# Diffuse3D

Jiang, Yutao; Zhou, Yang; Liang, Yuan; Liu, Wenxi; Jiao, Jianbo; Quan, Yuhui; He, Shengfeng

*DOI:*
10.1109/ICCV51070.2023.00826

# Diffuse3D: Wide-Angle 3D Photography via Bilateral Diffusion

Yutao Jiang[1,4*], Yang Zhou[1,4*], Yuan Liang[1,4], Wenxi Liu[2], Jianbo Jiao[3], Yuhui Quan[1], and Shengfeng He[4†]
[1]School of Computer Science and Engineering, South China University of Technology
[2]College of Computer and Data Science, Fuzhou University
[3]School of Computer Science, University of Birmingham
[4]School of Computing and Information Systems, Singapore Management University

(a) Input (b) AdaMPI [6] (c) 3D Photo [26] (d) Ours

Figure 1: We propose a new novel 3D photography approach that generates 3D viewing experience from a single image using bilateral diffusion. It allows injecting depth information into the denoising diffusion probabilistic inference, and leads to superior performances in wide-angle synthesis compared with state-of-the-arts.

## Abstract

*This paper aims to resolve the challenging problem of wide-angle novel view synthesis from a single image, a.k.a. wide-angle 3D photography. Existing approaches rely on local context and treat them equally to inpaint occluded RGB and depth regions, which fail to deal with large-region occlusion (i.e., observing from an extreme angle) and foreground layers might blend into background inpainting. To address the above issues, we propose Diffuse3D which employs a pre-trained diffusion model for global synthesis, while amending the model to activate depth-aware inference. Our key insight is to alter the convolution mechanism in the denoising process. We inject depth information into the denoising convolution operation with bilateral kernels, i.e., a depth kernel and a spatial kernel, to consider layered correlations among pixels. In this way, foreground regions are overlooked in background inpainting and only pixels close in depth are leveraged. On the other hand, we propose a global-local balancing approach to maximize both contextual understandings. Extensive experiments demonstrate that our approach outperforms state-of-the-art methods in novel view synthesis, especially in wide-angle scenarios. More importantly, our method does not require any training and is a plug-and-play module that can be integrated with any diffusion model. Our code can be found at* https://github.com/yutaojiang1/Diffuse3D.

*Both authors contributed equally to this research.
†Corresponding author: shengfenghe@smu.edu.sg.

## 1. Introduction

In our life, images play an important role in carrying and sharing visual memories. Animating a still image can further enhance the immersive experience of an impressive moment. 3D photography [26, 13, 10] is proposed for this purpose to generate 3D viewing experiences from a single image, by interactively changing the camera angles. These methods are all based on modular systems and leverage state-of-the-art depth estimation, inpainting, and segmentation models to understand the layer structure and fill in the holes of occlusions. This component-wise strategy shows robust performance when dealing with in-the-wild scenes.

This line of research focuses on the inpainting quality of disocclusion caused by camera movements, through taking neighboring contextual information into account. However, due to the limited and narrow boundary context for inpainting, prior works [26, 10, 6] are confined to novel view synthesis with small camera view angle changes (generally 5-10 consecutive frames for one scene), which can hardly create a sense of immersion to meet the realistic demand. Extending the camera angle reveals larger holes that cannot be simply filled by repeating neighboring textures (see Figure 1b). Indeed, human could conjecture the occluded areas not only from the nearby visible areas but also by connecting to our visual memories (*i.e.*, image synthesis). On the other hand, the way they inpaint disocclusions neglects [26, 6] or implicitly considers [10] the depth prior, which may leak foreground semantics into background recovery (see Figure 1c).

In this paper, we present a new method, called Diffuse3D, to address the above problems. We employ an off-the-shelf pre-trained diffusion model as the generative prior, and we further empower it to be depth-aware by introducing bilateral convolution into the denoising inference process. Specifically, diffusion models can be conditioned by the masks and generate coherent image contents through consecutive denoising steps, but all pixels are weighted equally during the process. Inspired by bilateral filter [1, 29, 7] and depth-aware learning [32], we decompose the denoising convolution kernel into two parts, a spatial kernel that averages local regions, and more importantly a depth kernel to assign different weights to the pixels according to the depth similarities. In this way, depth prior is injected in the diffusion model and only the surrounding pixels located in the nearby layers are considered in the denoising steps. On the other hand, we introduce two levels of inpainting in the framework, the global and local ones, to obtain diverse contextual knowledge from two essentially different aspects.

Our proposed method is a plug-and-play design that can easily work with arbitrary diffusion models, and the idea of bilateral diffusion can be easily extended to other diffusion applications to include information other than depth. Extensive experiments demonstrate the superior 3D photography performance of our method compared to the state-of-the-art methods. Especially in the wide-angle setting, our approach can effectively differentiate the contextual information of foreground/background and produce more semantically meaningful inpainted regions (see Figure 1d).

## 2. Related Work

### 2.1. 3D Photography

The process of generating 3D viewing experience from a 2D image is referred as 3D photography. This series of methods surpasses previous multi-view synthesis approaches [23, 34, 36, 35], which are limited to generating a fixed number of angles, by enabling unconstrained 3D synthesis. It can be briefly classified into two categories, the models based on the end-to-end network and modular system. The former aims to synthesize novel views on multi-view image datasets. These methods typically take a single image as input and represent the scene with different representations in an end-to-end fashion, like multi-plane images [39, 30], mesh [9], and point cloud [33]. Their main limitation is that it heavily relies on the training data and therefore cannot generalize to in-the-wild datasets.

