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

Supplementary Material

7. Quantitative evaluation experiments setting

In this section, we will provide a detailed description of the settings for each method. 3D Photography does not provide a direct method for generating stereo image pairs, its output is a mesh, which requires rendering to obtain images. Therefore, we manually set the left and right camera matrices as follows:

$$M_{\text{left}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad M_{\text{right}} = \begin{bmatrix} 1 & 0 & 0 & -0.04 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

After rendering with these settings, we obtained the left and right images of the stereo image pairs.

Given that Stable Diffusion can only generate images of 512×512 resolution, and the Middlebury dataset images are about 5 million pixels, we scaled both the dataset images and the corresponding depth maps to 512×512. For the Middlebury dataset, whose groundtruth disparity maps are noisy, we applied a Gaussian blur with a radius of 3 to smooth the disparity maps. Regarding the KITTI dataset, where the image size is 375×1242 with an aspect ratio of approximately 3.3, directly scaling images to 512×512 could lead to excessive stretching, negatively impacting many models' performance. Therefore, we proportionally scaled the images to 512×1696 and then applied a center crop to 512×512 to avoid excessive stretching. As null-text inversion technique is required, we used the Stable Diffusion version 1.5 for this test, setting the denoising steps to 50.

For the 3D photography method [31], we used the disparity map generated by the integrated MiDaS model [27] within its framework instead of the groundtruth disparity map. This was due to the extensive time required—up to two hours—for mesh reconstruction of a single image using the groundtruth disparity map with 3D photography. We hypothesize that this inefficiency arises when 3D photography attempts to reconstruct stereo image pairs from the disparity map, necessitating operations like breaking up discontinuous vertices in the mesh. Such processes become computationally intensive when the groundtruth disparity map is excessively noisy, leading to a proliferation of isolated vertices that consume substantial CPU resources. For the purpose of benchmarking and considering the rarity of obtaining groundtruth disparity maps in practical scenarios, we evaluated the results using both groundtruth disparity maps (denoted as GT disparity) and pseudo disparity maps generated by depth estimation models (denoted as Pseudo disparity). The depth estimation model we employed was DPT [26]. Since the use of Deblur results in lower scores, neither method employed deblur; details can be found in Section 4.3 Ablation study. When creating stereo image pairs using RePaint [16], we generate a mask for the blank areas left after moving the left-side image and then perform inpainting on the masked areas. The model *inet256* we utilized for this purpose was trained on ImageNet. Since RePaint's maximum supported output image size is 256×256, we downsized the images to 256×256 before conducting inpainting. However, considering that all other methods are evaluated at a 512×512 resolution, for fairness, we only upscale the inpainted area within the mask from 256×256 to 512×512, while maintaining the original resolution for the area outside the mask.

8. Analysis of Quantitative evaluation results

The use of null-text inversion [19] technique inherently causes distortion in images. On the Middlebury dataset, reference scores (for images generated by Stable Diffusion to be the same as the input) are: PSNR = 27.967, SSIM = 0.847, LPIPS = 0.046. The reference scores for the KITTI dataset are: PSNR = 25.615, SSIM = 0.762, LPIPS = 0.072. These scores represent the best possible outcomes achievable with the method we proposed. The quantitative analysis results, as seen in Table 1, indicate that our proposed method achieves state-of-the-art scores on both the datasets. Furthermore, as illustrated in Fig. 9, we selected images representing the best LPIPS, those closest to the average LPIPS, and the worst LPIPS from each method. This selection was made to visually demonstrate the differences in images generated by each method. Fig. 4 showcases images with the lowest SSIM, closest to the average SSIM, and the highest SSIM scores when using our method, compared to the outcomes when other methods are applied to the same images. We have also magnified some details to facilitate an intuitive comparison of the primary methods.

We also noted that the scores for the KITTI dataset are lower compared to those of the Middlebury dataset. However, if we convert the best scores into percentages relative to the Stable Diffusion reference scores, the results are as follows. For the Middlebury dataset, when SSIM = 0.551, it is 65.1% of the best score of 0.847, and for LPIPS = 0.173, the reference score of 0.046 constitutes 26.6% of the best score of 0.173 (the higher the percentage, the better). Similarly, for the KITTI dataset, SSIM is 63.1% of the reference score of 0.762, and the reference score for LPIPS of 0.072 is 35.1%

