StereoDiffusion: Training-Free Stereo Image Generation Using Latent Diffusion Models

Lezhong Wang Jeppe Revall Frisvad Mark Bo Jensen Siavash Arjomand Bigdeli
Technical University of Denmark, Kongens Lyngby
{lewa, jerf, mboje, sarbi}@dtu.dk

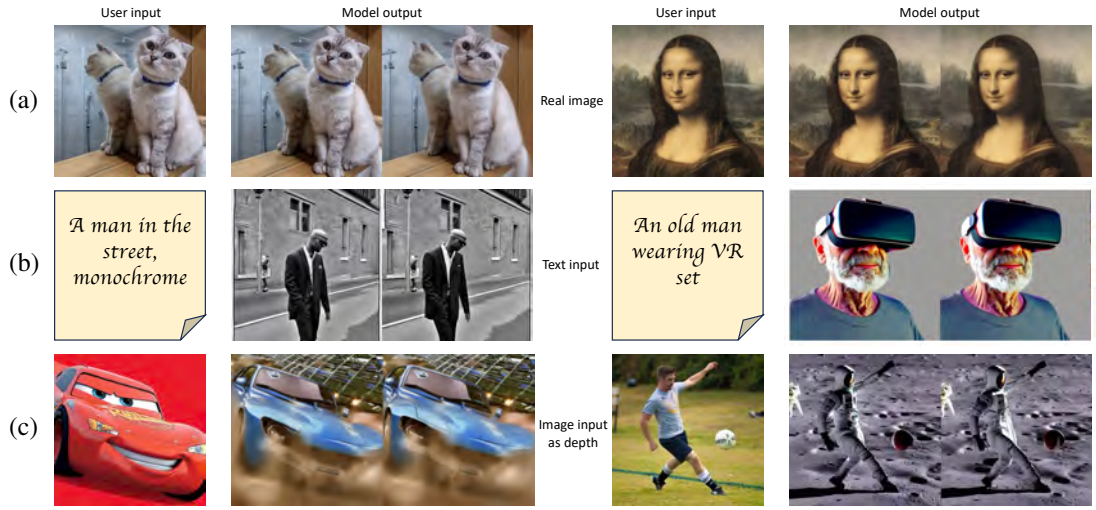

Figure 1. Our method takes one of three types of user input and generates a stereo image. Accepted user inputs: (a) a photo, (b) a text prompt, or (c) a user's image as a depth map and a prompt. We use a latent diffusion model pretrained on either images (a, b) or depth maps (c).

Abstract

The demand for stereo images increases as manufacturers launch more extended reality (XR) devices. To meet this demand, we introduce StereoDiffusion, a method that, unlike traditional inpainting pipelines, is training-free and straightforward to use with seamless integration into the original Stable Diffusion model. Our method modifies the latent variable to provide an end-to-end, lightweight method for fast generation of stereo image pairs, without the need for fine-tuning model weights or any post-processing of images. Using the original input to generate a left image and estimate a disparity map for it, we generate the latent vector for the right image through Stereo Pixel Shift operations, complemented by Symmetric Pixel Shift Masking Denoise and Self-Attention Layer Modifications to align the right-side image with the left-side image. Moreover, our proposed method maintains a high standard of image quality throughout the stereo generation process, achieving state-of-the-art scores in various quantitative evaluations.

1. Introduction

Large-scale language-image (LLI) models have become prominent in recent years, acclaimed for their advanced generative semantic and compositional abilities [6, 21, 25, 29, 45]. Their distinctiveness lies in their training on extensive language-image datasets, enabling them to interpret and generate content from diverse linguistic and visual contexts. Utilizing innovative image generative techniques such as auto-regressive and diffusion models [10] have significantly advanced the synergy between linguistic understanding and image generation. This has led to a new era of creative and semantically rich image synthesis, marking notable advancements in artificial intelligence and computer vision.

A significant recent development in the VR/AR field is Apple's introduction of Vision Pro, which has the potential to drive rapid advancements in this field. Despite the growing production of 3D content by various manufacturers, and related research [14, 15, 28, 41, 43] in recent years, the availability of stereo multimedia content, which offers a depth-enhanced visual experience, remains relatively scarce. As the VR/AR era looms, the limitations of existing image

© 2024 IEEE. This is the authors' version of the work.

Wang, L., Frisvad, J. R., Jensen, M. B., and Bigdeli, S. A. StereoDiffusion: training-free stereo image generation using latent diffusion models. In *Proceedings of Computer Vision and Pattern Recognition Workshop (CVPRW 2024)*. IEEE, June 2024. To appear.

generation models that are confined to producing 2D images become increasingly apparent. However, there is currently no relevant research that attempts to use image generation models to directly generate stereo image pairs. In response to this challenge, we introduce a novel methodology. Through modification of the Stable Diffusion model's latent variable, we have devised an efficient end-to-end approach, eliminating the need for additional models like inpainting [12, 31] for post-processing to generate stereo images. Figure 1 presents some examples. Code is available at https://github.com/lez-s/StereoDiffusion.

We address the constraints of traditional image generation models that employ an inpainting pipeline. Our approach is to generate stereo image pairs by adjusting the latent variable of the Stable Diffusion model, see Figure 2. We use Symmetric Pixel Shift Masking Denoise and Self-Attention layer modifications to align the generated right-side image with the left-side image. This method allows for a lightweight solution that can be seamlessly integrated into the original Stable Diffusion model without the need for fine-tuning. To the best of our knowledge, our approach represents the first instance of generating stereo images by modifying the latent variable of Stable Diffusion. Compared with other methods, our approach enables the training-free end-to-end rapid generation of high-quality stereo images using only the original Stable Diffusion model.

