StereoCrafter: Diffusion-based Generation of Long and High-fidelity Stereoscopic 3D from Monocular Videos

Sijie Zhao* Wenbo Hu* Xiaodong Cun* Yong Zhang† Xiaoyu Li†
Zhe Kong Xiangjun Gao Muyao Niu Ying Shan

Tencent AI Lab ARC Lab, Tencent PCG

Project page: http://stereocrafter.github.io

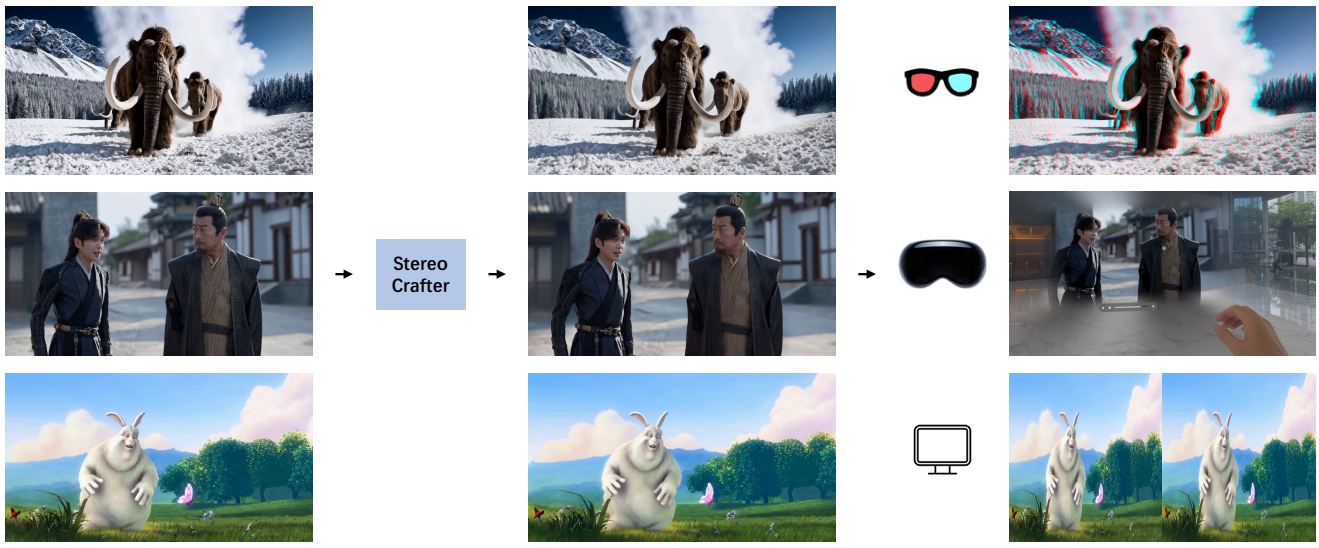

Figure 1. We propose a framework to convert any 2D videos to immersive stereoscopic 3D ones that can be viewed on different display devices, like 3D Glasses, Apple Vision Pro and 3D Display. It can be applied to various video sources, such as movies, vlogs, 3D cartoons, and AIGC videos. We hope this approach can be applied to revolutionize the way we experience digital media in the future.

Abstract

This paper presents a novel framework for converting 2D videos to immersive stereoscopic 3D, addressing the growing demand for 3D content in immersive experience. Leveraging foundation models as priors, our approach overcomes the limitations of traditional methods and boosts the performance to ensure the high-fidelity generation required by the display devices. The proposed system consists of two main steps: depth-based video splatting for warping and extracting occlusion mask, and stereo video inpainting. We utilize pre-trained stable video diffusion as the backbone and introduce a fine-tuning protocol for the stereo video inpainting task. To handle input video with varying lengths and resolutions, we explore auto-regressive strategies and tiled processing. Finally, a sophisticated data processing pipeline has been developed to reconstruct a large-scale and high-quality dataset to support our training. Our framework demonstrates significant improvements in 2D-to-3D video conversion, offering a practical solution for creating immersive content for 3D devices like Apple Vision Pro and 3D displays. In summary, this work contributes to the field by presenting an effective method for generating high-quality stereoscopic videos from monocular input, potentially transforming how we experience digital media.

* Equal contribution.
† Corresponding author.

1. Introduction

Pursuing a more immersive experience of digital content is gaining growing popularity due to the captivating and enjoyable psychological feeling of spatial presence. In contrast to traditional 2D digital content, immersive 3D content is becoming the next frontier, owing to the advancements in virtual reality (VR) and augmented reality (AR) technology in both hardware, such as the release of Apple Vision Pro, and software, including vision foundation models. However, a substantial volume of digital media like video clips, teleplays and movies on the internet are monocular which lack vividness for displaying in 3D, contrasted with the limited availability of 3D video content. Consequently, the conversion of these 2D videos into immersive 3D videos has been highly demanded.

The human visual system utilizes parallax between the left and right eye images to gain depth perception such that scenes are perceived in three dimensions rather than on a two-dimensional plane. Consequently, 3D videos are typically represented in a stereoscopic format. 2D-to-3D video conversion methods [26, 67] in the early stage usually consist of two main steps: the extraction of depth information from the input view and the rendering of a novel view based on the depth (Depth Image Based Rendering) to form a stereo pair. Deep3D [58] suggests directly regressing the right view using a pixel-wise loss, by predicting a probabilistic disparity-like map as an intermediary output. Nonetheless, owing to the limited training data and model capacity of convolutional neural networks, these methods tend to produce blurry results with limited generalization ability to real-world videos, which is far from the practice usage.

