = ct(u_n) - ct(u_{n-1}) \end{cases} \tag{6}$$

where $a = exp(-\sqrt{2}/\sigma_H)$, and which is also the feedback coefficient of this filter. $d$ denotes the distance between neighbor samples $u_n$ and $u_{n-1}$ in the transformed domain defined in Eq.(5), which controls the diffusion strength delivered from the transformed domain to influence maps.

By this means, taking $S_{anchor}$ as the original scene influence map to be processed, $S_c$ is produced through joint filtering with texture image $I_{gt}$ operated by the recursive domain transform filter as discussed in Eqs.(5) and (6).

### *C. Stage I_S2: Enhanced scene cues with directional constraints in the warped view*

The anchor points defined by Eq.(3) provide only a soft stereoscopic constraint. If the foreground contents around disocclusion holes have a rich texture, the current hybrid constraints in Eq.(5) are insufficient to completely prevent anchor points around layer boundaries from flowing into the neighboring foreground areas. To address this issue, this section further explores the scene context by restricting the influence map to the background area. More precisely, it is realized by adding a hard directional constraint to the domain transform-based diffusion equation. Then Eq.(5) can be rewritten as:

$$ct(u) = \int_0^u \frac{\sigma_H}{\sigma_s} + \frac{\sigma_H}{\sigma_r} e^{\beta S_l(p)} |I_{gt}(x)| dx \tag{7}$$

where $\beta$ represents a directional reference factor, which has been given a value of 5 in this work. $S_l$ represents the layer boundary map defined in Eq.(4), which indicates the layer boundaries between the foreground and background.

Therefore, when processing the layer boundary regions, the propagation chains would be stopped by this added constraint in the transformed domain, and the diffusion strength delivered from Eq.(6) is correspondingly decreased. Then, the generated influence map $S_{c\_d}$, which we define as a directional scene influence map, will mainly represent the relationships between the hole regions and backgrounds.

Furthermore, to enable the network to learn the implicit function mapping between the directional scene influence map and the warped view more easily, we also add the layer boundary map as an additional input to the scene cue extractor sub-network. Then, Eqs.(1) and (2) can be rewritten as:

$$S_{c\_d} = \mathscr{F}_1(I_w, M_w, S_l) \tag{8}$$

$$I'_w = \mathscr{F}_2(I_w, M_w, S_{c\_d}) \tag{9}$$

### *D. Network design and training strategy*

Fig. 2 displays the general design of the suggested approach. The process requires some key components, including a scene cue extractor network $\mathscr{F}_1$ and a GAN-based texture reconstruction network $\mathscr{F}_2$. We elaborate on each of these components in the following.

The scene cue extractor network $\mathscr{F}_1$ outputs a directional scene influence map from the concatenation of warped picture/mask along with the associated layer boundary map. In this step, pictures are downsampled twice by encoders, trailed with 8 residual blocks, and upsampled to their original size by decoders using an encoder-decoder network design. In the residual layers, dilated convolutions having a dilation

parameter equivalent to 2 are utilized in place of conventional convolutions.

The GAN-based texture reconstruction network $\mathscr{F}_2$ further restores the disocclusion holes using the scene influence map $S_{c\_d}$, extracted in the last stage as an additional key cue to guide the following inpainting process, that uses an adversarial model. In other words, the stage normally comprises and includes the generator-discriminator combination $\{G, D\}$. This scene cue extractor shares the same network architecture as Generator $G$. We employ the PatchGAN architecture, which detects if overlapping picture patches are genuine, for discriminator $D$. A $2 \times 2$ pixel stride is used by all convolutional layers to reduce picture resolution while maximizing the allowable output filters.

Although our approach consists of two different stages, with each stage aimed at a specific subtask, all the layers of the suggested framework are differentiable. Therefore, the entire network is capable of undergoing an end-to-end training.

