Eye2Eye: A simple approach for monocular-to-stereo video synthesis

Michal Geyer[1,2] Omer Tov[1] Linyi Jin[1,3] Richard Tucker[1]
Inbar Mosseri[1] Tali Dekel[1,2] Noah Snavely[1]
[1]Google DeepMind [2]Weizmann Institute of Science [3]University of Michigan

Figure 1. **3D anaglyph visualization of stereo videos produced by our method.** Our model takes as input a standard monocular video (representing a right-eye view) and produces a left-eye video, enabling stereoscopic viewing using 3D glasses or a VR headset. Our pipeline directly produces the new viewpoint, avoiding steps like explicit depth estimation and warping, and thus can plausibly handle videos with specular and transparent surfaces, like the wine glass in the left example and the shiny floor in the right example, where assumptions of a single depth per pixel do not hold.

Abstract

The rising popularity of immersive visual experiences has intensified interest in stereoscopic 3D video generation. Despite significant advances in video synthesis, creating 3D videos remains challenging due to the scarcity of 3D video data. We propose a simple approach for transforming a 2D text-to-video generator into a video-to-stereo generator. Given an input video, our system automatically produces the video frames from a shifted viewpoint, enabling a compelling 3D effect. Prior and concurrent approaches for this task typically begin with an analysis phase, estimating the video disparity or depth, then warping it accordingly to produce another view, and finally inpainting the disoccluded regions. These methods inherently fail when the input video contains specular surfaces or transparent objects. In such cases, single-layer disparity estimation is insufficient, resulting in artifacts and incorrect pixel shifts during warping. In this work, we directly leverage the pre-trained model's priors on geometry, object materials, optics, and semantics, without relying on external geometry models and manually disentangling geometry from the synthesis process. We demonstrate the advantages of our approach in complex, real-world scenarios featuring diverse object materials and compositions, thereby opening new directions for 3D video generation.

1. Introduction

Alongside the growing popularity of immersive experiences, there is increasing demand for methods that produce 3D content, such as stereoscopic video. Towards the goal of making it simple to create such content, we address the problem of up-converting monocular 2D videos to more immersive stereoscopic 3D videos, leveraging recent advances in generative video models. Many prior approaches to mono-to-stereo video conversion adopt a two-step process: they first estimate geometry for an input video via monocular disparity or depth models, and then use this geometry to re-project to a second view, inpainting dis-occluded regions

to generate a video for the second eye.

However, such *warp-and-inpaint* approaches have inherent shortcomings. For instance, even state-of-the-art disparity estimation and inpainting models can exhibit temporal inconsistencies. More fundamentally, using a disparity map to warp an image to a new viewpoint assumes that there is a single, distinct depth at every input pixel. For scenes that exhibit simple Lambertian reflectance, this assuming largely holds true. However, for scenes with more complex light transport—specular reflection, partial transparency, etc.—we often cannot characterize each pixel with a single depth. For instance, when viewing a store display through a plate glass window, many pixels will be a mixture of what we see inside the store, and objects behind us reflected in the glass—each at completely different (virtual) depths. To correctly handle such cases, methods based on explicit pixel warping would need to decompose the scene into multiple layers—e.g., the reflected and transmitted light through the window—warp each separately, then composite the results [29]. Without such special handling, these methods can produce physically implausible views, for instance, where reflections appear pasted on a reflective surface, rather than at their correct virtual depth. The effects of such artifacts have been widely studied in cognitive science, where it has been shown to affect the way shape, material, and geometry are perceived [4, 24, 26, 40, 41].

We propose to address these limitations by directly producing the output RGB view, sidestepping the need for explicit disparity estimation or pixel warping. We leverage recent video diffusion models for this goal, as well as the observation that stereo videos are relatively abundant online. Such videos represent ideal training data for mono-to-stereo methods, and allow us to learn to directly produce the desired output in a way that optimizes for the actual ground truth view, no matter how complex the underlying light transport is. We call this method **Eye2Eye**.

Our direct approach yields superior performance over warp-and-inpaint baselines in challenging real-world scenes featuring specular or transparent surfaces and dynamic lighting conditions. We validate these findings through extensive user studies.