Modular-based methods [26, 10, 6, 18, 17] utilize the off-the-shelf monocular depth estimation, segmentation, or inpainting models to produce reliable 3D viewing effects regardless the input data domain. They concentrate more on synthesis quality and computation efficiency. Shih *et al*. [26] introduce an edge context guided depth and color inpainting model to recover nearby disocclusions using LDI representation [25]. Jampani *et al*. [10] employ an efficient soft layering representation and a segmentation based matting technique to capture finer appearance details. For mobile devices, Kopf *et al*. [13] present a low resource-consumption system with elaborate depth estimation method. To mitigate the difficulty in producing efficiency and high-dimensional multi-plane images, Han *et al*. [6] propose to learn the adaptive MPI depth which guides the prediction of image planes in an interactive way. Zhou *et al*. [40] develop a self-rectified strategy to construct pseudo-stereo pairs, converting the monocular synthesis problem to a stereo synthesis problem, thereby reducing the learning ambiguity. Our work falls into this category and leverages the off-the-shelf components. By contrast, we propose a depth-injected bilateral diffusion to inpaint disocclusions in a global and depth-aware manner.

### 2.2. Diffusion Model

The seminal work [27] presents a novel way to construct a flexible and computationally tractable probabilistic model, *i.e.*, diffusion model. This is achieved by using a diffusion process to gradually permute the known initial distribution and then to learn a progressive denoising model to recover

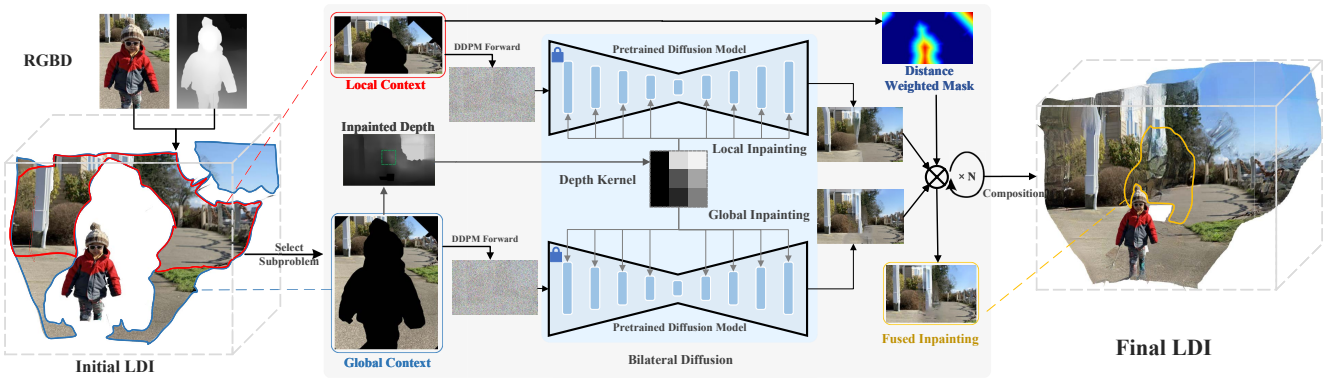



Figure 2: Overview of our framework. We take an RGB image and the corresponding depth map predicted using an off-the-shelf estimator as inputs, to generate 3D viewing experience. Specifically, we first convert the scene into layered depth images (LDI) and then select both local and global disocclusions for inpainting using our bilateral diffusion model. Depth kernel is injected into the denoising process during inference. Global and local inpaintings are then fused together to obtain the final result.

the original distribution. Because of high flexibilities, diffusion models have shown compelling results in image synthesis [8, 28, 22, 2], audio synthesis [12, 16, 14], and many other applications [15, 2, 11]. Our aim is to provide a depth-aware diffusion model in which a larger context inpainting can be implemented for the large area of disocclusion. Instead of directly applying the conditions as input for the encoder-decoder denoising models [22, 2], we opt to incorporate the depth prior into the multi-level spatial-wisely performed convolution layers, which alters the probabilistic representation in a bilateral manner.

# 3. Method

## 3.1. Overview

**Preliminaries.** Layered Depth Images (LDI) [25] represents a 3D scene by a layer-based representation. In LDI, each pixel lattice contains arbitrary number of color and corresponding depth values, *i.e.*, the foreground objects and occluded objects in the background. The sparse structure of LDIs makes it favorable to store a 3D scene and render views efficiently. [26] exploits a modified LDI with explicit local connectivity to achieve single-view 3D photography. In particular, they first create a trivial LDI from the RGBD input and determine the occluded regions along depth edges, which they utilize an edge inpainting network to make it more complete. Next, they propose to do context-aware inpainting on the occluded regions and then put them back to the original LDI. Consequently, novel views can be rendered from the inpainted LDI.

**Pipeline.** Figure 2 shows the overall pipeline of our framework. Given an RGB image with the depth predicted by an off-the-shelf depth estimator, following previous LDI-based solution [26], our method first constructs an initial layered depth images and then inpaint the occlusion areas by iteratively selecting subproblem from the initial LDI. Each subproblem has its own target disocclusion region, with the guidance of the corresponding contextual/referencing area along the depth edges (*i.e.*, discontinuities between layers) generated by the flood-fill algorithm. Specifically, we propose depth-injected bilateral diffusion with cross-layer inpainting strategy for the aforesaid inpainting, so as to facilitate wide-angle 3D photography. After integrating the inpainted results back to the initial LDI, we are able to render novel views from 3D scene representation. In the following sections, we describe the bilateral diffusion and cross-layer inpainting in detail.