2. Related work

Latent space of a Latent Diffusion Model. Diffusion models, notably the Denoising Diffusion Implicit Models (DDIM) [32], have made significant strides in image generation. The DDIM sampling algorithm revealed that using the same initial noise results in consistent high-level features across different generative paths, highlighting initial noise as a potent latent image encoding [32]. This discovery aids in modifying images by adjusting the Stable Diffusion latent variable. A key challenge in stereo image generation is maintaining content consistency between paired images. Researchers are focusing on image editing techniques using Stable Diffusion [3, 22, 35, 44], such as the "prompt-to-prompt" method [10], which involves altering the model's cross-attention during sampling for text-prompt-based image editing. ControlNet [46] is also a notable work in the field, but instead of utilizing the latent space, the authors trained ControlNet on a large dataset to better control the generation of desired images by Stable Diffusion. Additionally, ControlNet primarily focuses on pose control and lacks the capability for pixel-level modifications of images. Although effective, these methods are less suited for tasks needing precise pixel-level manipulation, like stereo image generation, due to their reliance on text prompts for image modification.

Video generation by Latent Diffusion Model. Ensuring the consistency of images within the same batch in Stable Diffusion has long been a challenge in video generation [2, 37]. VideoComposer addresses this by incorporating an STC-encoder into the Latent Diffusion Model's U-Net architecture, ensuring consistency in the generated image content [37]. Similarly, VideoLDM achieves impressive video generation results by introducing 3D convolution layers and temporal attention layers into the spatial and temporal layers of U-Net [2]. However, these methods require fine-tuning of the original models and substantial amounts of data. Generally, this is not an issue for video generation, but for achieving stereo image generation, the available stereo image data is quite limited, mostly comprising road traffic images initially intended for autonomous driving depth prediction services. In the video generation field, attempts such as Tune-A-Video [42] have explored zero-shot video generation [36, 42]. This work utilizes a technique called ST-Attn to maintain the continuity of videos. We employ a comparable approach to ensure consistency between the left and right images.

3D photography and inpainting. Traditional image-based reconstruction and rendering methods require complex capture setups, involving numerous images with significant baselines [8, 9, 13, 40]. Currently, limited research endeavors directly focus on generating stereo images. Many studies have concentrated on generating 3D photos, a technique allowing subtle changes in the camera angle for observing photos from different perspectives [12, 18, 31, 34, 47]. Among 3D image generation techniques, 3D Photography Inpainting is a notable approach [8, 31]. This method employs inpainting to generate 3D images. After passing the input image through a depth estimation model, they map the image onto a mesh and apply changes in perspective based on the depth map of the original image. Inpainting is then utilized to fill the gaps left by transformed pixels in the original image. This approach significantly differs from our modification of the Stable Diffusion latent space. Although this method could be adopted as post-processing of an image generated through Stable Diffusion, it requires additional steps and consumes more time.

3D scene generation by pretrained Stable Diffusion. Recent studies have used model distillation with pretrained Stable Diffusion models for text-based 3D reconstruction. DreamFusion [23] employes 'Score Distillation Sampling' (SDS) to initialize and render a NeRF model, improving with Imagen-surrounding score distillation loss. Variational Score Distillation (VSD) [38] further enhances 3D scene quality. These methods can create stereo images via rendering but are time-intensive. Without full 3D scenes, our method provides a faster solution for generating stereo image pairs.

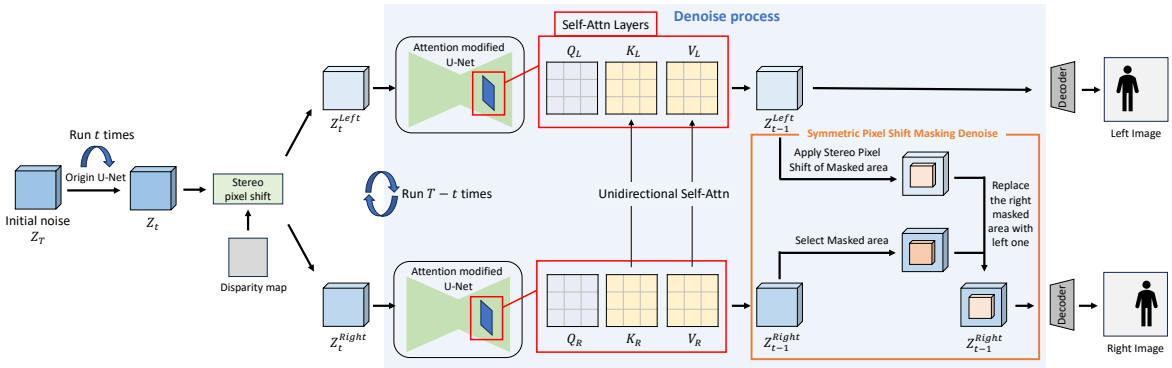

Figure 2. The pipeline of our Stereo Diffusion. The process starts with random noise and denoising of it to generate a stereo image pair. The operation of Stereo Pixel Shift is represented by Eq. 3. The Disparity Map for generating stereo image pairs can be obtained from depth models such as DPT [26] or MiDas [27]. The pipeline only shows the Unidirectional Self-Attention operation, designed to align the right-side image with the left-side image, a method that satisfies general needs. Bidirectional Self-Attention, being a mutual operation, would be represented by bidirectional arrows in the image. The orange box in the image depicts the concept of Symmetric Pixel Shift Masking Denoise, with details explained in Sec. 3.2. The cross attention part of the sampling process is omitted for brevity.

3. Method

Diverging from conventional inpainting methods, our approach is distinctively simple and training-free. With seamless integration into the original Stable Diffusion framework, we provide end-to-end generation of stereo image pairs, eliminating the need for post-processing. Our method leverages a disparity map in the early denoising stage to apply a Stereo Pixel Shift (Sec. 3.1) to the latent vector of the left image. This process generates the latent vector for the right image through disparity. To address the inconsistency issues between the left and right images during the denoising process, we employ a Symmetric Pixel Shift Masking Denoise (Sec. 3.2) technique and a Self-Attention module (Sec. 3.3) to align the right image with the left one. Since our method exclusively manipulates the latent variable, it can be applied across various image generation tasks in different Stable Diffusion models. This versatility stems from the technique's focus on latent space operations, making it adaptable to a wide range of scenarios within the Stable Diffusion framework (Sec. 3.4). Our method only requires a disparity map, which can be obtained by various depth estimation models like DPT [26], MiDas [27] etc. and does not require camera calibration.