In summary, our contributions are include: (1) Showing that medium-scale stereo datasets can be leveraged to tackle immersive video generation. (2) Demonstrating that video diffusion models can effectively capture geometry, appearance, and complex reflectance properties, enabling stereo generation in challenging scenarios. (3) Providing thorough quantitative and qualitative evaluations via user studies that highlight the advantages of our approach over existing warp-and-inpaint methods.

2. Related work

2.1. Multi-view video synthesis

Given the high-quality video generation capabilities of recent large-scale video diffusion models like Sora, Veo, and others [2, 5, 6], several works have aimed to leverage these models for 3D generation. CAT3D [9] trains a video diffusion model to take an arbitrary number of frames of a static scene as input, and generate as output a 360° set of views. They then showed that a 3D reconstruction can be estimated from these output views using off-the-shelf methods [14, 23]. While this approach works well for static scenes, it is computationally intensive and not scalable to the case of dynamic multi-view video generation. Scaling video-diffusion architectures to such a setting was tackled in [43]. While that model works fairly well on single object videos with simple motion, it fails to generate geometrically correct, high-quality views in more realistic settings. Other work that trains a video diffusion models to generate multi-viewpoint video outputs show some promise, but can fail to generate correct multi-view geometry in the face of arbitrarily complex scene motion [17, 38].

2.2. Stereo video synthesis

One notable work addressing video-to-stereo, Deep3D, uses a CNN to predict each right video frame from the input left, by first predicting a soft disparity map and using this to composite the output frame [42]. Since this model is trained from scratch, it does not benefit from the priors captured in large pre-trained models, and thus lacks a strong generative capability. Concurrent works on stereo-video generation include SVG [7] and StereoCrafter [45]. Both methods employ a multi-stage pipeline, involving (1) estimating video disparity (and temporally smoothing it), (2) using it to warp frames to the output view, and (3) inpainting the disoccluded regions. We refer to this strategy as *warp-and-inpaint*. While SVG leverages a pre-trained model without any further training, by devising a specific inpainting scheme, StereoCrafter fine-tunes an image-to-video model, modifying it (1) to be video- rather then image-conditioned, and (2) to inpaint left-right disocclusion regions. Our approach offers a key advantage over those pipelines. Namely, in many real-world scenarios, a single-layered disparity estimation simply does not suffice. While some progress on multi-layer flow prediction has been recently made [39], correctly estimating layered *video* disparity remains an overlooked challenge. Instead, we directly leverage a pre-trained model's priors on geometry, object materials, and light simultaneously, helping to to alleviate this issue.

2.3. Novel view synthesis with reflections and specularities

Another possible approach for stereo synthesis is to apply a 3D video reconstruction pipeline and render stereo views from it. Different lines of work have focused on improving the ability of 3D reconstruction methods to render and reconstruct scenes with specular reflections, including: (i) re-parameterizing the outgoing radiance as a function of the reflected view direction [19, 20, 35, 37], (ii) combining 3D reconstruction with inverse graphics (simultaneously estimating material properties) [3, 12, 21, 30], (iii) directly tracing reflection rays [36]. In the context of 3D *video* reconstruction, a recent work incorporates physically based rendering into a Gaussian-splatting 3D video reconstruction pipeline to handle specular reflections [8]. In contrast, our approach leverages the implicit modeling capabilities of a large pretrained video model, eliminating the need for explicit physics-based representations. Furthermore, existing 4D reconstruction pipelines rely heavily on the input video to constrain the learned geometry and appearance, and often fail when the input lacks sufficient information (for example, when the camera motion is minimal, as demonstrated in 6). These limitations make 3D video reconstruction pipelines less robust for stereo generation.

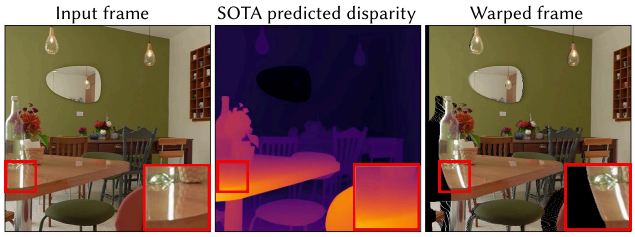

Figure 2. **Challenges faced by warp-and-inpaint methods for mono-to-stereo video conversion.** To demonstrate limitations of the common warp-and-inpaint style of method for stereo view synthesis, we take an input video frame (left) and use a state-of-the-art disparity estimation technique [11] to compute its disparity (middle). The predicted disparity map simply captures the surface of the table, without considering the reflection of other objects off of it, leading to a warped frame where the reflection is incorrectly skewed diagonally, and is not vertical as expected. When viewing such stereo frames in a VR headset, the reflection on the table appears "flat", as if it is a part of the table.