## 3.2. Bilateral Diffusion

The key of our method is to inpaint faithful image content in the occluded regions. This synthesis task can be accomplished by diffusion models, owing to its powerful ability on image inpainting. For a diffusion model, its forward process is to sequentially add Gaussian noise to the input $\boldsymbol{x}$ with variance schedule $\{\beta_t\}_{t=0}^{T}$ in $T$ steps, resulting in a noise $\boldsymbol{x}_T$:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) := \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}), \qquad (1)$$

which is essentially a Markov chain. Next, $\boldsymbol{x}_t$ of the arbitrary step $t > 0$ can be derived from $\boldsymbol{x}_0$ as:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \qquad (2)$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. In turn, the reverse process of a diffusion model is to gradually remove noise

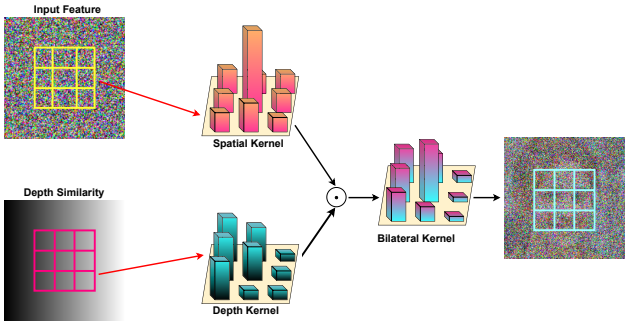



Figure 3: Illustration of the depth and spatial kernels in our bilateral diffusion. We show the convolution step of the denoiser with kernel size $3 \times 3$. Depth kernel is computed according to the depth similarities to the center pixels. Here we show the spatial kernel with the Gaussian function, and it is combined with the depth kernel to assign the final weights for the convolution operation.

from $\boldsymbol{x}_t$ and generate $\boldsymbol{x}_0$ finally:

$$p_\theta\left(\boldsymbol{x}_{0:T}\right) := p\left(\boldsymbol{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right), \tag{3}$$

$$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) := \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)\mathbf{I}). \tag{4}$$

In specific, the diffusion model iteratively runs the denoising autoencoder $\epsilon_\theta(\boldsymbol{x}_t, t)$ that predicts $\boldsymbol{x}_{t-1}$ from $\boldsymbol{x}_t$, *i.e.*,

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)). \tag{5}$$

Here we select the state-of-the-art model, Latent Diffusion [22] as our inpainting model. In general, diffusion models aim to iteratively predict a denoised variant from a noisy version of the input image. Usually the conditional denoising model is an autoencoder, it takes a noisy input $x$ and condition at the iteration step $t$ to produce the denoised output at the iteration step $t-1$. By repeating the denoising process for $T$ steps, the final output $y$ can be obtained. In particular, for the $l^{th}$ layer in the denoising network, a specific pixel at the spatial coordinate $p$ goes through the denoising process as followed:

$$y(p) = \sum_{p_i \in \Omega} x\left(p_i\right) g_s\left(\left\|p_i - p\right\|\right) \tag{6}$$

where $\Omega$ is the window centered at $p$ and $g_s$ is the spatial kernel of denoising model.

However, the convolution mechanism in denoising autoencoder of Latent Diffusion is of spatial invariance, and it treats the pixels of different depths equally, resulting in synthesis content of color-depth inconsistency near the depth edges. To relieve this concern, inspired by bilateral filter [1, 29, 7] and depth-aware learning [32], we present a depth-injected bilateral diffusion which leverage depth information for image inpainting to preserve color-depth consistency. Figure 3 shows the bilateral diffusion. In specific, we take the kernel of convolution as the spatial kernel similar in bilateral filter. Depth differences serve as the depth kernel to reweight the filter, so that depth information is injected into the convolution of diffusion process.

For the proposed bilateral diffusion, we start by calculating depth kernel from the depth at $p$ and its neighbors $p_i \in \Omega$, *i.e.* the depth differences. Then, we can apply the depth kernel to reweight the aforementioned filter:

$$y(p) = \sum_{p_i \in \Omega} x\left(p_i\right) g_s\left(\left\|p_i - p\right\|\right) f_d\left(\left\|D\left(p_i\right) - D(p)\right\|\right), \tag{7}$$

where $f_d$ is the depth kernel and $D$ denotes the depth. In practice, we apply Gaussian for the depth kernel $f_d$:

$$f_d\left(\left\|D\left(p_i\right) - D(p)\right\|\right) = \exp(-\frac{\left\|D\left(p_i\right) - D(p)\right\|^2}{2\sigma_r^2}), \tag{8}$$

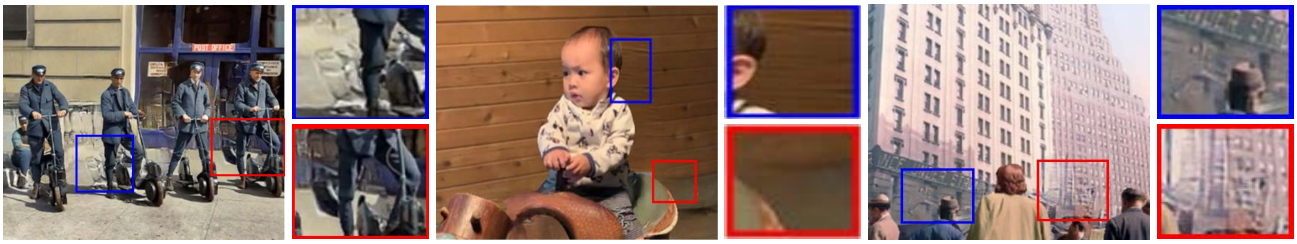
where $\sigma_r$ adjusts the depth differences. In this way, we empower the bilateral diffusion to concern more on the pixels with similar depth, leading to a color-depth consistent inpainted results.