3.1. Stereo Pixels Shift

For the task of generating stereo images, fine-tuning models on large stereo datasets like KITTI [17] seems intuitive. However, after fine-tuning the model using various methods such as ControlNet [4] and LoRA [11], the results of the generated images remains unsatisfactory. A major flaw of this approach is that even if we could generate a high-quality stereo image pairs, the types of images generated will be limited to driving scenes similar to KITTI, losing the most important feature of Stable Diffusion: its diversity. Inspired by the Denoising Diffusion Implicit Models (DDIM) sampling technique for Stable Diffusion [32], we present a new method, Stereo Pixels Shift, without the aforementioned drawbacks. Utilizing DDIM for sampling from generalized generative processes, a latent vector sample $\boldsymbol{x}_{t-1}$ is generated from a sample $\boldsymbol{x}_t$ via a noise predictor ϵ_θ:

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\, \epsilon_\theta^{(t)}(\boldsymbol{x}_t)}{\sqrt{\alpha_t}} \right)}_{\text{predicted } \boldsymbol{x}_0} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2}\, \epsilon_\theta^{(t)}(\boldsymbol{x}_t)}_{\text{direction pointing to } \boldsymbol{x}_t} + \underbrace{\sigma_t \epsilon_t}_{\text{random noise}}, \tag{1}$$

where ϵ_t is noise following a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, independent of $\boldsymbol{x}_t$, and α_t controls the noise scale at step t with $\alpha_0 := 1$. If we set $\sigma_t = 0$ for all t and the same model ϵ_θ is used, the generative results are consistent and identical, making the forward process deterministic, given $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_0$. Thus, the result of $\boldsymbol{x}_{t-1}$ depends solely on $\boldsymbol{x}_t$. During the denoise process at a certain step t', if we modify $\boldsymbol{x}_{t'}$ to $\boldsymbol{x}'_{t'}$, subsequently, $\boldsymbol{x}'_{t'-1}$ is denoised based on $\boldsymbol{x}'_{t'}$, eventually generating $\boldsymbol{x}'_0$ which is different from the original $\boldsymbol{x}_0$. This pivotal insight enables the practical application of Stable Diffusion for stereo image generation. To align with this approach, we scale down the disparity map to match the dimensions of the latent space. Subsequently, we manipulate the latent vector on a pixel-by-pixel basis, guided by the disparity map. Given the relatively small size of the latent vector, this process does not entail a substantial computational overhead.

Assuming that the two images have parallel optical axes,

we derive a disparity map from a depth map using

$$D(x, y) = \frac{fB}{Z(x, y)}\,, \tag{2}$$

where (x, y) is a point in image space, Z is the depth map, f represents the focal length, and B is the baseline distance (i.e., the distance between the two cameras). Typically, we normalize the range of the disparity map $D(x, y)$ to be in $[0, 1]$. When the disparity map is generated by a model rather than being measured by actual devices, the conversion process is unnecessary, since many depth estimation models are capable of directly generating disparity maps.

The Stereo Pixel Shift operation $\mathcal{S}$ is expressed by

$$\mathbf{x}_{\text{right}}(x, y) = \mathbf{x}_{\text{left}}(x - s\,D(x, y), y)\,, \tag{3}$$

where $\mathbf{x}$ (left or right) denotes the latent variable $\boldsymbol{x}_t$, $\mathbf{x}_{\text{left}}(x - s\,D(x, y), y)$ represents the position in the latent space that is shifted left by $D(x, y)$ pixels relative to the position (x, y) in the latent space, and s is a scaling factor that controls the range of disparity, i.e., the pixel shift distance of the point closest to the observer in the right image relative to the left. Within reasonable limits, a larger value of s enhances the stereo effect of the generated images, usually restricted to within 10% of the image width. Excessively large s values can cause discomfort or blurriness rather than a sense of depth. However, using this method on images directly can lead to problems like flying pixels, as it causes individual pixels to warp into the empty spaces between two depth surfaces [39]. But since we operate on pixels in the latent space, individual pixel issues are typically resolved in the subsequent denoising and decoding processes. Thus, our method is straightforward, requiring no additional processing such as sharpening of the moved pixels.

The reason why we can apply Stereo Pixel Shift to the latent variable is that, after a certain step, there is a spatial position correspondence between the latent variable and the generated image. According to diffusion process theory [1, 20, 24, 32, 33, 44], sampling can be represented as

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{(1 - \bar{\alpha}_t)}\epsilon\,, \tag{4}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. Applying the Fourier transform on both sides, we have

$$\mathcal{F}(\boldsymbol{x}_t) = \sqrt{\bar{\alpha}_t}\,\mathcal{F}(\boldsymbol{x}_0) + \sqrt{(1 - \bar{\alpha}_t)}\,\mathcal{F}(\epsilon)\,. \tag{5}$$

If the step t is small, then $\bar{\alpha}_t \approx 1$, which indicates that early-stage sampling involves low-frequency signals primarily defining the contours of the generated image. When the step t is large, $\bar{\alpha}_t \approx 0$, high-frequency signals in the later-stage sampling refine the image details. This results in a significant disparity between the generated image and the original image if pixel offsets are applied too early during the sampling steps. Applying pixels shifts too late maintains high consistency in image content, but results in noticeable artifacts in the generated images. We found through experiments that setting t to 20% of the total denoise step usually works well. Additionally, the appropriate sampling steps for pixel offsets vary depending on the size of the objects in the images, see Figure 3.

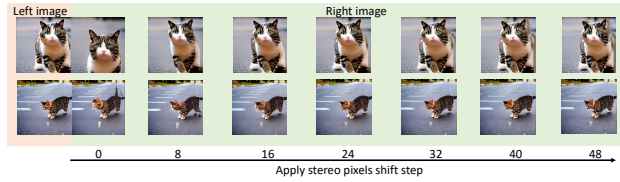

Figure 3. Comparing the outcomes of applying stereo shifts at different steps of denoising, reveals varying optimal configurations for different images. Implementing shifts too early could result in significant content alterations, while shifts applied too late might lead to noticeable artifacts in the images.