3. Preliminaries

3.1. Stereo geometry

The geometric relationship between corresponding points in a stereo pair is governed by epipolar geometry. For a rectified stereo setup with parallel camera projection planes, a 3D point (x, y, z) projects to image coordinates (u_L, v) in the left view and (u_R, v) in the right view, where the horizontal disparity $d = u_L - u_R$ is inversely proportional to depth: $d = \frac{fb}{z}$. Here, f is the focal length and b is the baseline distance between cameras.

3.2. Diffusion models

A diffusion model learns to reverse a noising process. Given a clean image x_0, the forward noising process adds Gaussian noise according to a variance schedule β_t, producing noisy samples $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon$, where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$ and $\epsilon \sim \mathcal{N}(0, I)$. The simplified diffusion objective minimizes:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right] \quad (1)$$

When conditioned on additional inputs c, the model learns the conditional distribution $\epsilon_\theta(x_t, t, c)$. During inference, the model iteratively denoises random noise x_T back to a clean sample.

3.3. Cascaded diffusion models

Cascaded diffusion models consist of two components: a base model that generates videos at low resolution, and a spatial-super-resolution (SSR) model that upsamples low-resolution outputs to a higher resolution. The SSR model is a conditional diffusion model that is trained to denoise high resolution videos conditioned on downsampled videos. At inference time, the SSR model iteratively denoises Gaussian noise into a high resolution video, conditioned on the low-resolution video produced by the base model.

4. Method

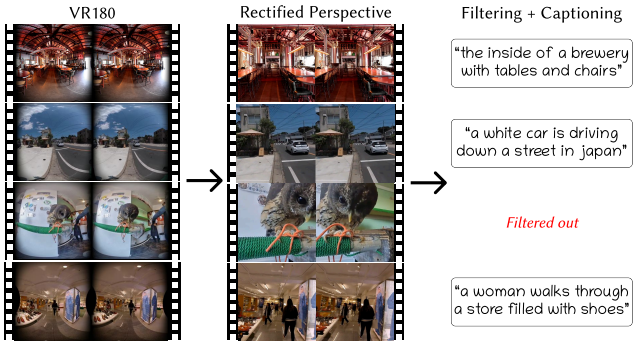

Figure 3. **Data processing pipeline.** We curate stereo VR180 footage captured with high-resolution cameras and stored in a equirectangular format. Following Stereo4D [13], we rectify the stereo videos and map the equirectangular format to perspective videos. We filter out videos with large disparity using RAFT [34] and caption the remaining videos with BLIP2 [18].

Given a monocular input video V, our goal is to synthesize its corresponding stereo pair by generating a left view V^{left}, as if captured by a camera horizontally shifted from

the original camera position by approximately human interpupillary distance (roughly 6.5cm), in accord with the rectified stereo geometry described in Section 3. This task presents two key challenges: (1) understanding the video geometry and light transport sufficiently well to determine how to transform the pixel content into the new view, and (2) generating realistic content for regions that are occluded in the original view, but become visible in the new viewpoint. Given that generative video diffusion models have been shown to capture priors on both scene geometry and occluded content [6, 10, 16, 28], we propose to leverage such models to jointly address both challenges. We extend Lumiere [2], a cascaded text-to-video diffusion model, to address this task. While we maintain Lumiere's two-stage process of low-resolution generation followed by super-resolution, we make two principal changes to adapt it to our task. First, our model takes a video as input, in addition to text. Second, we find that Lumiere's super-resolution design is not suitable for stereo synthesis, leading us to develop a different approach. We detail these modifications in the following sections, as well as our stereo dataset collection and processing. We call our overall method **Eye2Eye**.