An adversarial loss and the reconstruction loss together make up the total loss. The reconstruction loss $L_{rec}$ is composed of two parts $\{L_1, L_2\}$, which establishes the differences in errors between the forecasts and the ground truth in the two stages, respectively. The reconstruction loss function is defined as:

$$\begin{cases} L_{rec} = \gamma_1 L_1 + \gamma_2 L_2 \\ L_1 = \|S_{c\_d} - S_{gt}\| \\ L_2 = \|M_w \odot (I'_w - I_{gt})| \end{cases} \tag{10}$$

where $\odot$ is pixel-wise multiplication and $\|.\|$ is the Euclidean norm. $S_{gt}$ is the directional scene influence map ground truth, which is generated with texture image ground truth $I_{gt}$ conducted using recursive domain transform filter which has been further defined in Eqs.(5) and (6). $\gamma_1, \gamma_2$ are the loss term weights, which are adaptively set during training. We set $\gamma_1 = 2, \gamma_2 = 1$ in the 1st half of overall training iterations, and $\gamma_1 = 1, \gamma_2 = 1$ for the remaining iterations. This strategy helps the network to learn scene cues with higher priority at the beginning, which is beneficial for boosting the training phase.

The distribution of data is not guaranteed to be the same as that of the natural pictures since the reconstruction loss only penalizes the pixel-wise inaccuracy. Consequently, the outcomes of the inpainting might very well be hazy. By enforcing adversarial losses that are based on a generative adversarial network (GAN) [32], this can be reduced. The following is the definition of the adversarial loss:

$$\begin{aligned} L_{adv} = \max_{D} \mathbb{E} \left[ \log D\left(I_{gt}, M_w\right) \right. \\ \left. + \log\left(1 - D\left(I'_w, M_w\right)\right)\right] \end{aligned} \tag{11}$$

Thus, the total loss function can be obtained using the following:

$$L = \lambda_{rec} L_{rec} + \lambda_{adv} L_{adv} \tag{12}$$

where $\lambda_{rec}, \lambda_{adv}$ represent the weights of the loss terms. In this research, we chose $\lambda_{rec} = 1, \lambda_{adv} = 0.1$.

## III. Experimental evaluation

### A. Experimental setup

Our approach was implemented using the *TensorFlow* platform with a batch size of 16 using a PC having a GPU specification of 32GB NVIDIA Quadro. Adam optimizer has been employed in enhancing the model with $\beta_1 = 0.0005$ and $\beta_2 = 0.9$. We used open-source datasets to show how well the suggested hole-filling approach for stereoscopic vision synthesis performed: the *Middlebury* stereo vision and *KITTI* 2015 databases. The training was completed using 48236 sub-images at $256 \times 256$ resolution, some obtained by from the *Middlebury* datasets and the others by the *KITTI* datasets. The remaining data from both databases were adopted for testing to validate our proposed network.

In the following subsections, we first evaluate the experimental results qualitatively and quantitatively. Some experimental details of the proposed approach and an analysis of the outcomes of the collected directional scene restrictions are then reviewed.

### B. Visual image evaluations

In this part, we first visually evaluate the *Middlebury* and *KITTI* test sets that resulted from the following approaches:

- EdgeConnect (EC) [27], a representative, cutting-edge deep learning-based image inpainting method, that uses the processing scheme covered in [24] to fill disocclusion holes through generative image inpainting.
- The proposed approach with different scene cues. The first (Scheme1) uses the extracted scene influence map $S_c$ as the scene constraints, and the second (Scheme2) uses the enhanced directional influence maps $S_{c\_d}$.

Figs.3 and 4 each show a distinct illustration of the experimental findings. The original pictures of these test sets are regarded as left views. According to Figs.3(a) and 4(a), the newly revealed hole regions for the right views are on the right half of foreground objects. The structural information surrounding the holes is always unique and pertains to the foreground and background independently. With no right direction, it is challenging for EC to rebuild the holes by just using background texture signals. As seen by red-colored rectangles displayed in both Figs.3(b) and 4(b), content mixtures are introduced by EC at different levels. For comparison, our approach with Scheme1 generates more reasonable results. However, when we zoom in on the details, a mixture of shadows can still be seen in the yellow rectangles of Figs.3(c) and 4(c). The best results are generated by our approach with Scheme2, where the enhanced influence maps effectively prevented foreground contents from flowing into the hole regions, particularly in the green rectangles of Figs.3(d) and 4(d). As can be seen, the proposed approach provides more professional disocclusion hole filling in the warped views, even in some complex scenes, such as *Midd2* in Fig.3(d), where the boundaries on the right of the hat can be distinguished clearly after the neighboring hole regions are restored.