3.2. Symmetric Pixel Shift Masking Denoise

After applying the Stereo Pixel Shift, the right latent vector becomes inconsistent with the left one, potentially leading to discrepancies in the moved subject content following the denoising process.

According to Eq. 1, a pixel shift applied to $x_{t'}$ to obtain $x'_{t'}$ results in a slight difference between $x'_{t'-1}$ and $x_{t'-1}$. This difference accumulates during the subsequent denoising process, leading to variations in the final generated image. As a result, when the denoising algorithm is applied, it may interpret the shifted areas differently, potentially causing variations in how the subject matter appears after processing. This challenge is crucial in stereo image generation, as maintaining symmetry and coherence between the two sides is essential for creating a convincing and realistic stereo effect.

To circumvent the issue, inspired by the concept of inpainting, we propose a Symmetric Pixel Shift Masking Denoise method. We create a mask for the area where the stereo pixel shift is applied. At regular intervals, defined by specific steps t', the values from the masked region of the left latent space are copied to the corresponding area of the mask in the right latent space. Consequently, the denoising process for the right image can be reformulated from Eq. 1 as

$$\begin{aligned}\boldsymbol{x}'_{t'-1} = \sqrt{\alpha_{t'-1}}&\left(\frac{\boldsymbol{X}'_{t'} - \sqrt{1-\alpha_{t'}}\,\epsilon_\theta^{(t')}(\boldsymbol{x}'_{t'})}{\sqrt{\alpha_{t'}}}\right)\\ &+ \sqrt{1-\alpha_{t'-1}}\,\epsilon_\theta^{(t')}(\boldsymbol{x}'_{t'})\,,\end{aligned} \tag{6}$$

where $\boldsymbol{x}'_{t'}$ represents the right latent vector after undergoing a pixel shift, and the ith element of $\boldsymbol{X}'_{t'}$ is expressed by

$$\boldsymbol{X}'_{t',i} = \begin{cases} \mathcal{S}(\boldsymbol{x}_{t'-1,i}, D) & \text{if } \mathbf{M}_i = \text{True}, \\ \boldsymbol{x}'_{t',i} & \text{otherwise}, \end{cases} \tag{7}$$

where $\mathcal{S}$ represents the operation of Stereo Pixel Shift in Eq. 3, D is the corresponding disparity map of the image, and $\mathbf{M}$ is a Boolean matrix of the same shape as $\boldsymbol{x}$ that signifies the mask, with values set to True for the pixels that have been shifted. The variable $\boldsymbol{x}_{t'-1}$ denotes the latent vector of the left image at time step $t'-1$.

We note that if the area shifted is left blank (i.e., filled with zeros), the denoised region might become blurry. We address this blurriness by filling the shifted blank area with random noise using

$$\boldsymbol{x}_{t',i}^{(\text{deblur})} = \begin{cases} \boldsymbol{x}_{t',i} & \text{if } \mathbf{M}_i = \text{False}, \\ \epsilon_{t',i} & \text{otherwise}, \end{cases} \tag{8}$$

where $\epsilon_{t'}$ denotes random noise, $\mathbf{M}$ is the same mask as the one in Eq. 7. However, the effectiveness varies with different images. Sometimes, it may even lead to a decrease in the quality of the generated images. A detailed effects analysis of the Deblur technique is presented in the ablation studies described in Sec. 4.3.

3.3. Self-Attention layers modification

As numerous studies have attempted to modify the attention mechanisms within Stable Diffusion to achieve the goal of modifying the original images [3, 10, 22, 36, 42], we tackle this challenge by utilizing both Unidirectional and Bidirectional Self-Attention mechanisms. This method eliminates the need for fine-tuning the model to adjust its weights. Refer to supplementary materials for detailed explanation.

Algorithm 1 Bi/Uni-directional Attention Modification

Require: A text condition $\mathcal{C}$, a left latent variable z_{t-1} and a right latent variable z'_{t-1}.
Ensure: An edited right latent variable z'^{*}_{t-1} and an edited latent latent variable z^{*}_{t-1} if bidirection.

```
(z_{t-1}, z'_{t-1}), (M_t, M'_t) ← ε_θ((z_t, z'_t), t, C);
M̂_t, M̂'_t ← Edit (M_t, M'_t, t) ;
if Unidirection then
    (z_{t-1}, z'*_{t-1}) ← ε_θ((z_t, z'_t), t, C){M' ← M̂'_t}
    return (z_{t-1}, z'*_{t-1})
else if Bidirection then
    (z*_{t-1}, z'*_{t-1}) ← ε_θ((z_t, z'_t), t, C){M ← M̂_t, M' ← M̂'_t}
    return (z*_{t-1}, z'*_{t-1})
end if
```

Our modified approach is listed in Algorithm 1. The term $\epsilon_\theta((z_t, z'_t), t, \mathcal{C})$ represents the computation of a single step t of the diffusion process, which yields the noisy image z_{t-1} and the attention map M_t. Here, (z_t, z'_t) denote the left and right latent variables, respectively. In our implementation, these latent variables are stacked together along the batch size dimension. However, they are presented separately here for ease of explanation. The expression $\epsilon_\theta((z_t, z'_t), t, \mathcal{C})\{M' \leftarrow \widehat{M'_t}\}$ denotes the diffusion step where the attention map M is superseded by an additional map $\widehat{M}$. We define the function $\text{Edit}\,(M_t, M'_t, t)$ as a general edit function, designed to process the t-th attention maps of the left and right latent variables.