Recently, the emergence of 3D representations such as Neural Radiance Fields [41] (NeRF) and 3D Gaussian Splatting [25] (3DGS) have significantly transformed the field of novel view synthesis due to their high-quality results and simple reconstruction processes. Hence, an alternative approach for 2D-to-3D video conversion involves reconstructing the dynamic 3D scene from the input video and generating stereoscopic videos via novel view synthesis. However, these methods [31, 32, 35, 38, 43, 44] require estimating the camera pose of each frame from monocular videos, which heavily rely on the static parts in the video for calibration. For videos exhibiting large camera motion, sizable dynamic objects, or visual effects such as fog or fire, calibrating the cameras and reconstructing the scenes becomes a challenging task for these methods. In addition, these methods usually handle the occluded regions by blending the information from neighboring frames and cannot address the occlusion that does not appear in the neighboring frames. Consequently, we believe that dynamic 3D reconstruction methods are not a practical and effective solution for producing stereoscopic videos.

Meanwhile, trained from large-scale data, foundation models [7, 49, 52, 53] have emerged and garnered lots of attention due to their strong zero-shot performance in various downstream tasks. Benefiting from these basic models, recent works for monocular depth estimation from a single image [23, 46, 62, 66] have demonstrated remarkable results with substantial improvements compared with traditional methods. On the other hand, the utilization of basic video diffusion models has boosted the performance of related tasks such as video generation [5, 42], video editing [40, 48] and video inpainting [37, 71]. These developments inspire us to rethink the problem of 2D-to-3D video generation, which does not have a practical solution due to the limited performance in depth estimation and inpainting with traditional methods.

Therefore, by leveraging foundation models as model priors, we present a framework for converting 2D videos of various types to stereoscopic 3D which could be immersively experienced with devices like Apple Vision Pro and 3D displays as shown in Fig. 1. We introduce a practical solution for 2D-to-3D video conversion and achieve usable quality for the industry. Our system consists of two main steps: depth-based video splatting and stereo video inpainting. We first employ a depth estimation method to give us depth maps of the input video. Utilizing this depth map, we warp the input video from the left view to the right view via a depth-based video splatting method, which concurrently produces an occlusion mask. Subsequently, based on the warped video and its corresponding occlusion mask, we generate the final right-view video using our stereo video inpainting method.

To generalize our framework to input videos with various types, we first employ pre-trained stable video diffusion [5] as the backbone of our network. Leveraging this diffusion prior trained on a large-scale dataset, video quality and consistency of the results could be greatly ensured. Subsequently, we propose a fine-tuning protocol to adapt the model for the stereo video inpainting task, which requires data comprising occluded videos, occlusion masks and complemented videos. To reconstruct this dataset, we present a data processing pipeline that utilizes our video splatting approach based on collected stereo videos. Finally, to adapt the model to the input videos with varying lengths and resolutions, an auto-regressive strategy and tiled processing are explored in our work.

Our major contributions can be summarized as follows:

- We develop a framework for converting 2D videos to stereoscopic 3D with immersive experience leveraging diffusion model priors and our reconstructed dataset.
- We design a data processing pipeline to facilitate the training of our approach with high-quality data.
- We introduce a depth-based video splatting that could produce accurate warped videos and occlusion masks in

parallel for each pixel, which has a fast processing speed running on modern GPUs.
- We propose a stereo video inpainting network with an auto-regressive strategy and tiled processing to handle input video with different lengths and resolutions.

2. Related Work

2D-to-3D Video Conversion. 2D-to-3D video conversion has been an active area of research in computer vision and graphics since the popularity of head-mounted displays for 3D content and 3D movie production. Early approaches [26, 67] for 2D-to-3D video conversion relied on depth estimation and image-based rendering techniques. Specifically, Deep3D [58] introduces an end-to-end network to directly generate the right view from the left view. Lang et al. [28] propose a method to retarget stereoscopic 3D video automatically to a novel disparity range. Other approaches leverage deep learning methods for video depth estimation [27, 68] and utilize the depth for novel view synthesis. Recently, some methods have explored the use of diffusion models for various tasks including stereo generation [11, 55]. Despite these advancements, generating high-quality, consistent stereoscopic videos from monocular input remains challenging, especially for scenes with complex motion or occlusions.

Dynamic View Synthesis from Monocular Videos. Recent advances in 3D reconstruction, such as NeRF [41] and 3D Gaussian Splatting [25], have greatly improved the quality of novel view synthesis, which also facilitate the view synthesis for a dynamic scene from monocular videos [15, 30, 32, 35, 38, 43, 54, 56]. By reconstructing dynamic scenes, these methods can synthesize space-time results, which can also be utilized for creating stereoscopic videos. However, relying on camera poses as input or jointly optimizing camera poses within the method makes it challenging to handle complex dynamic scenarios where camera poses are hard to optimize.

Video Diffusion Models. The video generation methods have rapid development in recent studies due to the stronger abilities of the diffusion model [17]. Early works [16, 19] directly train the multi-scale video diffusion models from the video data. Thanks to the large-scale pre-trained text-to-image model, *i.e.*, Stable Diffusion [47, 52], adding temporal layers to the text-to-image models for text-to-video generation are also popular [8, 9, 18, 21, 57, 70]. More recently, 3D-VAE [69] based video diffusion model, *i.e.*, Sora [6] show more advanced results on this topic. These text-to-video diffusion models provide a strong visual backbone for other conditional generation tasks. *e.g.*, the image-to-video generation [5, 42], video-editing [40, 48], video-to-video translation and enhancement [59, 64], frame-interpolation [65]. In this paper, we also utilize the pre-trained knowledge from the video diffusion model for our stereo video generation task.