As illustrated in Figure 2, we first select one edge from the depth edges and then expand the synthesized regions and context regions along the edge by the flood-fill algorithm. The context regions are the visible background nearby the depth edge, while the synthesized regions refer to the inpainted regions behind the foreground. After that, a depth inpainting network [26] is employed to synthesize depth $D$ on the occluded regions. Unlike [26] that carries out color inpainting independently from depth, we take advantage of depth to encourage the color inpainting results with color-depth consistency.

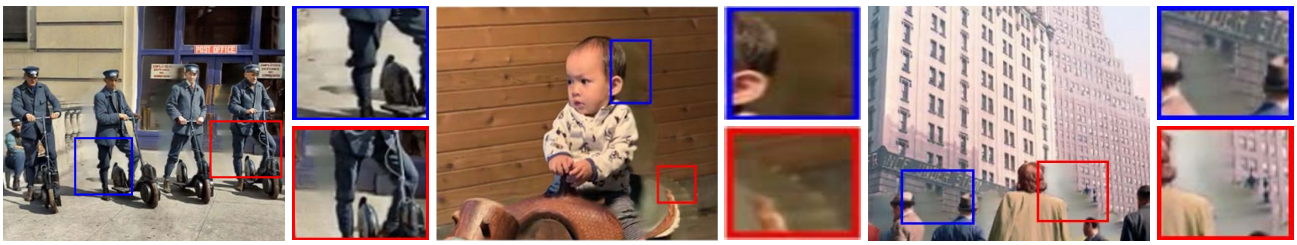
### 3.3. Cross-layer Inpainting

To better model the global and local relationship for inpainting, we come up with a cross-layer inpainting strategy. Similar to [31], we extend the context regions from the layer close in depth to all the layers farther than the current layer. It is intuitive that the foreground layers may contain irrelevant content that may mislead the inpainting of background, so that they should be excluded from the context regions. Besides, rather than concatenating all context regions and inpainting them together, we feed them into the local branch and global branch separately to alleviate the large change on depth of background. At last, we merge the outputs from both branches and obtain the final result.

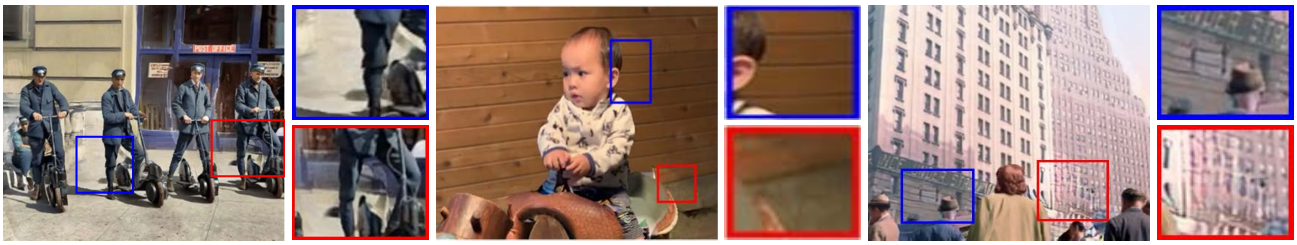
Specifically, when we process the context regions, we regard the regions close in depth and synthesis regions as



(a) 3D Photo



(b) AdaMPI



(c) Ours

Figure 4: **Qualitative comparisons with state-of-the-art methods.** (a) 3D Photo blends foreground into the background inpainting. (b) AdaMPI tends to produce blurry inpainted results. (c) Our method synthesizes realistic and geometrically consistent content image content in the scenario of wide-angle viewpoint changes.

the local information that considers the neighboring pixels only, and the other regions far in depth as the global reference. We apply the inpainting model to each branch independently, and fuse the inpainted images into a composite image by a distance-based weighted fusion, similar to [5], as below.

$$w = \frac{d_{edge}}{max(d_{edge})}, y = w \cdot y_{\text{local}} + (1 - w) \cdot y_{\text{global}} \quad (9)$$

where $w$ is the weight for the local-global fusion and $d_{edge}$ denotes the distance between the occluded pixel and the depth edge. With this cross-layer operation for global and local processing, we are able to produce more realistic inpainting results, and further boost the performance of novel view synthesis.

### 3.4. Implementation Details

For the depth inpainting model, we leverage the pretrained model from [26]. We revise the pretrained latent diffusion model (LDM) in [22] as our color inpainting model. In detail, we replace the denoising kernel in the denoising autoencoder with our proposed bilateral diffusion kernel. In purpose of sharing the depth consistency in information contraction and expansion, we perform this operation in both input block and output block of the denoising autoencoder. Besides, we adopt distance-transform in [3] for distance function in Section 3.3.