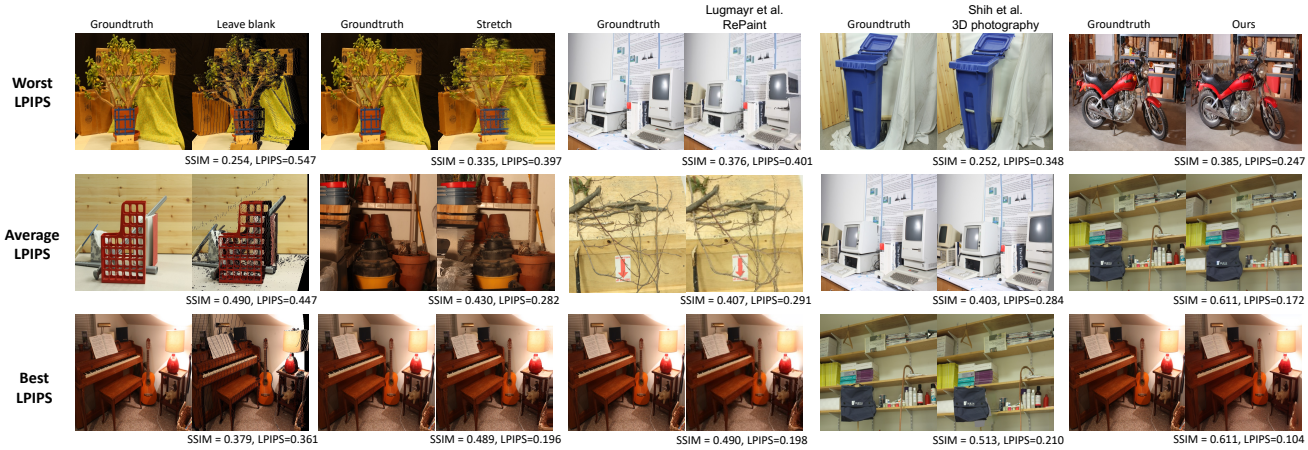


Figure 9. Comparing different methods by Perceptual Image Patch Similarity (LPIPS) scores. We evaluate the right-side images generated from left-side images and disparity maps using various methods: 'Worst LPIPS', 'Average LPIPS', and 'Best LPIPS'. These represent, respectively, the images with the highest (worst) LPIPS score, the image closest to the average LPIPS score, and the image with the lowest (best) LPIPS score for each method. We also annotate each image with its Structural Similarity Index Measure (SSIM) for reference.

of the best score. The model actually performs better on the KITTI dataset in terms of LPIPS. Another possible reason for this is the larger baseline distance B of the cameras used to capture the KITTI dataset images, which in turn requires a larger scale factor s (KITTI $s = 20$, Middlebury $s = 9$). This larger scale factor means that, when generating stereo image pairs, the corresponding pixels in the KITTI dataset images have to move a greater distance, resulting in more extensive blank areas.

9. Analysis of Ablation

Deblur has a certain negative impact on LPIPS and SSIM scores on Middlebury dataset, with a more pronounced effect on SSIM. This is because blurred images contain fewer high-frequency details, implying less noise and finer details. Since SSIM focuses more on large-scale structural features at lower frequencies, these features might appear more pronounced and consistent in blurred images, leading to higher SSIM scores. Unlike traditional metrics like SSIM or PSNR, LPIPS emphasizes perceptual differences rather than just pixel-level discrepancies, hence the lesser impact of Deblur on LPIPS scores. A lower LPIPS score with highter SSIM scores indicates closer approximation to the original image.

On the KITTI dataset, the scores for Groundtruth and Pseudo disparity maps are more aligned with general expectations. Compared to the high-precision and complex Groundtruth disparity maps in the Middlebury dataset, the Groundtruth disparity maps in the KITTI dataset are relatively straightforward, mostly depicting driving scenes. Therefore, stereo images guided by Groundtruth disparity maps scored higher than those guided by Pseudo disparity maps. We believe that the positive effect of Deblur in the KITTI data set is due to the large scale factor s, which makes the larger blank area left after Pixel shift unable to be filled during denoise. It's also important to note that the LPIPS score is a better indicator of the overall similarity of images. Therefore, a higher SSIM score accompanied by a higher LPIPS score does not necessarily imply a greater similarity to the original image, as demonstrated in Fig. 5. However, when the LPIPS scores are comparable, the SSIM score becomes a more effective measure for assessing the similarity of images.

10. Attention module modification details

Within the Stable Diffusion model, the denoising U-Net is structured as a series of basic blocks. Each basic block incorporates a residual block, a self-attention module, and a cross-attention module which can be represented as [24, 32, 44].