### C. Quantitative evaluations

We employed a number of common measures for quantitative evaluations to further analyze the performances of the suggested ways: peak signal-to-noise ratio (PSNR), Frechet inception distance (FID), and structural similarity index (SSIM), according to [27].

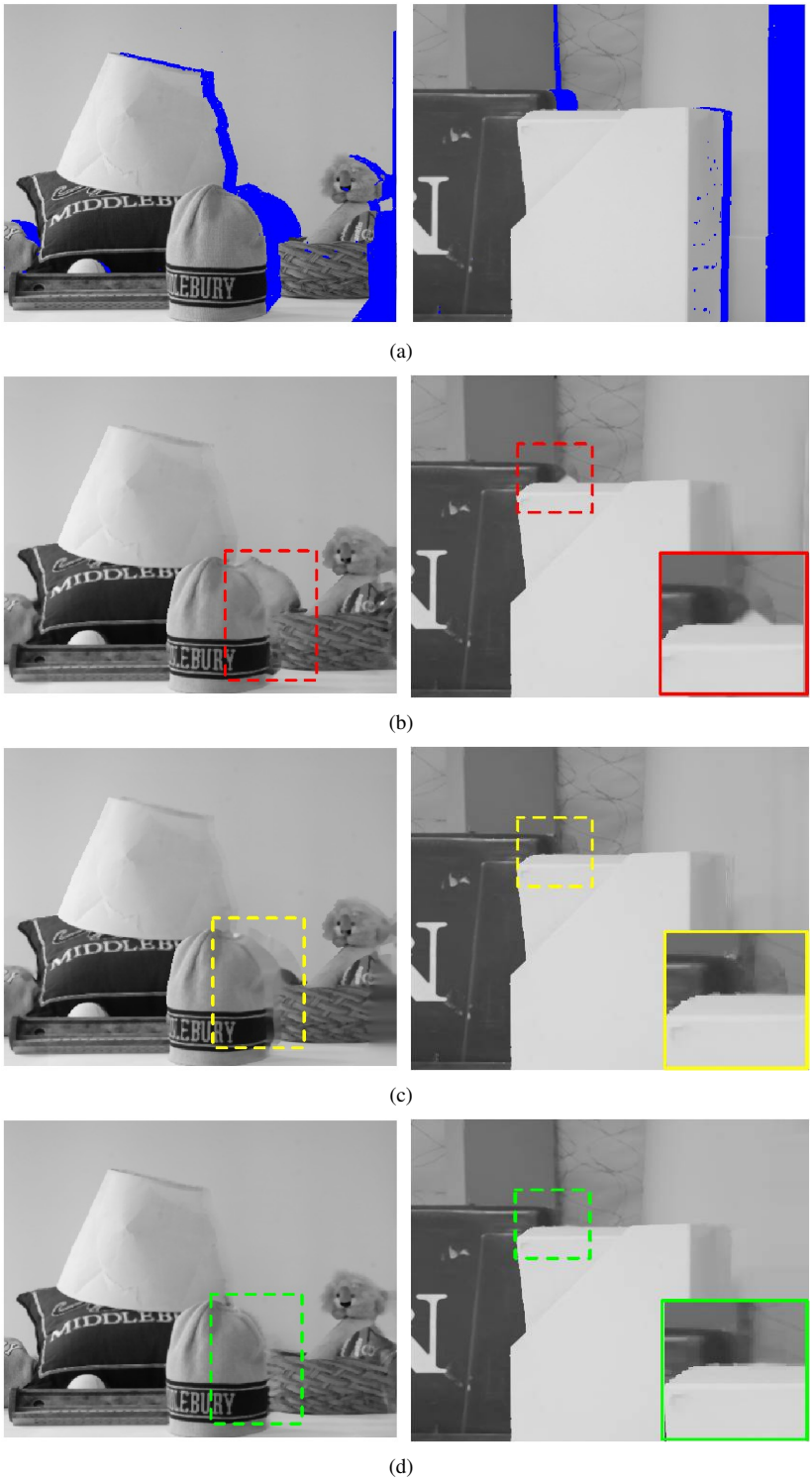

Fig. 3. Virtually observed pictures through the use of the *Middlebury* datasets including (from left to right) *Midd2* and *Plastic*: (a) warped pictures with holes, and holes filled with (b) EC, (c) Scheme1, and (d) Scheme2.