3.4. Application scenarios

As shown in Figure 1, our method is compatible with various types of Stable Diffusion models, enabling it to: (a) produce the corresponding right-side image from an existing left-side image; (b) generate stereo images from text prompts; (c) produce the corresponding right-side image from an existing left-side image, where the pair shares the same composition but differs in content. For text-to-image and depth-to-image tasks, the initial noise is randomly generated. Thus, it is sufficient to apply a pixel shift to the denoised noise after a specific denoising step, as illustrated in Figure 2. However, for generating stereo image pairs of an existing image, it is necessary to use null-text inversion [19] to obtain the latent space of the original image. A straightforward inversion technique was proposed for DDIM sampling [5, 32]. This technique is grounded in the hypothesis that the ordinary differential equation (ODE) process is reversible, especially in scenarios involving small step sizes. The diffusion process is executed in reverse, meaning the transition is from z_0 to z_T, contrary to the typical z_T to z_0 progression:

$$z_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}\, z_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \varepsilon_\theta(z_t, t, \mathcal{C})\,. \tag{9}$$

Here, ε_θ is a noise predictor including an embedding of a text condition $\mathcal{C}$, while z_0 is the encoding of the provided real image. A guidance scale parameter w is used to blend between a noise predictor with no text condition ($w = 0$) and ε_θ with $\mathcal{C}$.

To address the inefficiency of mapping each noise vector to a single image, we start with a default DDIM inversion at $w = 1$ as the pivot trajectory. Subsequently, we optimize around this trajectory using a standard guidance ratio of $w > 1$. In practical applications, individual optimizations are conducted for each step t during the diffusion process, aiming to closely approximate the initial trajectory z^*:

$$\min \left\| z^*_{t-1} - z_{t-1} \right\|_2^2 , \tag{10}$$

where z_{t-1} represents the intermediate result of the optimization. We thus substitute the default blank text embedding with an optimized embedding. This is because the generated results are significantly influenced by the unconditional prediction [19].

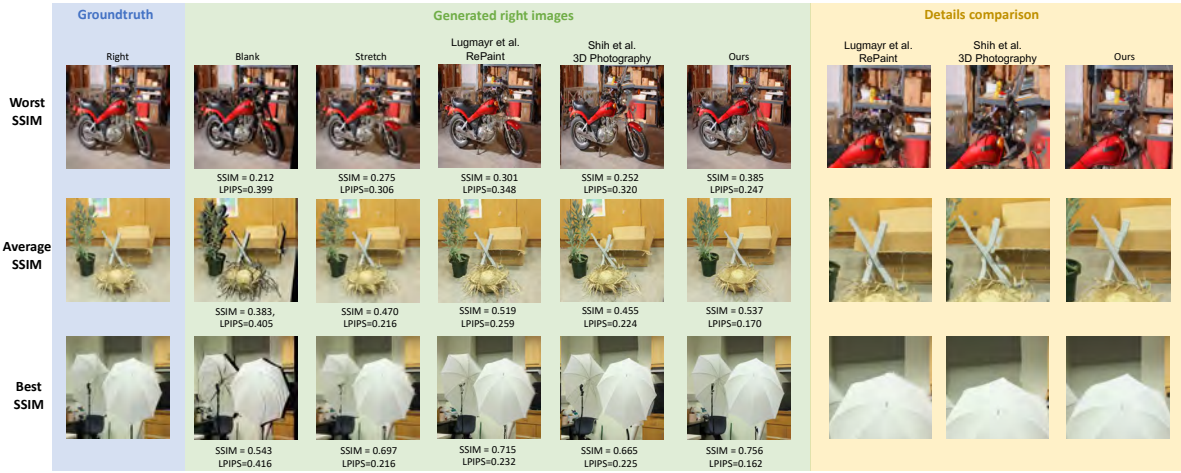

Figure 4. The same image generated using different methods. The rows present, respectively, the image with the lowest (worst) SSIM score, the image closest to the average SSIM score, and the image with the highest (best) SSIM score generated using our method. The other methods are represented solely by their results on these specific images, which do not necessarily reflect the best, average, or worst SSIM scores achievable by those methods. We do this to facilitate a direct comparison of the effects of each method on the same image. We also provide LPIPS scores for reference and close-ups of the images generated by the primary benchmark methods for inspection of details.

Table 1. Quantitative evaluation results for Middlebury and KITTI: the results of generating right-side images from left-side images and disparity maps using different methods. We assess the similarity between the generated and the original images using PSNR, SSIM, and LPIPS. 'GT' indicates the use of ground truth disparity maps, while 'pseudo' denotes the use of disparity maps generated by a depth estimation model. Scores presented in bold indicate the best performance. The numbers in the top right represent the best scores, while those in the bottom right indicate the worst scores.

Methods	Middlebury			KITTI		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Leave blank	$11.328^{+2.483}_{-3.489}$	$0.315^{+0.230}_{-0.154}$	$0.450^{-0.089}_{+0.097}$	$12.980^{+4.919}_{-3.831}$	$0.374^{+0.286}_{-0.251}$	$0.313^{-0.109}_{+0.222}$
Stretch	$14.842^{+2.714}_{-2.753}$	$0.432^{+0.265}_{-0.190}$	$0.285^{-0.089}_{+0.112}$	$14.757^{+5.375}_{-4.694}$	$0.429^{+0.287}_{-0.271}$	$0.212^{-0.100}_{+0.145}$
3D Photography [31]	$14.190^{+2.464}_{-2.798}$	$0.427^{+0.238}_{-0.175}$	$0.275^{-0.065}_{+0.073}$	$14.540^{+7.256}_{-4.023}$	$0.398^{+0.270}_{-0.323}$	$0.210^{-0.073}_{+0.099}$
RePaint [16]	$15.102^{+2.909}_{-2.802}$	$0.462^{+0.253}_{-0.184}$	$0.311^{-0.079}_{+0.090}$	$15.056^{+5.366}_{-4.897}$	$0.462^{+0.268}_{-0.285}$	$0.251^{-0.095}_{+0.128}$
Ours (with GT disparity)	$15.456^{+2.669}_{-3.313}$	$0.468^{+0.252}_{-0.205}$	$0.231^{-0.088}_{+0.096}$	$\mathbf{15.679}^{+5.888}_{-5.487}$	$\mathbf{0.481}^{+0.245}_{-0.310}$	$\mathbf{0.205}^{-0.099}_{+0.135}$
Ours (with pseudo disparity)	$\mathbf{16.980}^{+4.737}_{-3.818}$	$\mathbf{0.551}^{+0.208}_{-0.166}$	$\mathbf{0.173}^{-0.069}_{+0.074}$	$15.589^{+8.061}_{-5.016}$	$0.479^{+0.241}_{-0.300}$	$0.209^{-0.116}_{+0.114}$