3. Methodology

3.1. Overview

We propose a stereo video generation framework that converts a monocular video into a stereo video for an immersive experience, which can be viewed in VR/AR devices or 3D display. As shown in Fig. 2, the framework consists of two main stages: depth-based video splatting and stereo video inpainting. We first determine the depth of the input monocular video by utilizing a video depth estimation model and perform depth-based video splatting to warp the input video from the left view to the right view, simultaneously obtaining the corresponding occlusion mask. Then, we train a diffusion model for stereo video inpainting to fill the holes of the warped video based on the occlusion mask, resulting in the final right view. The input left and completed right videos can be viewed on stereo display devices like Vision Pro. Finally, we present our data processing pipeline designed for constructing the training dataset, which significantly contributes to our success.

3.2. Depth-based Video Splatting

We utilize disparity maps to synthesize right-view videos from the left-view input, necessitating a depth estimation method to predict the depth of the input video. Numerous depth estimation methods have been proposed in the past [3, 12, 13, 29, 34, 45, 60], and significant progress has been made recently in this field [4, 14, 24, 46, 51, 61, 63, 66] benefiting from the utilization of model priors, such as stable diffusion and DINO. Consequently, we adopt the state-of-the-art depth estimation method DepthCrafter [20] or Depth Anything V2 [63] to obtain detailed video depth for the input video. DepthCrafter [20] is capable of producing more temporally consistent video depth results than Depth Anything V2 [63], making it a better fit for our task.

After estimating the depth of the input video, our goal is to warp the input left video to the right view, using the disparity calculated from the depth. The disparity map indicates the target position for each pixel in the left image, as it transitions from the left to the right image. It is important to note that disparity is a forward mapping and common resampling techniques such as backward warping and interpolation cannot be applied to it. Consequently, we propose a forward splatting method capable of mapping each source pixel to the target image based on its disparity as shown in Fig. 3.

During forward splatting, we map each pixel in the source image to its target position and splat it onto the four

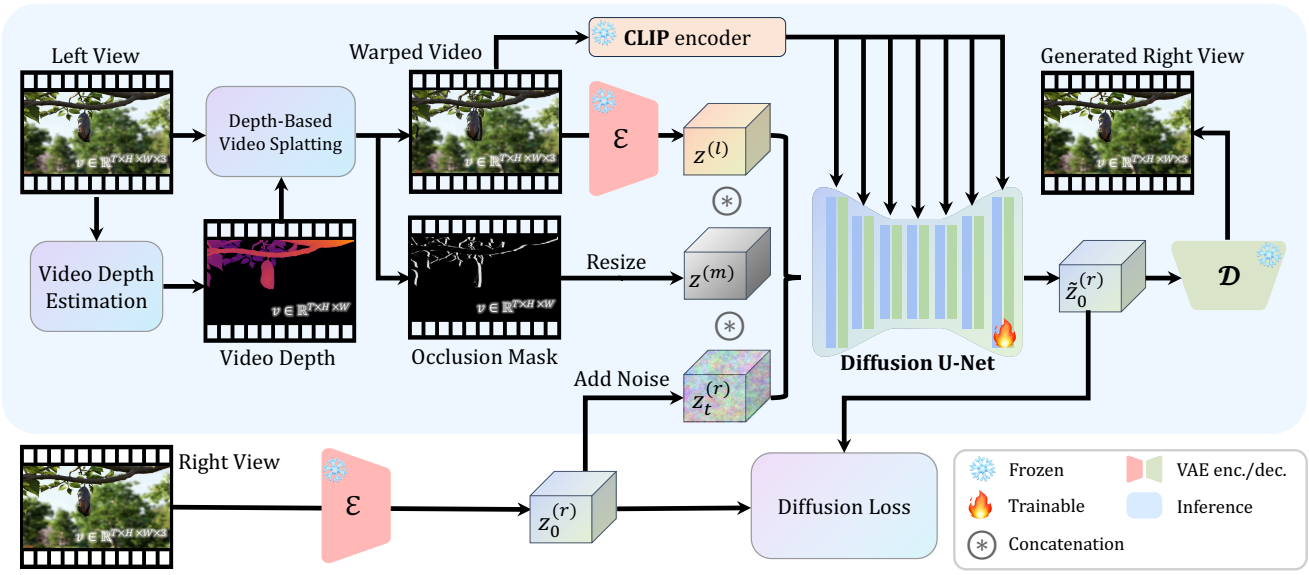

Figure 2. Overall framework of StereoCrafter, which contains two main stages. In the first stage, the video depth is estimated from the monocular video and we obtain the warped video and its occlusion mask through depth-based video splatting with the left video and the video depth as input. Then, we train a stereo video inpainting model to fill in the hole region of the warped video according to the occlusion mask to synthesize the right video.