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V\,, \qquad (11)$$

where Q represents the query, while K and V represent the key and value, respectively, and d is the output dimension of the key and query features. The values are obtained through linear projection. When there is an input context, it functions as cross-attention. In the absence of context, it operates as self-attention. Cross-attention is commonly employed in tasks involving text-guided image editing [3, 10].

In the case of self-attention, non-rigid editing cannot be performed as the semantic layout and structures are maintained. Similar to sharing semantic information between different samples in the same batch using 3D convolution to align content across batches in video generation tasks [2, 7, 37], applying self-attention between samples within the same batch has a comparable effect [4, 42].

Querying the left-side image using the key and value of the right-side image in a unidirectional manner, enhancing the alignment from right image to left, is termed unidirectional self-attention. In contrast, employing queries from both the left and right sides to mutually query each other is referred to as bidirectional self-attention. However, bidirectional self-attention has a significant drawback: it aligns the left and right images with each other, thereby altering the input left-side image. Although this can enhance alignment, it is not a suitable option when users wish to keep the input image unchanged. Thus, despite its potential to improve alignment, the bidirectional approach may not be preferable if it is crucial to maintain the integrity of the input image. The algorithm is explained in appendix.

We apply this attention control to all layers of the U-Net to achieve the best alignment results. Although another study observed that applying attention control to all layers results in exactly the same images [3], in our method, stereo shifts have already been applied, which leads to content consistency while the main subject is shifted to different positions, precisely the outcome we desire.

11. Attempts of fine-tuning Stable Diffusion model to genreate stereo image pairs

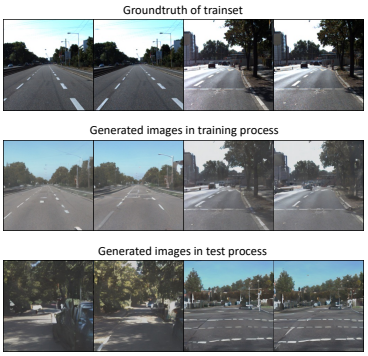


Figure 10. Example of images generated by stereo fine-tuned Stable Diffusion: The images reveals that while the generated left and right images exhibit certain similarities, the extent of this resemblance falls significantly short of the requirements for stereo imaging. Even during training, maintaining pixel-level consistency between the left and right images proves challenging, and the quality of images generated during test exhibits notable deficiencies.

In this section, we briefly present our initial attempts at fine-tuning Stable Diffusion for generating stereo image pairs. This approach was unsuccessful in producing high-quality stereo image pairs.

ControlNet [4], known for its capability to manipulate the posture of images generated by Stable Diffusion, produces images that are structurally similar to the input image but with different content. We hypothesized that this might be beneficial for generating stereo images. Consequently, we adopted an architecture similar with ControlNet. A neural network block $F(\cdot;\Theta)$ with a set of parameters Θ transforms a feature map $\boldsymbol{x}$ into another feature map $\boldsymbol{y}$.

$$\boldsymbol{y} = F(\boldsymbol{x};\Theta)\,. \tag{12}$$

We have frozen all the parameters Θ of the original Stable Diffusion and created a trainable copy Θ_c. The neural network blocks are interconnected through a distinctive convolution layer, which is initialized with zero weights and biases. The operation can be represented by the following equation

$$\boldsymbol{y}_{\mathrm{c}} = \mathcal{F}(\boldsymbol{x};\Theta) + \mathcal{Z}\left(\mathcal{F}\left(\boldsymbol{x} + \mathcal{Z}\left(\boldsymbol{c};\Theta_{\mathrm{z1}}\right);\Theta_{\mathrm{c}}\right);\Theta_{\mathrm{z2}}\right), \tag{13}$$

where y_c represents the output of this neural network block. The operation $Z(\cdot;\cdot)$ denotes a zero convolution operation, and $\{\Theta_{z1}, \Theta_{z2}\}$ represents two instances of parameters, each corresponding to a distinct instance of the zero convolution operation.

Using ControlNet only maintains the general content of the images, which is insufficient for generating stereo image pairs. We aim for Stable Diffusion to generate stereo image pairs concurrently. To achieve this, we align even-numbered images in the batch with their adjacent odd-numbered counterparts, such as 0 with 1, and 1 with 2, to create a stereo effect between each adjacent pair. Inspired by VideoLDM[2], we introduce a 3D convolution layer and a temporal attention layer into the Stable Diffusion architecture. These layers are added after Stable Diffusion's existing spatial layers in the U-Net. The function of 3D convolution layer's is to break the information isolation between different samples in the same batch. Before feeding the intermediate features to the 3D convolution layer, we reshape the features from `[b c h w]` to `[b/2 2 c h w]`, where `b`, `c`, `h`, `w` represent batch size, color channel, height, and width, respectively. The `2` in the reshaped second item represents the left and right images, allowing the newly added 3D convolution block to learn the distribution of the left and right stereo image pairs. The structure of the temporal attention layer is same as that in Stable Diffusion, assisting the 3D convolution layer in distinguishing different timesteps during the denoise process.