4. Experiments

We have compared our results with traditional methods such as 'leave blank' and 'stretch'. Additionally, we have selected the 3D Photography techniques of Shih et al. [31] for comparison, as well as the RePaint method of Lugmayr et al. [16], which involves using Stable Diffusion for inpainting images processed by the traditional 'leave blank' method. It is important to emphasize that RePaint is not inherently designed for generating stereo image pairs. However, we believe that employing inpainting techniques to fill in the blank areas after creating stereo images is a very straightforward and common approach. Thus, we have chosen to compare with the latest model that achieves good results in various metrics within the same Stable Diffusion framework. This comparison is intended to demonstrate the innovation and advantage of our method.

4.1. Quantitative evaluation

Since currently no metrics exists specifically for the comparison of stereo image pair generation, we quantitatively evaluate our results using the Middlebury [30] and KITTI [17] datasets. We evaluate the performance by generating the right-side image from the left-side image and its disparity map. We then compare the model-generated right-side image with the ground truth image. We calculated the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) between the generated image and the ground truth. The results are in Table 1 with some visuals in Figure 4. We provide the settings used for each method, more comparison figures, and a detailed explanation of the quantitative evaluation results in the supplemental document.

In Table 3, we compare the time consumption of different methods for generating a single stereo image pair using an

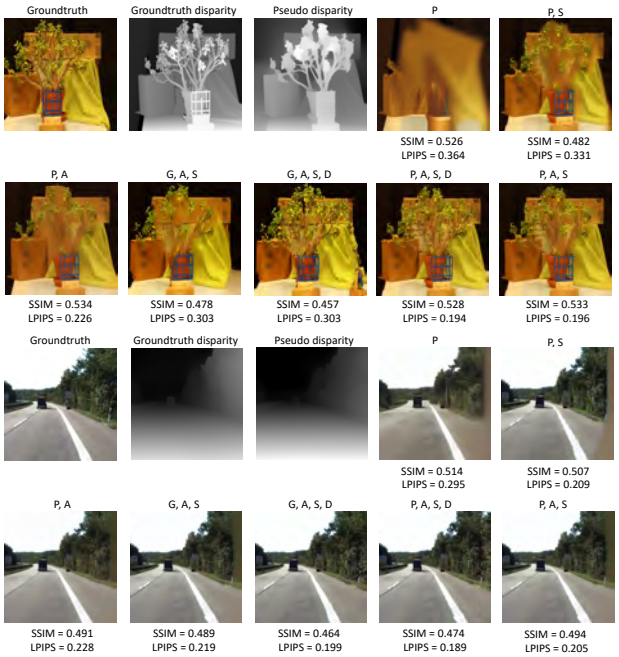

Figure 5. Ablation example of Middlebury (up) and KITTI (down). In the images, 'P' and 'G' respectively denote whether the image has been guided by a Pseudo disparity map or a Groundtruth disparity map. 'A', 'S', and 'D' indicate the use of Attention layers modification, Symmetric Pixel Shift Masking Denoise, and Deblur technique, respectively. The lower scores associated with the use of Groundtruth disparity maps in Middlebury may be attributed to their generally higher precision and complexity. This heightened detail can render pixel shift operations during image generation more intricate and sensitive. Our Stereo Pixel Shift operation is executed within a smaller latent space (64×64), where minor pixels, such as those around tree trunks and leaves, might be overlooked. In contrast, disparity maps generated by depth estimation models, with their lower precision, are more conducive to Pixel Shift in the latent space without sacrificing image detail.

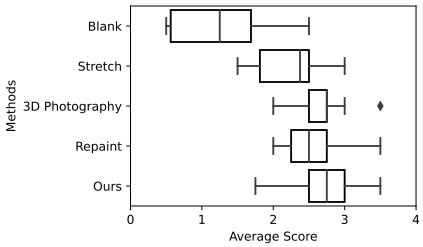

Figure 6. User evaluation results. Distribution of the scores provided by users on the test scenes.

NVIDIA RTX3090 graphics card. Our method offers the capability to quickly generate high-quality stereo image pairs with direct integration into Stable Diffusion.

4.2. User evaluations

In our user tests, we adopted a more practical and user-centric approach. The users' input were text prompts used to generate stereo image pairs using Stable Diffusion. For benchmarking, we compared this with other methods by generating the left-side images using Stable Diffusion, obtaining the corresponding disparity maps via a depth estimation model, and then using the respective methods to generate stereo image pairs. We utilized Google Cardboard and presented the stereo images on mobile phones, inviting participants to assess the image quality and correctness of the 3D perception. Ratings ranged from 0 to 5, with 5 being the highest and 0 being the lowest. Some test pictures are shown in Figure 12 of the supplemental document.

The results of the user tests showed that our method has the highest average but it did not significantly outperform the others. This was anticipated, as when viewing stereo images, people tend to focus more on the overall image rather than the details. In terms of ease of use, our proposed method has a clear advantage. It is simpler, does not require an additional inpainting model, and can be seamlessly integrated with Stable Diffusion.

4.3. Ablation study

We conducted ablation studies on the proposed method to evaluate the impact of images guided by either Groundtruth disparity maps or Pseudo disparity maps (generated by a depth estimation model), as well as the effects of using Symmetric Pixel Shift Masking Denoise, Attention Layer Modification, and Deblur techniques on PSNR, SSIM, and LPIPS scores. The results are shown in Table 2. Figure 5 presents a visual representation of an example from the Middlebury dataset and KITTI to intuitively demonstrate the impact of each factor on the image generation outcomes, explaining the reason that scores using Groundtruth disparity maps in the Middlebury dataset are unexpectedly lower than those using Pseudo disparity maps.