4.1. Low-Resolution stereo generation

Our first step focuses on fine-tuning the base Lumiere model to produce left-from-right views. We do so by modifying its architecture to accept additional conditioning channels in its first input convolution layer. The model is trained to denoise the left view while being conditioned on the clean right view, following the standard conditional diffusion training formulation (as described in Section 3). This results in a model that produces novel left views at 128-pixel resolution (Fig. 4 top left). We call this model the **base Eye2Eye generator**.

4.2. High-Resolution stereo refinement

While the base stereo generator successfully creates left-from-right views, achieving high-resolution stereo synthesis presents additional challenges. Directly applying the pre-trained Lumiere super-resolution (SSR) model to the low-resolution left-view outputs would inappropriately alter details relative to the original input video, as the SSR denoises Gaussian noise based solely on the downsampled video. Our attempts to modify the Lumiere SSR model to produce left-from-right views resulted in degraded quality, which we attribute to the SSR model's fully convolutional, simpler architecture. Therefore, we focus on modifying the base model for high-resolution synthesis.

Consider a pixel in a video with a resolution of 128×128 that has a disparity of d pixels between the original and generated view. When generating a video at 512×512 resolution (4× higher), the disparity should scale proportionally ($4d$) pixels to maintain the same real-world depth effect. However, we observe that when sampling from our fine-tuned base model at different input resolutions, the pixel disparity remains at d pixels rather than scaling with the resolution. This leads to an undesirable effect: sampling at higher resolutions effectively reduces the perceived 3D depth in the stereo pairs, as shown in Fig. 5, columns 2 and 3. This behavior is analogous to changing the scale of the disparity itself.

To address this issue, we instead train on *high-resolution crops* of size 128×128 to learn correct-scale disparity and inpainting (Fig. 4 top right). We observe that although training on high-resolution crops indeed allows high-resolution sampling with larger pixel shifts, this approach introduces its own challenge: small crops often contain limited disparity variation and distant content. This causes the model to develop a bias toward largely and uniformly shifting the input view (Fig. 5 column 1).

What we would like is for the output geometry produced by the model trained on downsampled views, but with high-resolution details learned by the model trained on crops. We can achieve this desired effect by combining the two models in a two-stage inference pipeline, which exploits a fundamental property of diffusion models—early denoising steps establish global layout and structure, while later steps refine local details [22]. Specifically, we use the base Eye2Eye generator to produce a low-resolution layout with correctly scaled disparity, and use the model trained on crops as an *Eye2Eye refiner model*. That is, our method:

1. generates an initial low-resolution stereo video using the base Eye2Eye generator,
2. upsamples this output to the target resolution and noises the upsampled output,
3. and denoises the noised upsampled-resolution video using the *stereo refiner* model.

In other words, we perform SDEdit [22] on the upsampled low-resolution outputs. This combined approach preserves correct large-scale disparity from the LR generation while enabling high-resolution refinement of fine details and textures (Fig. 5, rightmost column). The result is a pipeline that consistently balances stereo disparity with high-resolution detail, effectively bridging the gap between training and inference resolution.

4.3. Training dataset

We construct our training data from the Stereo4D [13] dataset, which contains over 100k high-resolution, rectified stereo videos capturing diverse scenes and moving objects. As shown in Fig.3, this dataset provides real-world video data that naturally includes challenging cases such as reflective surfaces, which are difficult to simulate in synthetic datasets. Following Stereo4D, we project VR180 videos to rectified perspective projection videos of resolution 512×512. We filter out examples with excessively