## 4. Experiment

### 4.1. Experimental Setup

We implement our method using Pytorch [19] and evaluate it on a single Nvidia GeForce RTX 3090. Our approach can achieve 3D photography without any training, and the running time mainly depends on the pre-trained diffusion model. When applying stable diffusion, it takes 8-15 mins to generate all the viewpoints (including 240 frames). For reference, state-of-the-art method 3D Photo [26] takes 5 mins to render all the frames. Our execution time can be accelerated by adopting a more efficient diffusion model.

**Metrics.** As for the dataset, the RealEstate-10K [39] dataset contains a large number of video clips. We ran-

domly sample 100 video clips from the test set for evaluation. To evaluate our performance in wide-angle scenarios, we use a more challenging setting than previous methods. Rather than using an interval of 10 frames as in 3D Photo [26], we use the frame t = 0 as the source view and the frame t = 20 as the target view to cope with wide-angle movement. In other words, the synthesized angles are approximately two times larger than the previous setting. We use the SSIM, PSNR and LPIPS [38] scores to quantitatively measure the quality of the novel views. The image resolution is set to $384 \times 512$ when quantitatively compare. Note that we crop 5% border when we calculate the metrics, following the setting of [6].

**Baselines.** We mainly compare our method with two state-of-the-art methods with publicly available source code: 3D Photo [26] and AdaMPI [6]. For 3D Photo, we use the pretrained model provided by the authors. For AdaMPI, the performance of AdaMPI is related to the number of planes $N$, we use the pretrained model with 64 planes for the best performance. As a fair comparison, all those methods use the same depth map predicted by DPT [20].

## 4.2. Comparisons

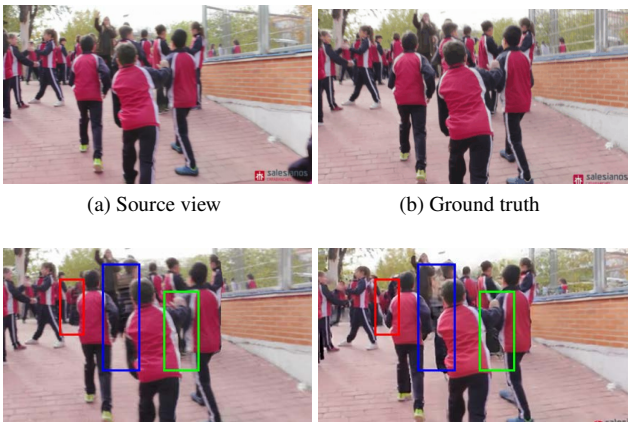
### 4.2.1 Qualitative Evaluation

For qualitative comparisons, we use the photos provided by 3D Photo [26], which consist of challenging examples in the wild. Figure 4 shows some examples of them. While other methods may be able to synthesize the correct contents in some disoccluded regions, they often produce artifacts and blurs in the discontinuous regions between the foreground and background when there is a wide camera angle change. In contrast, our method, which benefits from the proposed bilateral diffusion, is able to reduce the misdirection from the foreground and prioritize the background. This results in images that are both faithful to the original and geometrically consistent.

We also show the comparison with another state-of-the-art method SLIDE [10]. Since the code of SLIDE is not publicly available, we qualitatively compare the result from their paper, as shown in Figure 5. SLIDE shows blurry and repetitive patterns in highlighted regions, while our method shows rich and harmonious details.

### 4.2.2 User Study

We conducted a user study to evaluate the subjective image quality of the novel views generated by our proposed method and other comparison methods. Fifty participants were recruited and each was asked to rate the quality of 30 images produced by each method, resulting in 90 images rated per participant. The images were presented in a random order to eliminate potential biases in the participants' ratings. Participants assigned a score between 1 and 5, with 1 indicating the worst image quality and 5 indicating the best. The ratings were averaged across all participants to obtain a mean opinion score (MOS) for each method.



(a) Source view (b) Ground truth

(c) SLIDE [10] (d) Ours

Figure 5: Qualitative comparison with SLIDE [10]. Note that the result is directly obtained from their paper. Our method shows rich and clear patterns.

Table 1: Quantitative Comparisons with state-of-the-art methods. ↑ denotes the higher the better and ↓ denotes the opposite. Column E.S. denotes the metrics calculated in the entire scene, and column D.R. denotes the metrics calculated in disocclusion regions. The best results are marked in **bold**. The proposed method outperforms the robust 3D Photo, and achieves the best perceptual score (LPIPS) among all the competitors.

| Methods | PSNR ↑ | | SSIM ↑ | | LPIPS ↓ | |
|---|---|---|---|---|---|---|
| | E.S. | D.R. | E.S. | D.R. | E.S. | D.R. |
| 3D Photo | 19.80 | 21.78 | 0.691 | 0.889 | 0.153 | 0.111 |
| AdaMPI | **22.57** | **24.42** | **0.785** | **0.915** | 0.185 | 0.098 |
| Ours | 20.72 | 23.56 | 0.726 | 0.894 | **0.132** | **0.083** |

### 4.2.3 Quantitative Evaluation

To quantitatively compare with state-of-the-art methods, 3D Photo [26] and AdaMPI [6], we conduct experiments on the testset of RealEstate-10K [39] and calculate the metrics on two different scales: the entire scene and only the disocclusion region. The quantitative result is shown in Table 1. In particular, it is worth noting the comparison with 3D Photo since both methods are based on the same LDI-based 3D representation. Our proposed method outperforms 3D Photo in all three metrics, indicating that the proposed bilateral diffusion approach is more effective than the inpainting method used in 3D Photo. While AdaMPI leads in PSNR and SSIM, it performs worse in LPIPS due to its tendency to generate blurry content in the disoccluded regions, which