5. Limitations and Discussion

Our method relies on the disparity map. If the results generated by other depth estimation models are inaccurate, our method will also be unable to produce high-quality stereo images. Furthermore, when using high-precision disparity maps obtained from device measurements, the results may not be entirely satisfactory, as shown in Table 2 and Figure 5.

When using our depth to stereo image model, one may observe overlapping areas in the generated images. This issue might stem from the *LatentDepth2ImageDiffusion* model we used, which tends to fill blank areas with pixels from adjacent main subjects rather than background elements. In such cases, a better-quality image can be generated by first generating a single image using the *Depth2Image* model, and then applying our Image to Stereo Image Pairs method, as illustrated in Figure 7.

We found that our method can be used for inpainting tasks with the original text prompt to a Stable Diffusion image

Table 2. Ablation study on Middlebury and KITTI. In the 'Disparity Map' column, 'GT' and 'Pseudo' respectively indicate the use of groundtruth disparity maps or disparity maps generated by a Depth Estimation Model. In the 'Technique Applied' column, 'Attn Layer,' 'SPSMD,' and 'Deblur' represent the use of Self-Attention Layers Modification, Symmetric Pixel Shift Masking Denoise, and Deblur techniques, respectively. The symbol '✓' denotes the adoption of these respective techniques. Bold numbers represent the best scores for that column. When employing Attn Layer and SPSMD together, LPIPS has a better score, but the effect of Deblur varies from image to image. When the LPIPS scores are comparable, the higher SSIM score indicates the better similarity. An example is shown in Figure 5.

Disparity map		Technique Applied			Middlebury			KITTI		
GT	Pseudo	Attn Layers	SPSMD	Deblur	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
	✓				$16.352^{+2.330}_{-2.139}$	$0.514^{+0.241}_{-0.169}$	$0.378^{-0.149}_{+0.181}$	$15.286^{+5.389}_{-4.622}$	$0.476^{+0.229}_{-0.346}$	$0.364^{-0.206}_{+0.265}$
	✓		✓		$\mathbf{17.076}^{+4.450}_{-3.652}$	$0.549^{+0.190}_{-0.174}$	$0.191^{-0.073}_{+0.082}$	$15.305^{+6.013}_{-4.772}$	$0.474^{+0.228}_{-0.303}$	$0.230^{-0.115}_{+0.114}$
	✓	✓			$15.421^{+2.849}_{-3.657}$	$0.478^{+0.245}_{-0.199}$	$0.255^{-0.123}_{+0.116}$	$\mathbf{15.868}^{+7.775}_{-5.068}$	$\mathbf{0.483}^{+0.240}_{-0.301}$	$0.212^{-0.113}_{+0.114}$
✓		✓	✓		$15.456^{+2.669}_{-3.313}$	$0.468^{+0.252}_{-0.205}$	$0.231^{-0.088}_{+0.096}$	$15.679^{+5.888}_{-5.486}$	$0.481^{+0.245}_{-0.310}$	$0.205^{-0.099}_{+0.135}$
✓		✓	✓	✓	$15.149^{+2.769}_{-3.174}$	$0.444^{+0.263}_{-0.231}$	$0.234^{-0.097}_{+0.117}$	$15.360^{+5.787}_{-5.332}$	$0.461^{+0.250}_{-0.306}$	$\mathbf{0.200}^{-0.092}_{+0.127}$
	✓	✓	✓	✓	$16.753^{+4.815}_{-3.783}$	$0.540^{+0.197}_{-0.169}$	$0.174^{-0.071}_{+0.066}$	$15.269^{+7.923}_{-5.053}$	$0.458^{+0.261}_{-0.295}$	$0.204^{-0.111}_{+0.106}$
	✓	✓	✓		$16.980^{+4.737}_{-3.818}$	$\mathbf{0.551}^{+0.208}_{-0.166}$	$\mathbf{0.173}^{-0.069}_{+0.074}$	$15.589^{+8.061}_{-5.016}$	$0.479^{+0.241}_{-0.300}$	$0.209^{-0.116}_{+0.114}$

Table 3. Time cost for different methods in seconds. We measure the total time consumed for each usage scenario, including the time taken to generate the images using Stable Diffusion. The scenarios are text to stereo image (T2SI), depth to stereo image (D2SI), and image to stereo image (I2SI). The time in parenthesis is the cost excluding the time spent on generating images with Stable Diffusion. For D2SI, our method, being directly integrated into Stable Diffusion, requires only a single pass of sampling to generate stereo image pairs. In I2SI, our method requires the use of null-text inversion [19] for x_t, resulting in an extra 23 seconds.

Methods	T2SI	D2SI	I2SI
3D Photography [31]	245 (231)	247 (231)	231
Repaint[16]	338 (324)	340 (324)	324
Ours	32 (18)	18	40 (17)

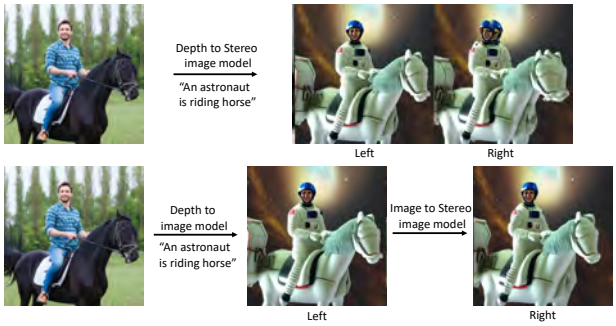

Figure 7. Limitation: Generating compositionally similar stereo images directly from a disparity map may sometimes fail. However, this issue can be mitigated by first generating a compositionally similar left image using the disparity map, and then employing the Image to Stereo Image method to generate the right image. This two-step process helps avoid such failures.

model. We conducted a simple test where, after obtaining $\boldsymbol{x}_t$ using null-text inversion, we applied various masking ratios to the right side and tested whether Stable Diffusion could

Figure 8. Results for inpainting tasks using our proposed method, the red-colored areas represent the masked regions.

fill in the blank areas within the mask during denoising. The results, as shown in Figure 8, indicate that our method is somewhat effective for inpainting when a smaller area of the image is masked. However, when a larger portion of the image is masked, the inpainting results exhibit a strong patchwork appearance. Applying our method to inpainting tasks might require further modifications to both the model and the technique.