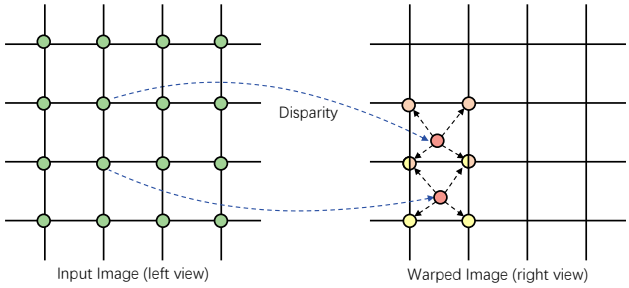

Figure 3. Illustration of our depth-based forward splatting. The image on the right is created by splatting the input pixels according to the disparity. And we use a depth-aware method to resolve any ambiguity when multiple pixels are splatted to the same pixel in the right view.

nearest pixels in the target image's grid based on its distance to the target position. However, multiple pixels from the source image may be mapped to the same pixel in the target image, creating ambiguity that requires a proper solution as illustrated in Fig. 3. To address this issue, we calculate the weights of the splatted pixel based on the disparity value of its source pixel. We then accumulate all the splatted pixels corresponding to the same target pixel and blend them according to their weights to obtain the final color of the target pixel. It is important to note that a large disparity signifies a smaller depth of the pixel, indicating that it is closer than other pixels. As a result, our method assigns a larger blending weight to such pixels. Therefore, we calculate the weight according to the formula $w = \sqrt{2}^{disp}$. For target pixels without any splatted pixels, we mark them as occlusion pixels and calculate the occlusion mask for the subsequent inpainting process. We have implemented a parallel version of our splatting approach on the GPU, which can run in real-time on modern GPUs.

3.3. Stereo Video Inpainting

Given the warped video and corresponding occlusion mask, we introduce a stereo video inpainting method to address the occluded pixels and synthesize the output right-view video. As shown in Fig. 2, we extend the Stable Video Diffusion (SVD) for stereo video inpainting, which includes: (1) changing the condition of SVD from image to warped frames; (2) adding an extra channel in the input layer of Unet (i.e., increasing from 8 to 9) to input the occlusion mask, and we set the parameters corresponding to this channel in the first convolutional layer to 0 for zero initialization.

Based on our video inpainting model, we propose the following methods to achieve stereo video inpainting of arbitrary length and resolution while keeping the consistency of the generated results. (1) **Auto-regressive modeling**. A common video clip may have hundreds of frames, while the SVD released version can only generate 25 frames. Therefore, multiple inferences are needed when the input video is longer. However, simply splitting the video along the time dimension and processing it independently will result in inconsistencies in the inpainting area between adjacent blocks. Therefore, we propose auto-regressive mod-

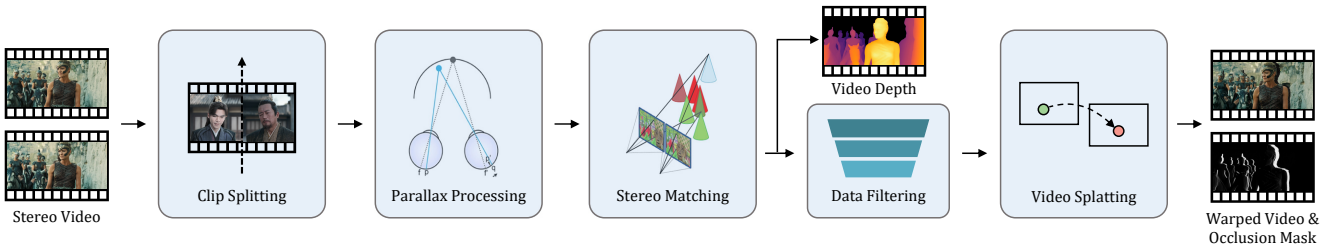

Figure 4. The pipeline of our approach for constructing the training dataset. After curating a large number of stereo videos, we generate the video depth/disparity, warped left video, and occlusion mask for each data sample, while using the right video as the ground truth.

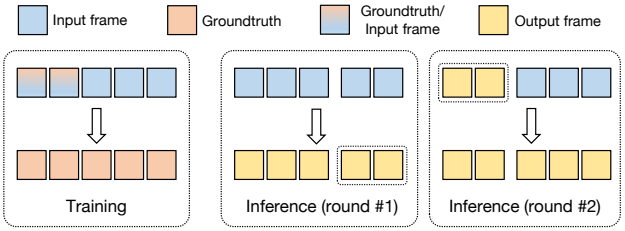

Figure 5. Illustration of our approach for handling videos of arbitrary length.

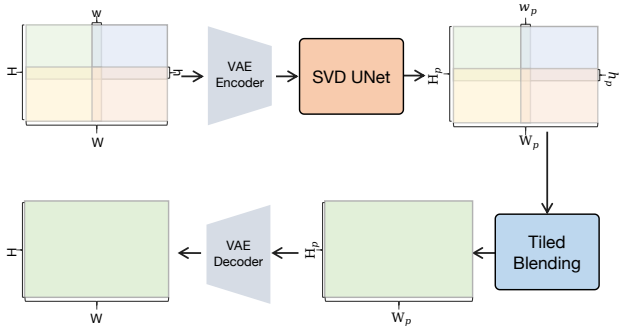

Figure 6. Illustration of our approach for handling high resolution input.

eling to deal with videos of arbitrary length. As shown in Fig. 5, during the training process, we randomly sample a value n from 0 to N and replace the first n input frames with ground truth. Then we use the standard diffusion loss for training. During the inference process, we concatenate the last m frames generated in the previous round with the subsequent input frames as the input for the next round. Therefore, the model can generate more temporally consistent contents by combining the inpainting results from the previous round. (2) **Tiled diffusion**. Video diffusion models require a large amount of memory during the inference process, making it difficult to process high-resolution videos in limited memory. Therefore, we propose tiled diffusion for high-resolution video inpainting. As shown in Fig. 6, We first divide the high-resolution video into blocks along the spatial dimension, then use the video diffusion model to independently infer each block, and then blend the overlap areas of adjacent blocks in the latent space. Taking horizontal blending as an example, $blended = \mathbf{w} * left + (1 - \mathbf{w}) * right$, where $\mathbf{w}$ increases from 0 to 1 in the horizontal direction. Afterwards, we use VAE for decoding to obtain the inpainting results at the target resolution. With this tiled diffusion processing, we could break through the memory limitation to process videos at high resolution.