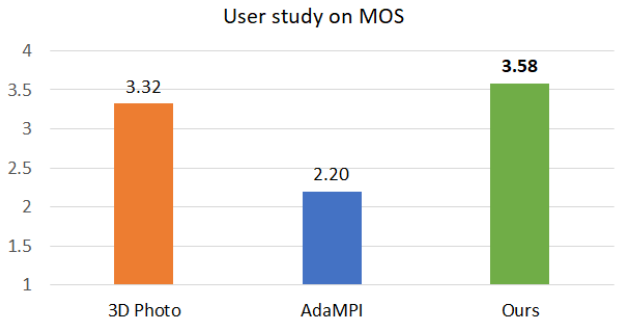



Figure 6: **User study on mean-opinion-score.** Our method outperforms other methods due to less artifacts in synthesized images.

Table 2: **Ablation Study on different variants of our method.** ↑ denotes the higher the better and ↓ denotes the opposite. Best results are marked in **bold**. All the components contribute to the final performance.

| Methods | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| w/o bilateral diffusion | 19.83 | 0.690 | 0.151 |
| w/o local context | 20.02 | 0.707 | 0.145 |
| w/o global context | 20.32 | 0.711 | 0.136 |
| Ours | **20.72** | **0.726** | **0.132** |
| Input block | 20.41 | 0.718 | 0.136 |
| Output block | 20.58 | 0.720 | 0.135 |

is preferred by the first two pixel-level metrics. This is also evident in super-resolution evaluation [4], where a blurry image usually has a higher pixel-wise score, but is detected by the perceptual metric LPIPS. Our results demonstrate a much better perceptual score than AdaMPI, indicating that bilateral diffusion improves the generation of realistic and coherent image content. Moreover, the cross-layer inpainting leverages multi-level information to ensure consistent synthesis even when there are wide camera angle changes.

As shown in Figure 6, our proposed method achieves the highest MOS score, outperforming the other comparison methods. Participants found our method to produce images with fewer artifacts and a higher degree of realism compared to the other methods. These results are consistent with our objective evaluations, which show that our method generates images that are more faithful to the original and geometrically consistent than the other methods.

### 4.3. Ablation Study

To validate the performance of the proposed bilateral diffusion and cross-layer inpainting, We conduct qualitative and quantitative experiments to compare our complete method with the other three variants: 1) w/o bilateral diffusion, inpainting with traditional spatial diffusion kernel; 2) w/o local context, inpainting without local reference content; 3) w/o global context, inpainting without global reference content.

(a) Source view (b) w/o bilateral diff. (c) Ours

Figure 7: **Qualitative ablation comparisons of bilateral diffusion.** (b) shows that without the injected depth information, background or out-of-view inpainting suffers badly from the ambiguous surrounding context. (c) recovers both background and out-of-view regions well.

The results in Table 2 demonstrate that incorporating any of the three components leads to an improvement in the quantitative measurements. In particular, the use of bilateral diffusion has a greater impact on the novel view synthesis, indicating that depth information is a critical prior for color inpainting. By replacing the spatial diffusion kernel with a bilateral diffusion kernel, the depth prior can help the diffusion model to focus on areas with similar depth, resulting in more accurate geometry and realistic generation. Moreover, the negative impact of the lack of context information on the final synthesis view provides evidence of the effectiveness of cross-layer inpainting, especially the local context. The local context refers to the areas closest in depth to the occluded regions. By adopting both strategies, we achieve the best performance on the novel view synthesis.

Figure 7 provides a qualitative comparison between our model with bilateral diffusion and the variant without it, which only uses spatial diffusion kernel. The results show that the model without bilateral diffusion generates distorted images, whereas the bilateral diffusion is able to generate color-depth consistent results in the disoccluded regions. This is evident in the ground behind the astronaut


(a) Source view (b) Global (c) Local (d) Ours

Figure 8: **Qualitative ablation comparisons of cross-layer inpainting.** (b) and (c) show two different types of inpaintings with different emphasis. Our method (d) can combine their advantages.

and the background sheltered from the shrubs in the source view. The bilateral diffusion effectively exploits the information of visible background close in depth. Furthermore, Figure 8 illustrates that combining global and local information leads to the best visual results by utilizing multi-level information.

We conduct another experiment to determine the optimal location for injecting the depth kernel, comparing three methods: bilateral diffusion only in the input block, only in the output block, and in both input and output blocks. The quantitative results presented in Table 2 show that the complete method, where bilateral diffusion is used in both input and output blocks, achieves the best results. This suggests that our approach benefits from depth consistency in both feature contraction and expansion, leading to superior performance.

## 4.4. Discussions

### 4.4.1 Challenging Scenarios

**Wide-angle Movement Evaluation.** To showcase the effectiveness of our method in handling wide-angle camera movements, we present the variation of synthesis quality as the view changes. For ease of analysis, we utilize spatial camera displacements to represent camera movements.