However, the use of ControlNet combined with 3D convolution layers is still insufficient to generate stereo image pairs. Despite a certain degree of consistency between the left and right images, the main objects within these images do not maintain a strict correspondence. For example, a car appearing in the center of the left image may appear in a considerably random position in the right image. Although the KITTI dataset is captured with the same devices and, in theory, 3D convolution blocks should be able to learn the devices' parameters and estimate the displacement of objects in the right image relative to the left, this proves to be quite challenging in practice. Hence, we introduced a disparity

map as an additional condition. Our purpose was to use the disparity map of the left image as guidance to assist the 3D convolution blocks in estimating the pixel displacement in the right image. Using the disparity map as an additional condition for Stable Diffusion significantly improved the quality of the generated images, but the detail quality still did not meet our standards. Even when limiting the generation type to driving scenes, the probability of producing flawed images remained high. Therefore, we abandoned this approach. Fig. 10 shows the example of images generated using fine-tuned Stable Diffusion.

12. Ablation of Bidirectional attention and Stereo Pixel Shift

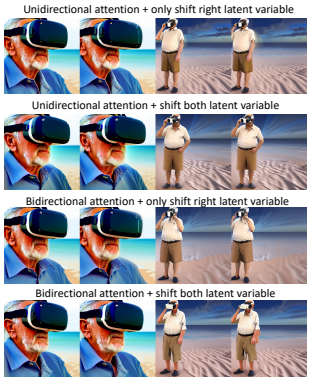
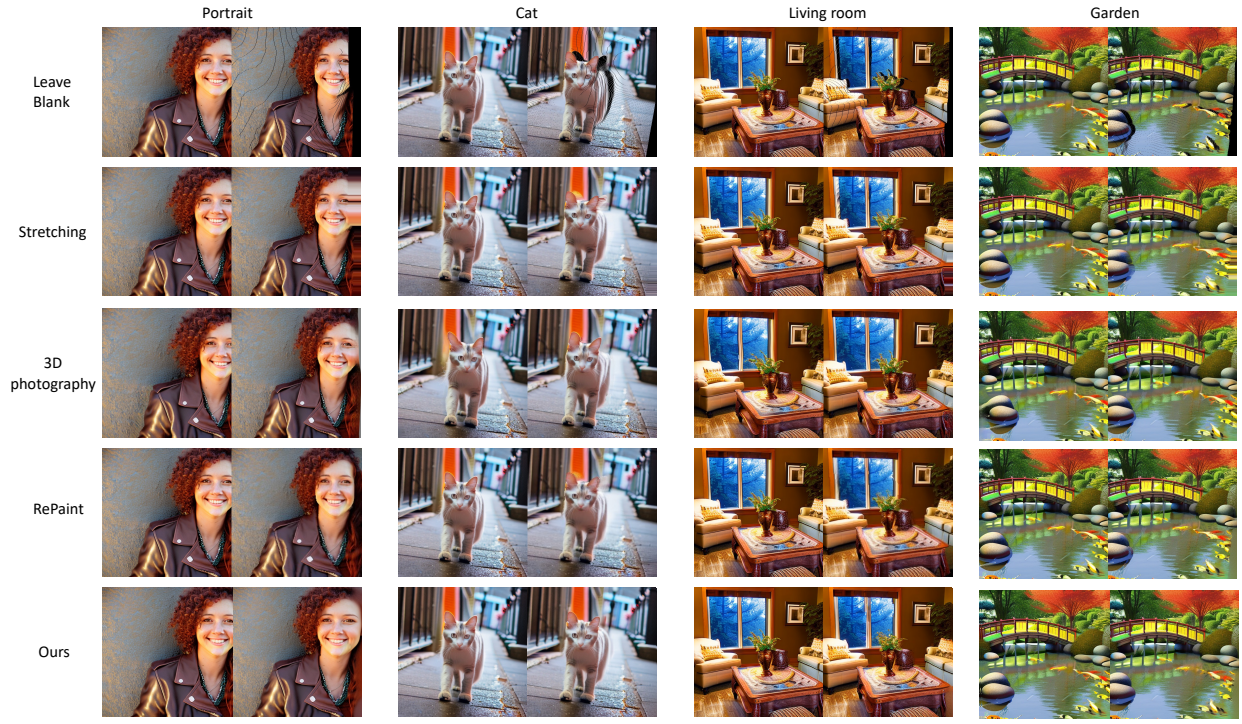


Figure 11. Ablation of Bidirectional attention and Stereo Pixel Shift: The implementation of Bidirectional attention and the simultaneous application of Stereo Pixel Shift to the left and right latent variables theoretically enhances the consistency between the two images. However, this approach may induce certain changes in the original images, which are currently uncontrollable.