6. Conclusion

We proposed a novel method for generating stereo image pairs by modifying the latent vector of Latent Stable Diffusion. We implement Stereo Pixel Shift on the left latent vector and its corresponding disparity map, and during the denoising process, we ensure consistency between the left and right images through Symmetric Pixel Shift Masking Denoise and Self-Attention Layer Modification. Our approach differs fundamentally from traditional inpainting pipelines and can be seamlessly integrated into existing Stable Diffusion models, offering end-to-end capabilities for text prompt to stereo image, depth to stereo image, and image to stereo image generation, all without the need for fine-tuning any parameters and using only the original Stable Diffusion model. Our method achieved better scores on both the KITTI and Middlebury datasets.

References

[1] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *International Conference on Learning Representations (ICLR)*, 2022. 4

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: high-resolution video synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575. IEEE, 2023. 2, 3

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: tuning-free mutual self-attention control for consistent image aynthesis and editing. In *International Conference on Computer Vision (ICCV)*, pages 22503–22513. IEEE, 2023. 2, 5, 3

[4] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: controllable text-to-video generation with diffusion models. arXiv:2305.13840 [cs.CV], 2023. 3, 2

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8780–8794, 2021. 5

[6] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. CogView2: Faster and better text-to-image generation via hierarchical transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 16890–16902, 2022. 1

[7] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. AnimateDiff: animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[8] Peter Hedman, Suhib Alsisan, Richard Szeliski, and Johannes Kopf. Casual 3D photography. *ACM Transactions on Graphics*, 36(6):234:1–234:15, 2017. 2

[9] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics*, 37(6):257:1–257:15, 2018. 2

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations (ICLR)*, 2023. 1, 2, 5

[11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 3

[12] 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 *International Conference on Computer Vision (ICCV)*, pages 12528–12537. IEEE, 2021. 2

[13] Johannes Kopf, Fabian Langguth, Daniel Scharstein, Richard Szeliski, and Michael Goesele. Image-based rendering in the gradient domain. *ACM Transactions on Graphics*, 32(6): 199:1–199:9, 2013. 2

[14] Yuhan Li, Yishun Dou, Xuanhong Chen, Bingbing Ni, Yilin Sun, Yutian Liu, and Fuzhen Wang. Generalized deep 3D shape prior via part-discretized diffusion process. In *Computer Vision and Pattern Recognition (CVPR)*, pages 16784–16794. IEEE, 2023. 1

[15] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: any single image to 3D mesh in 45 seconds without per-shape optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 22226–22246, 2023. 1

[16] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: inpainting using denoising diffusion probabilistic models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471. IEEE, 2022. 6, 8, 1, 5

[17] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070. IEEE, 2015. 3, 6

[18] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics*, 38(4):29:1–29:14, 2019. 2

[19] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6038–6047. IEEE, 2023. 5, 8, 1

[20] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of Machine Learning Research (PMLR)*, pages 8162–8171, 2021. 4

[21] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of Machine Learning Research (PMLR)*, pages 16784–16804, 2022. 1

[22] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *SIGGRAPH 2023 Conference Proceedings*, pages 11:1–11:11. ACM, 2023. 2, 5

[23] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: text-to-3D using 2D diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[24] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: toward a meaningful and decodable representation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10619–10629. IEEE, 2022. 4, 2

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. arXiv:2204.06125 [cs.CV], 2022. 1

[26] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *International Confer-*

ence on Computer Vision (ICCV), pages 12179–12188, 2021. 3, 1

[27] 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 Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 3, 1

[28] Chris Rockwell, David F. Fouhey, and Justin Johnson. PixelSynth: generating a 3D-consistent experience from a single image. In *International Conference on Computer Vision (ICCV)*, pages 14104–14113. IEEE, 2021. 1

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 36479–36494, 2022. 1

[30] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition (GCPR)*, pages 31–42. Springer, 2014. 6

[31] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3D photography using context-aware layered depth inpainting. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8028–8038. IEEE, 2020. 2, 6, 8, 1

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 3, 4, 5

[33] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 4

[34] Pratul P. Srinivasan, Richard Tucker, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 175–184. IEEE, 2019. 2

[35] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930. IEEE, 2023. 2

[36] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. arXiv:2303.17599 [cs.CV], 2023. 2, 5

[37] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. VideoComposer: compositional video synthesis with motion controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7594–7611, 2023. 2

[38] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: high-fidelity and diverse text-to-3D generation with variational score distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8406–8441, 2023. 2

[39] Jamie Watson, Oisin Mac Aodha, Daniyar Turmukhambetov, Gabriel J. Brostow, and Michael Firman. Learning stereo from single images. In *European Conference on Computer Vision (ECCV)*, pages 722–740. Springer, 2020. 4

[40] Thomas Whelan, Michael Goesele, Steven J. Lovegrove, Julian Straub, Simon Green, Richard Szeliski, Steven Butterfield, Shobhit Verma, and Richard A. Newcombe. Reconstructing scenes with mirror and glass surfaces. *ACM Transactions on Graphics*, 37(4):102:1–102:11, 2018. 2

[41] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3D reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9065–9075. IEEE, 2023. 1

[42] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: one-shot tuning of image diffusion models for text-to-video generation. In *International Conference on Computer Vision (ICCV)*, pages 7623–7633. IEEE, 2023. 2, 5

[43] Yue Wu, Sicheng Xu, Jianfeng Xiang, Fangyun Wei, Qifeng Chen, Jiaolong Yang, and Xin Tong. AniPortraitGAN: animatable 3D portrait generation from 2D image collections. In *SIGGRAPH Asia 2023 Conference Proceedings*, pages 51:1–51:9. ACM, 2023. 1

[44] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: a comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):105:1–105:39, 2023. 2, 4

[45] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, (11), 2022. 1

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision (ICCV)*, pages 3836–3847. IEEE, 2023. 2

[47] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4):65:1–65:12, 2018. 2