3.4. Dataset Construction

To boost the performance of our stereo video generation methods, an appropriate dataset is required for training, as the quality of the dataset significantly influences the fidelity of the results produced by diffusion models. However, there is no existing dataset that we could use directly, which should include a warped left video and an occlusion mask as inputs, a completed right video as the ground truth for each data sample.

As illustrated in Fig. 4, the data pipeline begins with curating a diverse collection of stereo videos spanning various categories. We employed the PySceneDetect tool to perform shot detection within each stereo video, facilitating automated segmentation into individual clips. Subsequently, the stereo matching method outlined in [22] was applied to predict disparity maps between the left and right views of each clip, enabling accurate reconstruction of stereo content.

However, stereo videos in the wild may exhibit varying disparity ranges based on the distinct definitions of the zero disparity plane in the scene. Directly feeding these videos to video stereo matching methods can result in inaccurate results, as the learning-based methods are typically trained on a specific dataset within a particular disparity range. To address this issue, we employ a parallax processing step before stereo matching. Specifically, we shit the right view video to the left by a certain amount and crop the left view video accordingly until the disparity for all pixels is negative and the maximum disparity is nearly zero. After this processing, we align the disparity distribution of the stereo

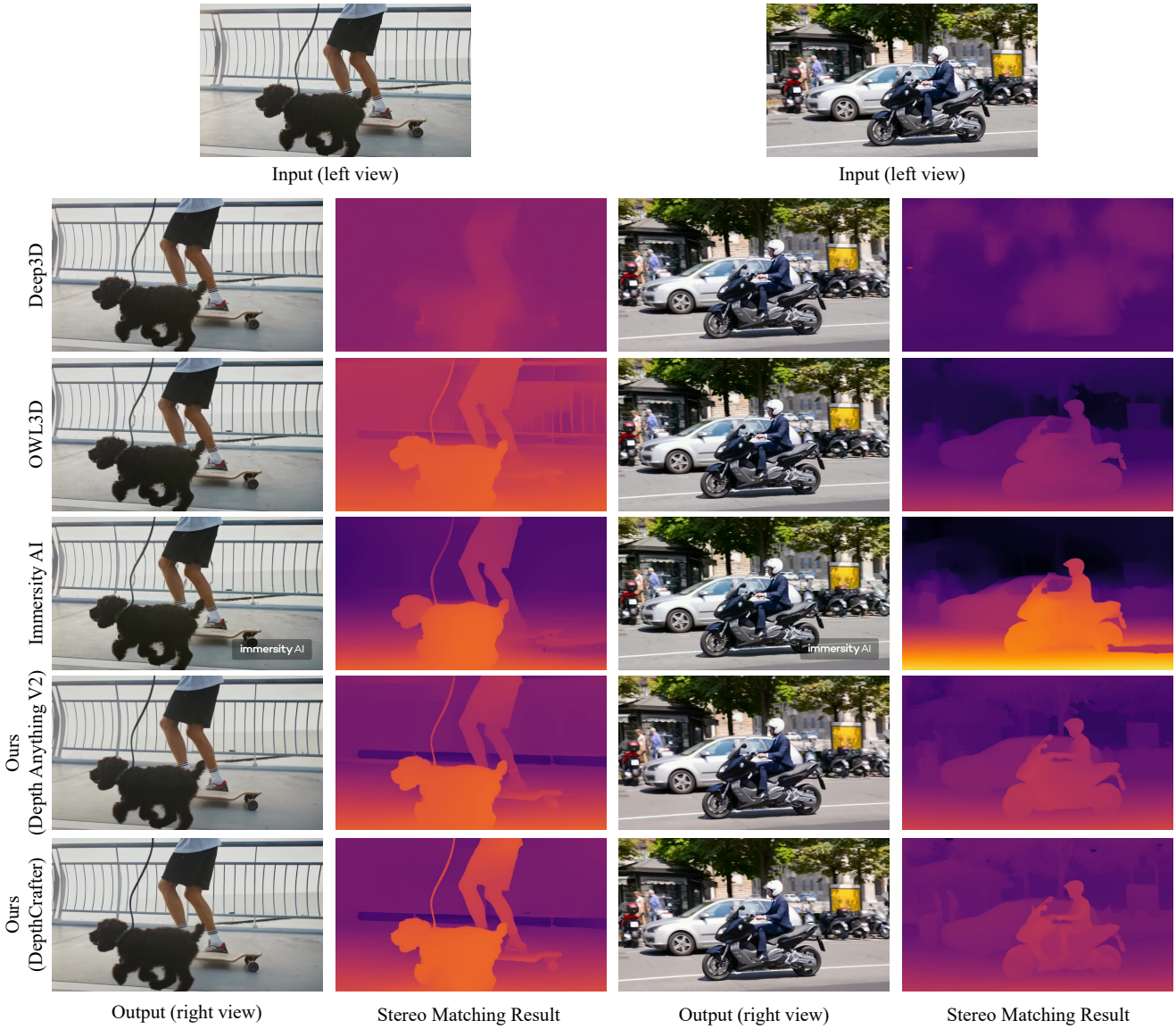

Figure 7. Qualitative comparison results of our approach with different 2D-to-3D conversion methods. Our approach could synthesize high-quality results using different depth estimation methods like Depth Anything V2 [63] or DepthCrafter [20], while maintaining consistency with the input left view as shown by the matching results.

video to the training data of the video stereo matching methods, thereby yielding more accurate matching results.