Incorporating Bidirectional Attention and applying Stereo Pixel Shift to the both left and right latent variables can alter the original image, making it unsuitable for quantitative analysis. Therefore, we only partially showcase the results of the text prompt to stereo image generation, as depicted in Fig. 11. The simultaneous application of Bidirectional Attention and Stereo Pixel Shift to both left and right latent variables may induce changes in the original image. These modifications are currently uncontrollable. However, this may suggest a new potential of our approach: a method of controlling the generated images, akin to ControlNet, but without the need for fine-tuning.

13. User test images

In Fig. 12, we show the example images used for our user evaluation.



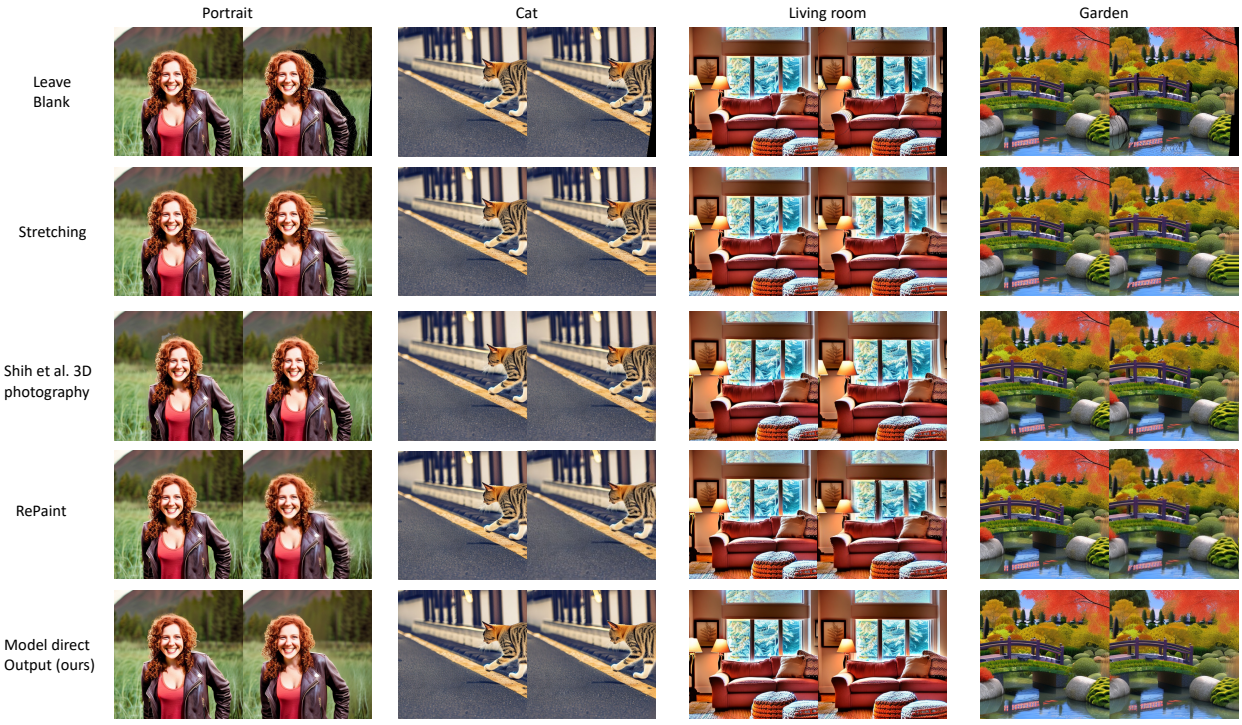


Figure 12. Comparison of stereo image generation techniques. RePaint et al. [16] indicates using their inpainting model to fill the blank area. HINT: The images can be viewed using the autostereogram technique to achieve a 3D effect. (Keep your eyes steady and maintain the unfocused gaze, try adjusting eyes' focus and the distance between the autostereogram and your eyes slightly.)