Figure 9 reveals that while 3D Photo [26] and AdaMPI [6] are capable of generating results with small angle changes, the synthesis quality rapidly deteriorates when faced with wide-angle settings. In contrast, our proposed method not only performs well with narrow camera movements, but also exhibits promising results when the camera angle changes extensively. This indicates that bilateral diffusion effectively enforces the denoising model to focus on pixels that are close in depth, and filter out irrelevant foreground information, resulting in improved synthesis quality. Additionally, our proposed global-local inpainting strategy further enhances synthesis quality.

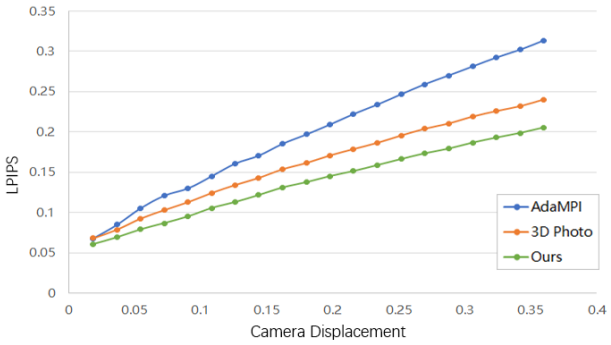
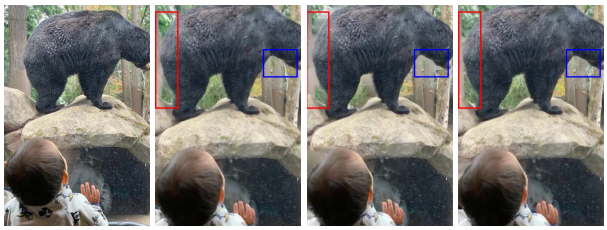



Figure 9: **Comparison of wide-angle movement.** Our method can achieve superior results with wide-angle changes (the lower the better).



(a) Source View (b) 3D Photo (c) AdaMPI (d) Ours

Figure 10: Our method can handle multiple foregrounds.

(a) Source view (b) Latent diff. (c) Palette (d) RePaint

Figure 11: Our method is flexible and plug-and-play that can be applied to any diffusion model.

**Scene with Multiple Foregrounds.** Figure 10 shows a challenging case with multiple foregrounds. The result demonstrates our method understands the geometrical relations well and produces more realistic results than others.

### 4.4.2 Bilateral Diffusion with Different Backbones

The proposed bilateral kernel is a versatile plug-and-play module that can be seamlessly integrated with any diffusion model. To showcase its compatibility, we present additional

(a) Source view (b) DPT (c) Midas (d) NewCRFs

Figure 12: Our method is capable to handle depth maps from different sources.

results in Figure 11 using various diffusion models such as Latent diffusion [22], Palette [24], and RePaint [15]. The figures illustrate that our bilateral kernel can effectively collaborate with different diffusion models, thereby extending the scope of our method's applicability.

#### 4.4.3 Tolerability of Different Depth maps

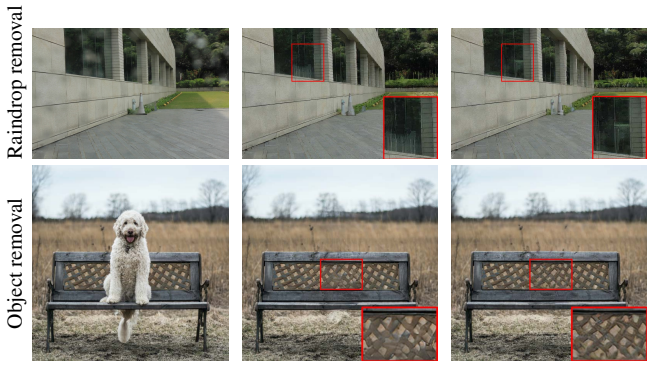
We conducted an experiment to assess the effectiveness of depth maps obtained from different depth estimators. Specifically, we evaluated our model using depth maps estimated from three different methods, namely DPT [20], Midas [21], and NewCRFs [37]. The results are presented in Figure 12, and show that our method can adapt to the variations in depth maps and produce plausible results.

### 4.5. Applications on Other Tasks

We propose the first feasible attempt to intervene convolutional operations of a pretrained diffusion model. Our idea can be easily applied to other applications. We show two applications of our proposed bilateral diffusion, depth-aware raindrop and object removals, in Figure 13. In the task of raindrop removal, our proposed method removes raindrops well and synthesizes clear content in covered regions (the top row of Figure 13c). However, the diffusion model without bilateral weights generates blurry results (the top row in Figure 13b). Besides, the task of object removal shows similar results. In the bottom row of Figure 13b, the traditional diffusion model generates blurs and discontinuous geometry. In contrast, with the assistance of bilateral weights, our method generates images that are more faithful and geometrically consistent (Figure 13c).

## 5. Conclusion

In this paper, we resolve the problem of wide-angle 3D photography from a pure depth-aware synthesis perspective. In particular, we employ a diffusion model as the generative prior for holistic inpainting, and we further alter the




(a) Input (b) w/o bilateral weights (c) Ours

Figure 13: Applications on raindrop removal and object removal. Our incorporated bilateral weights can produce better results against traditional diffusion.

diffusion inference process by proposing a bilateral diffusion that takes depth into the denoising consideration. We reformulate the convolution operations of the denoisers into two components, a spatial kernel that assigns equal weights to all the samples, and a depth kernel that weights pixels differently according to the depth similarities. As a consequence, inpainting the disocclusions will not show foreground leakage because of our depth-aware diffusion. We further involve both global and local contexts in disocclusion inpainting. Our method is a plug-and-play design that can easily be applied in arbitrary diffusion models to include not only depth information. We show superior performances in wide-angle 3D photography.