After obtaining the disparity map, we warp the left view video to the right view using our depth-based video splatting method, which will output the warped video and its corresponding occlusion mask. In addition, to exclude the data with large stereo matching errors, we filter the data by calculating the PSNR between the right view video and the warped left view video and only retain samples with PSNR greater than 25dB. Through the above processing steps, we have collected approximately 180k training sequence samples with about 25 million frames.

4. Experiments

4.1. Implementation Details

Datasets. We curate a large number of stereo videos as the data sources, which are cropped from the long video into different clips and decoded into left view videos V_{left} and right view videos V_{right}. To get pseudo ground truth depth, we calculate the disparity map between the left view video and right view video using the video stereo matching method [22], i.e., $D_{left} = Matching(V_{left}, V_{right})$. Subsequently, based on this disparity map, we perform forward splatting on the left view video according to the method in

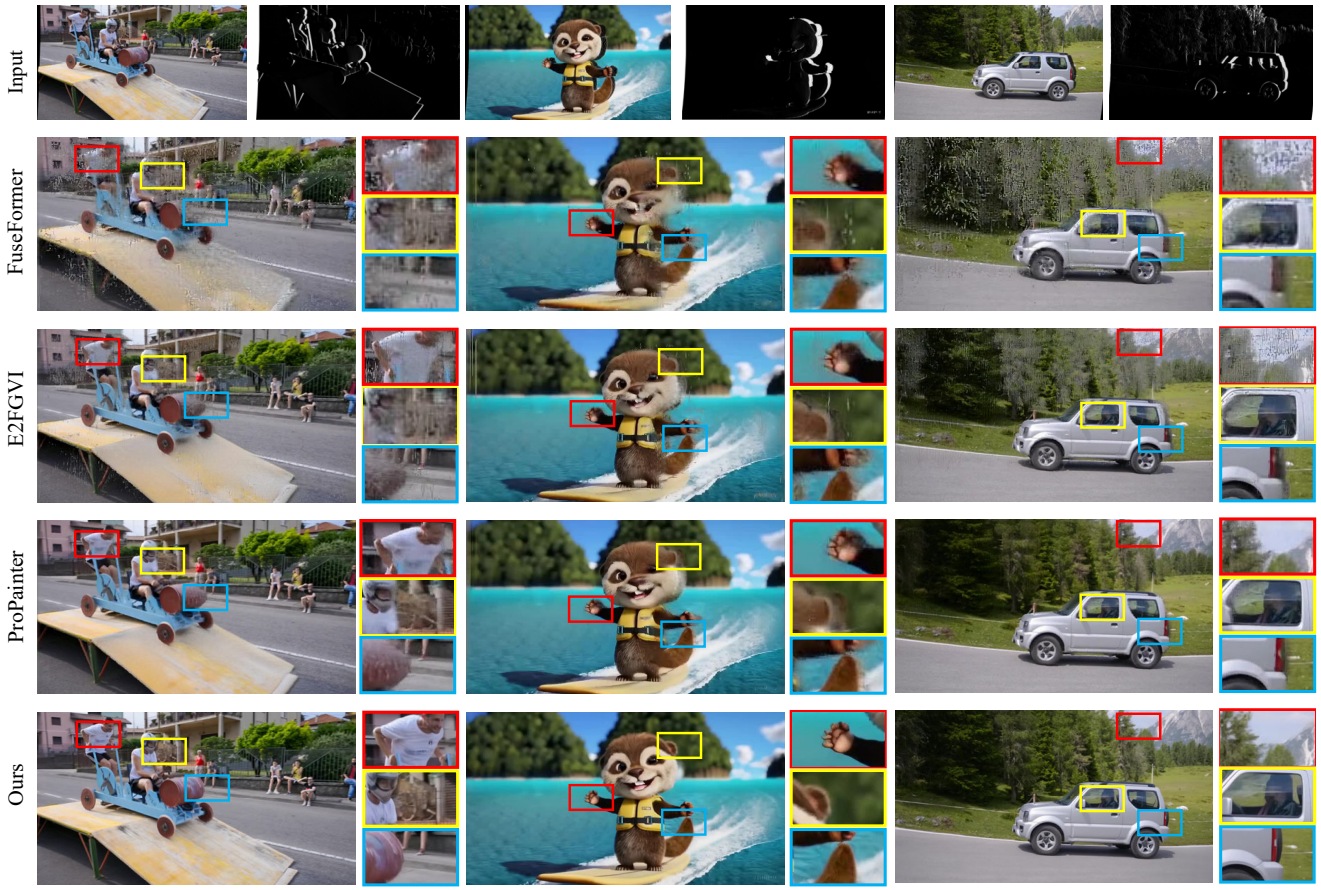

Figure 8. Qualtitative comparison results of different video inpainting models. Our method is capable of producing sharper results in the occluded areas.

Sec. 3.2, obtaining the warped video and occlusion mask, i.e., $V_{warped}, M_{occlusion} = Splatting(V_{left}, D_{left})$. Ultimately, the training data pairs required for stereo video inpainting are formed by $(V_{warped}, M_{occlusion}, V_{right})$.

Training Details. We train our stereo inpainting model based on the aforementioned dataset. We initialize the model with pre-trained weights from SVD and only fine-tune the spatial layers in the U-Net. We sample the training data from the dataset with a resolution of $25 \times 576 \times 1024$ and a frame stride ranging from 1 to 6. We used a constant learning rate of 1e-5 with the AdamW [39] optimizer. The training is conducted on 8 A100 GPUs with a batch size of 1 per GPU and 26K iterations. For training efficiency, we employ deepspeed stage 2 [50], gradient checkpointing [10] techniques, and train with float16 precision.