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

Fig. 4. Virtually observed pictures through the use of the *KITTI* dataset: (a) warped picture with holes, and holes filled with (b) EC, (c) Scheme1, and (d) Scheme2

TABLE I
QUANTITATIVE EVALUATION RESULTS
(↑:HIGHER IS BETTER,↓:LOWER IS BETTER)

| | Database | EC | Scheme1 | Scheme2 |
|---|---|---|---|---|
| **PSNR(dB)↑** | *Middlebury* | 27.23 | 29.24 | **30.96** |
| | *KITTI* | 27.56 | 29.52 | **31.41** |
| **SSIM↑** | *Middlebury* | 0.8818 | 0.9084 | **0.9193** |
| | *KITTI* | 0.8873 | 0.9107 | **0.9232** |
| **FID↓** | *Middlebury* | 6.12 | 3.87 | **2.76** |
| | *KITTI* | 5.43 | 3.16 | **2.24** |

According to Table I, our approach surpasses the generative inpainting method EC in all the metrics. The proposed approach with Scheme2 produced the best scores. Like the experimental results that we previously discussed regarding the visual qualities, the lower performance of EC is unsurprising because it does not use any special stereoscopic consistent constraints in the warped view, such as the scene/enhanced influence maps extracted by our approaches.

TABLE II
SUBJECTIVE QUALITY EVALUATION RESULTS

| | Database | EC | Scheme1 | Scheme2 |
|---|---|---|---|---|
| **Test1** | *Middlebury* | 4.0 | 4.2 | **4.3** |
| | *KITTI* | 3.9 | 4.1 | **4.2** |
| **Test2** | *Middlebury* | 3.8 | 4.2 | **4.4** |
| | *KITTI* | 3.8 | 4.2 | **4.3** |

Using these test datasets, ten people having normal or

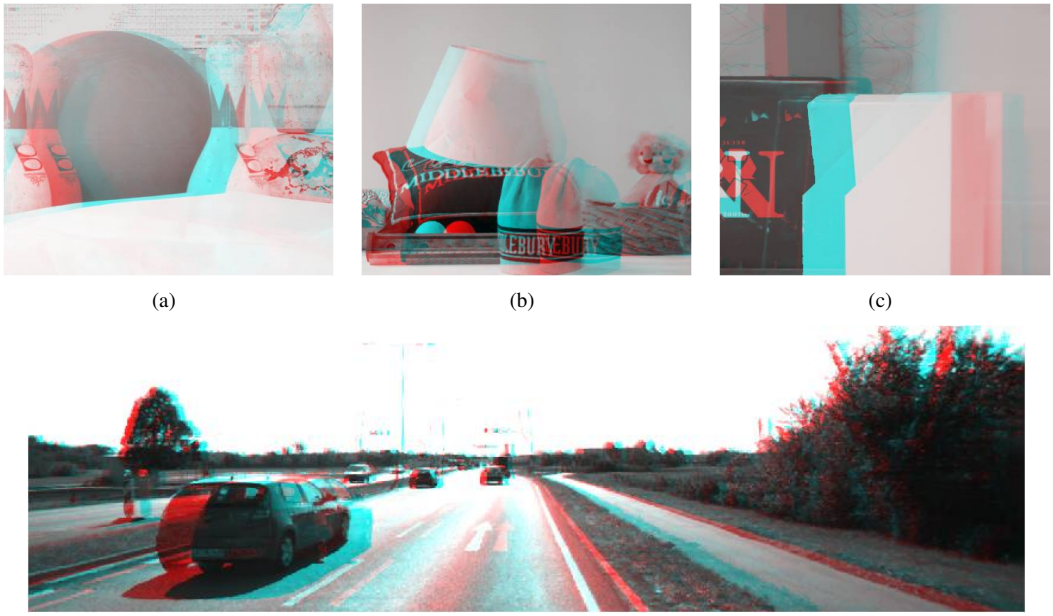

Fig. 5. Chosen synthesized 3D anaglyph images that were obtained from the test samples through the suggested method: (a) *Bowling1*, (b) *Midd2*, (c) *Plastic*, and (d) *KITTI*.