**Limitation.** Our method shares the same limitation as in diffusion models, and one is the heavy computational cost. Although our method does not involve any training, which is a big advantage, it requires many iterations of the denoising steps. On the other hand, our method can be easily integrated into a new efficient diffusion model in the future. Besides, our method also shares similar limitations with LDI. Our entire framework has a high tolerance to depth variations, but when the depth map is particularly inaccurate, the generation quality will be degraded.

**Potential Negative Impact.** Our main contribution is a method to improve the quality of 3D photography, which could impact automation. As such, our work inherits the general ethical risks of Al, like the question of how to address the potential of increased automation in society.

**Acknowledgement.** This paper is partially supported by the National Natural Science Foundation of China (No. 61972162); Guangdong Natural Science Funds for Distinguished Young Scholars (No. 2023B1515020097); and Singapore Ministry of Education Academic Research Fund Tier 1 (MSS23C002).

## References


[1] Volker Aurich and Jörg Weule. Non-linear gaussian filters performing edge preserving diffusion. In *Mustererkennung 1995, 17. DAGM-Symposium*, pages 538–545, 1995. 2, 4

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, pages 18208–18218, 2022. 3

[3] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 5

[4] Kelvin C.K. Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, pages 14245–14254, 2021. 7

[5] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *CVPR*, 2022. 5

[6] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Single-view view synthesis in the wild with learned adaptive multi-plane images. In *SIGGRAPH*, 2022. 1, 2, 6, 8

[7] Shengfeng He, Qingxiong Yang, Rynson WH Lau, and Ming-Hsuan Yang. Fast weighted histograms for bilateral filtering and nearest neighbor searching. *IEEE TCSVT*, 26(5):891–902, 2015. 2, 4

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 3

[9] Ronghang Hu, Nikhila Ravi, Alexander C Berg, and Deepak Pathak. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *ICCV*, pages 12528–12537, 2021. 2

[10] Varun Jampani, Huiwen Chang, Kyle Sargent, Abhishek Kar, Richard Tucker, Michael Krainin, Dominik Kaeser, William T Freeman, David Salesin, Brian Curless, et al. Slide: Single image 3d photography with soft layering and depth-aware inpainting. In *ICCV*, pages 12518–12527, 2021. 2, 6

[11] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, 2022. 3

[12] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*, 2020. 3

[13] Johannes Kopf, Kevin Matzen, Suhib Alsisan, Ocean Quigley, Francis Ge, Yangming Chong, Josh Patterson, Jan-Michael Frahm, Shu Wu, Matthew Yu, et al. One shot 3d photography. *ACM TOG*, 39(4):76–1, 2020. 2

[14] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *AAAI*, volume 36, pages 11020–11028, 2022. 3

[15] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pages 11461–11471, 2022. 3, 9

[16] Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. *arXiv preprint arXiv:2103.16091*, 2021. 3

[17] Fangzhou Mu, Jian Wang, Yicheng Wu, and Yin Li. 3d photo stylization: Learning to generate stylized novel views from a single image. In *CVPR*, pages 16273–16282, 2022. 2

[18] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM TOG*, 38(6):1–15, 2019. 2

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035. 2019. 5

[20] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021. 6, 9

[21] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3):1623–1637, 2020. 9

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3, 4, 5, 9

[23] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *ICCV*, pages 14356–14366, 2021. 2

[24] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 9

[25] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *SIGGRAPH*, pages 231–242, 1998. 2, 3

[26] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *CVPR*, pages 8028–8038, 2020. 1, 2, 3, 4, 5, 6, 8

[27] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265, 2015. 2

[28] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020. 3

[29] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *ICCV*, pages 839–846, 1998. 2, 4

[30] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, pages 551–560, 2020. 2

[31] Qianqian Wang, Zhengqi Li, David Salesin, Noah Snavely, Brian Curless, and Janne Kontkanen. 3d moments from near-duplicate photos. In *CVPR*, 2022. 4

[32] Weiyue Wang and Ulrich Neumann. Depth-aware cnn for rgb-d segmentation. In *ECCV*, pages 135–150, 2018. 2, 4

[33] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, pages 7467–7477, 2020. 2

[34] Cheng Xu, Keke Li, Xuandi Luo, Xuemiao Xu, Shengfeng He, and Kun Zhang. Fully deformable network for multi-view face image synthesis. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2

[35] Xuemiao Xu, Keke Li, Cheng Xu, and Shengfeng He. Gdface: Gated deformation for multi-view face image synthesis. In *AAAI*, volume 34, pages 12532–12540, 2020. 2

[36] Yangyang Xu, Xuemiao Xu, Jianbo Jiao, Keke Li, Cheng Xu, and Shengfeng He. Multi-view face synthesis via progressive face flow. *IEEE TIP*, 30:6024–6035, 2021. 2

[37] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3916–3925, 2022. 9

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[39] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM TOG*, 37(4):1–12, 2018. 2, 5, 6

[40] Yang Zhou, Hanjie Wu, Wenxi Liu, Zheng Xiong, Jing Qin, and Shengfeng He. Single-view view synthesis with self-rectified pseudo-stereo. *International Journal of Computer Vision*, pages 1–12, 2023. 2