4.2. Comparison to 2D-to-3D Video Conversion

Firstly, We compare our framework with traditional 2D-to-3D video conversion methods Deep3D [58] and some 2D-to-3D conversion software Owl3D [2] and Immersity AI [1]. In particular, Deep3D [58] proposes a fully automatic 2D-to-3D conversion approach that is trained end-to-end to directly generate the right view from the left view using convolutional neural networks. Owl3D [2] is an AI-powered 2D to 3D conversion software and Immersity AI is a platform converting images or videos into 3D. For Owl3D and Immersity AI, we upload the input left view videos to their platform and generate the right view video for comparison. The qualitative comparison results are shown in Fig.7. In addition to showing the right view results, we also employ a video stereo matching approach [22] to estimate the disparity between input left view video and output right view video to verify its spatial consistency. As shown in Fig.7, Deep3D [58] could generate overall promising right view results, but is not spatially consistent with the input video according to the stereo matching results. On the other hand, Owl3D and Immersity AI could generate more consistent results, but some artifacts appear in the images, such as the handrail in the first example. In the end, our method could synthesize high-quality image

results while keeping consistency with the left view images from the stereo matching results using different depth estimation methods. With more temporally consistent video depth predicted by DepthCrafter, our method could achieve even better results.

4.3. Comparison to Video Inpainting

We show the qualitative results of our method on the stereo video inpainting and compare it with previous video inpainting models, including FuseFormer [36], E2FGVI [33] and ProPainter [71]. As shown in Fig. 8, previous inpainting models suffer from the problem of generating blurry content in the occluded areas. In addition, FuseFormer and E2FGVI also face serious image quality issues in non-occluded areas, making these methods difficult to apply in real-world video inpainting scenarios. On the other hand, our method maintains high consistency with warped videos in non-occluded areas while generating pleasing results in occluded areas.

4.4. Ablation Study

Auto-regressive modeling. We evaluated the effectiveness of auto-regressive modeling through the following experiments: (1) **'w/o overlap'**, where each round of video frames is inferred independently and then concatenated together; (2) **'w/ overlap'**, where each round after the first uses the last n frames of inpainting results from the previous round as input, with $n = 3$ in this experiment. The experimental results are shown in Fig. 9. It can be observed that in the 'w/o overlap' case, inconsistencies appear in the inpainting area between the front and back frames of adjacent rounds, while 'w/ overlap' can solve this problem.

Tiled diffusion. We validate the effectiveness of tiled diffusion through the following experiments: (1) when the resolution is low, e.g. 512×960, perform global inpainting inference in the spatial dimension; (2) when the resolution is high, e.g. 1024×1920, use tiled diffusion in the spatial dimension for multiple inferences and then obtain inpainting results through blending. The results are shown in Fig. 10. It can be observed that using tiled diffusion at high resolution can achieve better detailed results within the same memory constraints. Without tiled diffusion, inferring videos with a resolution of 1024×1920 is not feasible due to the high memory usage of GPUs.

5. Conclusion and Future Work

We have introduced a novel framework for converting 2D videos into stereoscopic 3D content to meet the growing demand for immersive digital experiences driven by advancements in VR and AR technologies. Our approach leverages foundation models as priors to enhance video depth estimation, achieving high-quality and detailed video depth

Figure 9. Ablation results of auto-regressive modeling. We concatenate the last n frames generated from the previous round with the warped frames of current round as input. When $n = 0$, i.e., without overlap, the inpainting results of adjacent rounds cannot maintain temporal consistency, as shown in the second row.

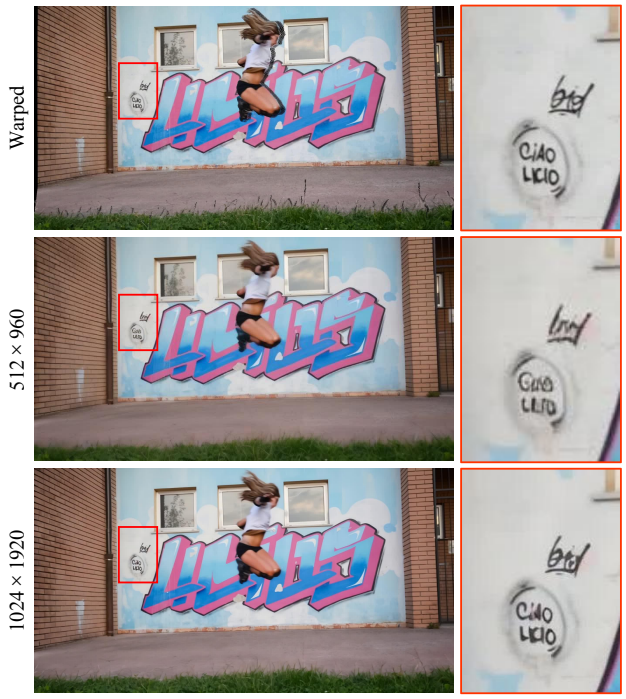

Figure 10. Comparison results of our method at different resolutions. We use the tiled diffusion described in Sec. 4.3 to handle high-resolution videos, Which maintains more details in the generated videos.

maps. By combining video depth estimation, video splatting, and stereo video inpainting, our system successfully converts 2D videos into stereoscopic 3D videos that can be experienced with different devices like Apple Vision Pro.