those falling within the description of corrected-to-normal visuality and stereo vision conducted subjective human testing. Furthermore, we ran two experiments to gauge the picture quality for restored virtual views (Test1) and the stereoscopic perception for synthetic 3D anaglyph pictures (Test2). Each was graded using a scale range 0-5, where a higher number denotes better picture quality or a more stereoscopic experience. Within every test, users were requested to score their pleasure after viewing the virtual or synthetic 3D pictures in a random sequence. As stated in Table II, the mean scores received had been utilized to represent the gauge of the subjective assessment. Bold font is used to display the optimum findings. This table demonstrates how the suggested method using Scheme2 outperformed other methods in terms of picture quality for virtual views and stereoscopic perception for artificially created 3D anaglyph images. Note that the generative inpainting method EC was scored closer to ours in Test1 than in Test2, where the scoring differences were larger. This occurred because, as we discussed concerning Fig.1(b), in some scenes, depth perception had been affected despite the repaired quality of the visual image due to content mixtures in the warped perspective. The assessment test sets' demonstrations of synthetic 3D anaglyph pictures are shown in Fig.5.

### *D. Implementation details*

In this subsection, we present further experiments in Fig.6 to demonstrate the effects of the implementation details in our approach. In Figs.6(a) and (d), hole regions are outlined in blue and the green rectangle areas are regions of interest (ROIs), which we primarily focus on in the intermediate experimental results. For right views, the layer boundary maps, marked with red lines, appear along the left side of each hole. With their guidance, the enhanced influence maps generated for the ROIs in Figs.6(c) and (f) show more directional characteristics than those without the constraints in Figs.6(b) and (e). For example, for test set *Midd2*, the enhanced influence map in the ROI of Fig.6(c) is only diffuse on the right side of the corresponding layer boundary line shown in Fig.6(a), unlike the one which is diffuse on both sides in the ROI of Fig.6(b). Thus, in the texture reconstruction stage, just as the results displayed in the experiments of Fig.3(d), the contents in the hat were no longer introduced into the hole regions. Furthermore, this experiment proves that the extracted directional scene influence map can provide enough information to realize efficient disocclusion hole filling under the consistent stereoscopic constraints in our proposed scheme.

## IV. Conclusion

Throughout this research, a stereoscopic vision synthesis method was suggested on the basis of a deep learning-based disocclusion hole-filling methodology with directional scene cue guidance. Scene cue extraction and texture reconstruction are the two steps of our method for breaking the work down. In the first stage, a directional influence map, which diffuses in a novel anisotropic manner, is defined and extracted. In the second stage, the extracted scene influence map provides key additional heuristic cues and plays an important role in efficient disocclusion hole filling under consistent stereoscopic constraints. Experimental results verified that the suggested approach greatly improves the disocclusion hole-filling performance in the warped views over the results of conventional generative inpainting methods.

## References

[1] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," in *Proceedings of SPIE - Stereoscopic Displays and Virtual Reality Systems XI*, vol. 5291, 2004, pp. 93–104.

[2] Z. Liang, C. Vazquez, and S. Knorr, "3D-TV content creation: Automatic 2d-to-3d video conversion," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 372–383, 2011.

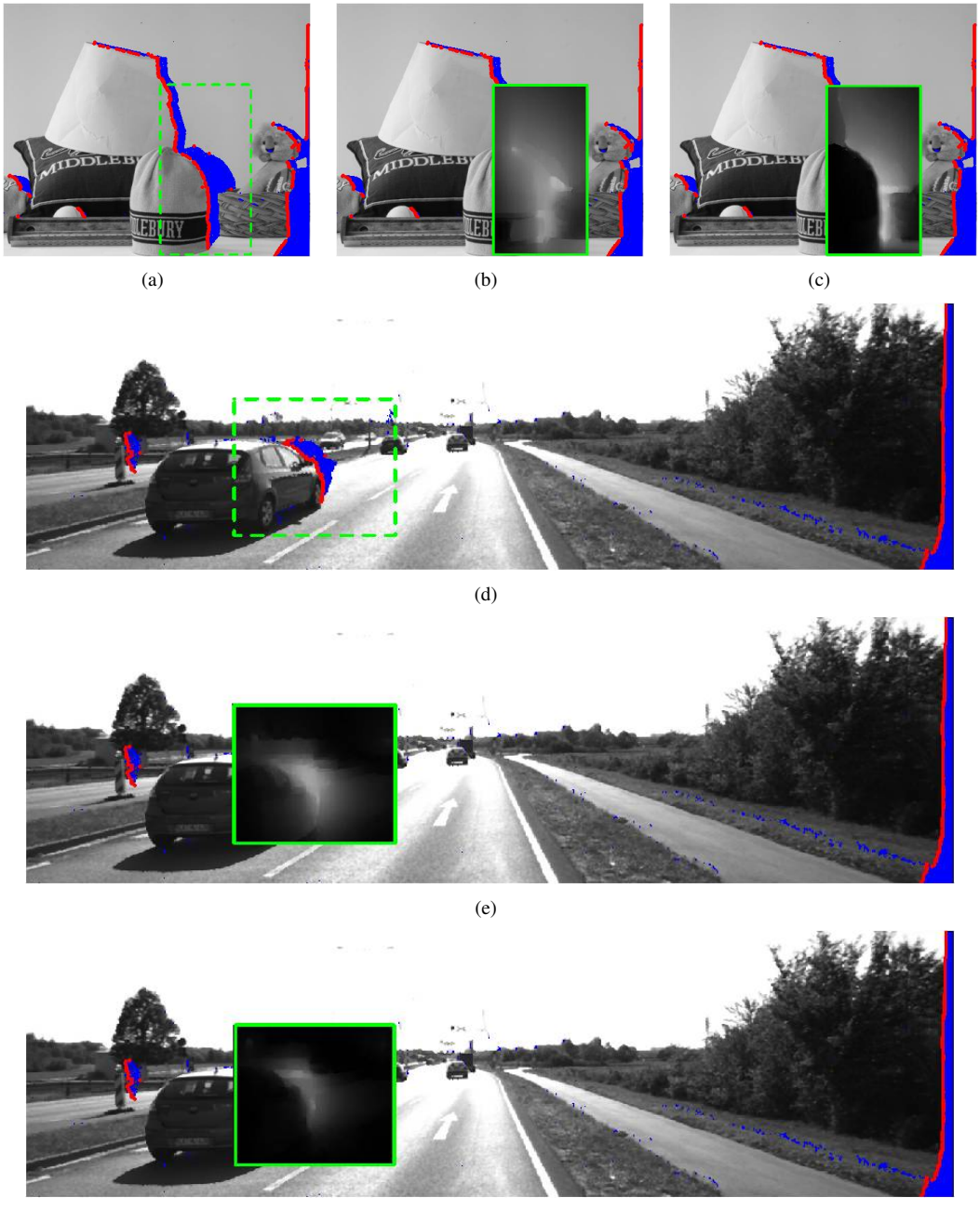

Fig. 6. Intermediate experimental results on *Midd2* (a-c) and *KITTI* (d-f): (a),(d) layer boundary maps; (b),(e) influence maps generated in the regions of interest (ROIs) without enhanced scene constraints; and (c),(f) influence maps generated in the ROIs with enhanced directional scene constraints.


[3] M. P. Leong, M. Y. Yeung, C. K. Yeung, C. W. Fu, P. A. Heng, and P. Leong, "Automatic floating to fixed point translation and its application to post-rendering 3d warping," in *IEEE Symposium on Field-programmable Custom Computing Machines*, 1999, pp. 1–9.

[4] W. Liu, L. Ma, B. Qiu, and M. Cui, "Stereoscopic view synthesis based on region-wise rendering and sparse representation," *Signal Processing: Image Communication*, vol. 47, pp. 1–15, 2016.

[5] W. J. Tam, G. Alain, L. Zhang, T. Martin, and R. Renaud, "Smoothing depth maps for improved steroscopic image quality," in *Proceedings of SPIE - Three-Dimensional TV, Video, and Display III*, vol. 5599, 2004, pp. 162–172.