Future Work. While our proposed framework achieves promising results, several areas for future work remain to further enhance 2D-to-3D video conversion: (1) Future research could focus on developing more advanced depth estimation techniques that can provide even higher accuracy and consistency, particularly in challenging scenarios involving high motion or complex visual effects. (2) Optimization of the framework to support real-time video conversion would be a significant advancement, making the technology more practical for live streaming and real-time applications.

References

[1] Immersity ai: The ai platform converting images and videos into 3d, https://www.immersity.ai/.

[2] Owl3d: Ai-powered 2d to 3d conversion software, https://www.owl3d.com/.

[3] Shubhra Aich, Jean Marie Uwabeza Vianney, Md Amirul Islam, and Mannat Kaur Bingbing Liu. Bidirectional attention network for monocular depth estimation. 2021.

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[9] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

[10] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

[11] Peng Dai, Feitong Tan, Qiangeng Xu, David Futschik, Ruofei Du, Sean Fanello, Xiaojuan Qi, and Yinda Zhang. Svg: 3d stereoscopic video generation via denoising frame matrix. *arXiv preprint arXiv:2407.00367*, 2024.

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 27, 2014.

[13] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018.

[14] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *ECCV*, 2024.

[15] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021.

[16] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022.

[20] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024.

[21] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *CVPR*, 2022.

[22] Junpeng Jing, Ye Mao, and Krystian Mikolajczyk. Match-stereo-videos: Bidirectional alignment for consistent dynamic stereo matching. *arXiv preprint arXiv:2403.10755*, 2024.

[23] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[24] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024.

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[26] Janusz Konrad, Meng Wang, Prakash Ishwar, Chen Wu, and Debargha Mukherjee. Learning-based, automatic 2d-to-3d image and video conversion. *IEEE Transactions on Image Processing*, 22(9):3485–3496, 2013.

[27] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021.

[28] Manuel Lang, Alexander Hornung, Oliver Wang, Steven Poulakos, Aljoscha Smolic, and Markus Gross. Nonlinear disparity mapping for stereoscopic 3d. *ACM Transactions on Graphics (TOG)*, 29(4):1–10, 2010.

[29] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[30] Yao-Chih Lee, Zhoutong Zhang, Kevin Blackburn-Matzen, Simon Niklaus, Jianming Zhang, Jia-Bin Huang, and Feng Liu. Fast view synthesis of casual videos with soup-of-planes. *arXiv preprint arXiv:2312.02135*, 2023.

[31] Yao-Chih Lee, Zhoutong Zhang, Kevin Blackburn-Matzen, Simon Niklaus, Jianming Zhang, Jia-Bin Huang, and Feng Liu. Fast view synthesis of casual videos. *arXiv preprint arXiv:2312.02135*, 2023.

[32] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021.

[33] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[34] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *Machine Intelligence Research*, 20(6):837–854, 2023.

[35] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4284, 2023.

[36] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14040–14049, 2021.

[37] Weihuang Liu, Xiaodong Cun, Chi-Man Pun, Menghan Xia, Yong Zhang, and Jue Wang. Coordfill: Efficient high-resolution image inpainting via parameterized coordinate querying. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1746–1754, 2023.

[38] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13–23, 2023.

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[40] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023.

[41] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[42] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. *arXiv preprint arXiv:2405.20222*, 2024.

[43] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.

[44] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

[45] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *CVPR*, 2022.

[46] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024.

[47] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[48] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv:2303.09535*, 2023.

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[50] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.

[51] 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.

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[53] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[54] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021.

[55] Lezhong Wang, Jeppe Revall Frisvad, Mark Bo Jensen, and Siavash Arjomand Bigdeli. Stereodiffusion: Training-free stereo image generation using latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7416–7425, 2024.

[56] Shizun Wang, Xingyi Yang, Qiuhong Shen, Zhenxiang Jiang, and Xinchao Wang. Gflow: Recovering 4d world from monocular video. *arXiv preprint arXiv:2405.18426*, 2024.

[57] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023.

[58] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 842–857. Springer, 2016.

[59] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *arXiv preprint arXiv:2306.00943*, 2023.

[60] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *ICCV*, 2021.

[61] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

[62] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024.

[63] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024.

[64] Qinyu Yang, Haoxin Chen, Yong Zhang, Menghan Xia, Xiaodong Cun, Zhixun Su, and Ying Shan. Noise calibration: Plug-and-play content-preserving video enhancement using pre-trained video diffusion models. *arXiv preprint arXiv:2407.10285*, 2024.

[65] Shaoshu Yang, Yong Zhang, Xiaodong Cun, Ying Shan, and Ran He. Zerosmooth: Training-free diffuser adaptation for high frame rate video generation. *arXiv preprint arXiv:2406.00908*, 2024.

[66] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023.

[67] Liang Zhang, Carlos Vazquez, and Sebastian Knorr. 3d-tv content creation: automatic 2d-to-3d video conversion. *IEEE Transactions on Broadcasting*, 57(2):372–383, 2011.

[68] Zhoutong Zhang, Forrester Cole, Richard Tucker, William T Freeman, and Tali Dekel. Consistent depth of moving objects in video. *ACM Transactions on Graphics (ToG)*, 40(4):1–12, 2021.

[69] Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan. Cv-vae: A compatible video vae for latent generative video models. *https://arxiv.org/abs/2405.20279*, 2024.

[70] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

[71] Shangchen Zhou, Chongyi Li, Kelvin C.K Chan, and Chen Change Loy. ProPainter: Improving propagation and transformer for video inpainting. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2023.