[6] X. Chen, H. Liang, H. Xu, S. Ren, H. Cai, and Y. Wang, "Virtual view synthesis based on asymmetric bidirectional DIBR for 3D video and free viewpoint video," *Applied Sciences*, vol. 10, no. 5, pp. 1–19, 2020.

[7] L. Zhang and W. J. Tam, "Stereoscopic image generation based on depth images for 3D TV," *IEEE Transactions on Broadcasting*, vol. 51, no. 2, pp. 191–199, 2005.

[8] S. Zhu, H. Xu, and L. Yan, "An improved depth image based virtual view synthesis method for interactive 3D video," *IEEE Access*, vol. 7, pp. 115 171–115 180, 2019.

[9] S.-B. Lee and Y.-S. Ho, "Discontinuity-adaptive depth map filtering for 3D view generation," in *2nd International ICST Conference on Immersive Telecommunications*, 2010, pp. 1–6.

[10] I. Daribo, C. Tillier, and B. Pesquet-Popescu, "Distance dependent depth filtering in 3D warping for 3DTV," in *IEEE 9th Workshop on Multimedia Signal Processing*, 2007, pp. 312–315.

[11] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and Y. Akbari, "Image inpainting: A review," *Neural Processing Letters*, no. 5, pp. 2007–2028, 2019.

[12] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1200–1212, 2004.

[13] I. Daribo and H. Saito, "A novel inpainting-based layered depth video for 3DTV," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 533–541, 2011.

[14] S. M. Muddala, M. Sjöström, and R. Olsson, "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,

2016.
[15] I. Ahn and C. Kim, "A novel depth-based virtual view synthesis method for free viewpoint video," *IEEE Transactions on Broadcasting*, vol. 59, no. 4, pp. 614–626, 2013.
[16] H. Liang, X. Chen, H. Xu, S. Ren, H. Cai, and Y. Wang, "Local foreground removal disocclusion filling method for view synthesis," *IEEE Access*, vol. 8, pp. 201 286–201 299, 2020.
[17] A. Q. de Oliveira, M. Walter, and C. R. Jung, "An artifact-type aware DIBR method for view synthesis," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1705–1709, 2018.
[18] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, "Deep learning on image denoising: An overview," *Neural Networks*, vol. 131, pp. 251–275, 2020.
[19] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.
[20] D. Yuan and Y. Xu, "Lightweight vehicle detection algorithm based on improved YOLOv4." *Engineering Letters*, vol. 29, no. 4, pp. 1544–1551, 2021.
[21] Y. S. Tan, K. M. Lim, and C. P. Lee, "Wide residual network for vision-based static hand gesture recognition." *IAENG International Journal of Computer Science*, vol. 48, no. 4, pp. 906–914, 2021.
[22] H. Zhang and J. Zhao, "Traffic sign detection and recognition based on deep learning." *Engineering Letters*, vol. 30, no. 2, pp. 666–673, 2022.
[23] H. T. Lim, H. G. Kim, and Y. M. Ro, "Learning based hole filling method using deep convolutional neural network for view synthesis," *Electronic Imaging*, vol. 28, pp. 1–5, 2016.
[24] C. Li, X. Sang, D. Chen, and D. Zhang, "Innovative hole-filling method for depth-image-based rendering (DIBR) based on context learning," in *Proceedings of SPIE - Optoelectronic Imaging and Multimedia Technology V*, vol. 10817, 2018, pp. 26–30.
[25] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–14, 2017.
[26] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 85–100.
[27] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 1–10.
[28] M. A. Hedjazi and Y. Genc, "Image inpainting using scene constraints," *Signal Processing: Image Communication*, vol. 93, no. 4, pp. 1–10, 2021.
[29] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "Structureflow: Image inpainting via structure-aware appearance flow," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 181–190.
[30] L. Liao, J. Xiao, Z. Wang, C.-W. Lin, and S. Satoh, "Uncertainty-aware semantic guidance and estimation for image inpainting," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 310–323, 2020.
[31] A. Bapat and J.-M. Frahm, "The domain transform solver," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6014–6023.
[32] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.