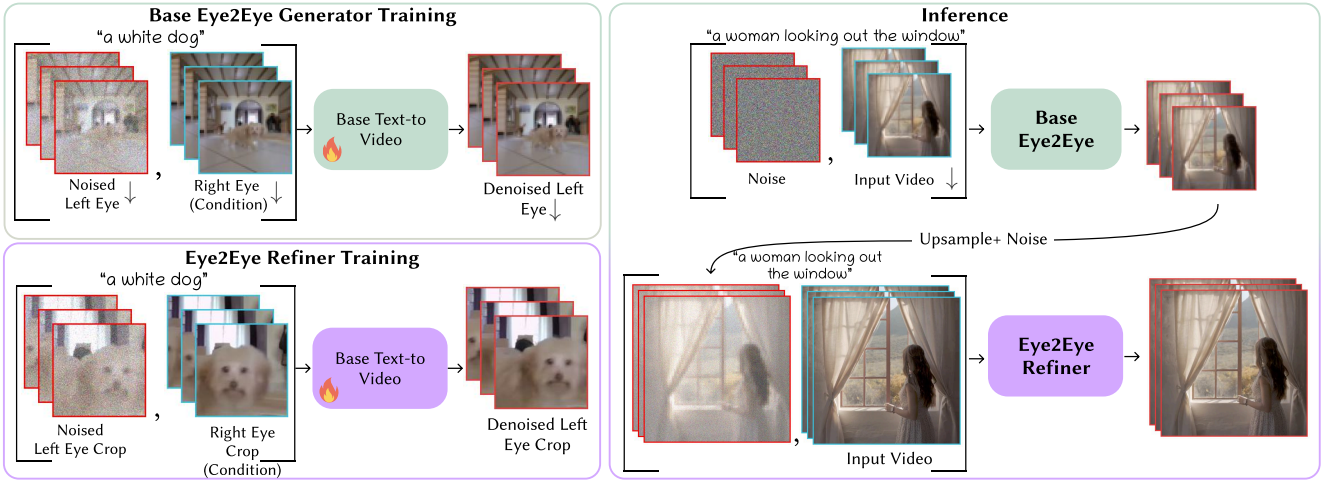

Figure 4. **Our mono-to-stereo pipeline.** We leverage the pre-trained Lumiere cascaded text-to-video model, as well as a curated rectified stereo pairs dataset, to perform mono-to-stereo synthesis. We fine tune two different copies of a base (low-resolution) pre-trained Lumiere model, in two different contexts. For the first base model, we add additional input channels to condition the model on an input right eye, and train the base Eye2Eye generator on downsampled, low-resolution 128×128 stereo pairs (top left). We call the resulting trained model the **base Eye2Eye generator** model. We train the second model to be a refinement model with the same conditioning mechanism, but instead trained on 128×128 crops from the full, high-resolution images (bottom left). We call the resulting model the **Eye2Eye refiner** model. The base Eye2Eye model models the correct pixels disparity at a low resolution, and the Eye2Eye refiner has better inpainting capabilities and can model the larger pixel disparities corresponding to high-resolution imagery. At inference time, our sampling process (right) combines both models' strengths by first generating a low-resolution output from the base Eye2Eye model to establish appropriate stereo disparity for a compelling 3D effect, then noising and denoising it with the Eye2Eye refiner to achieve high visual quality.

large disparities caused by objects being too close to the camera, as these often lead to stereo window violations [46] and are challenging for the model to learn. Specifically, we compute optical flow between the left and right frames with RAFT [32, 34] to estimate pixel disparities and discard videos where the disparity exceeds a specified threshold (60 pixels). Additionally, we use BLIP2 [18] to generate captions with the middle frame of each video. During training, we sample 80 frames per clip to align with Lumiere's input video length.

5. Results

5.1. Baselines

The most prominent baselines to our approach are (1) SVG [7], a training-free method that uses a depth model to estimate video disparity, warp the frames, and devise a specific sampling scheme to perform consistent inpainting; (2) Stereocrafter citestereocrafter, where a simliar warp-and-inpaint approach is adopted, but an inpainting model is trained specifically for the inpainting of left-right dis-occlusion. Since both methods do not have open source weights or implementations, we implement (2) and train an inpainting model on warped views. We use Lumiere as the pre-trained model, fine-tuning the LR model and then using the Lumiere-super resolution stage with a blended diffusion [1] approach, to ensure that the details of the original warped videos are preserved. See appendix for more details. We additionally include a qualitative comparison with Deep3D [42], a deep CNN trained for mono-to-stereo prediction; and (2) Dynamic Gaussian Marbles (DGM) [31], a method for novel view synthesis of casual monocular videos.

5.2. Evaluation data

We assess our method on a dataset of 30 publicly sourced videos encompassing diverse scenes, camera motions, and dynamic content. These videos feature complex lighting conditions and varied materials, including specular surfaces that introduce challenges such as reflections. Some of the videos are taken from the data provided in [44], which proposed a method to decompose the different layers of reflected and refracted light. See a sample of the evaluation videos in Figure 7.

5.3. Qualitative comparisons

Figure 6 shows qualitative comparisons to the baselines. As can be seen, the warp-and-inpaint baseline (column b) shifts content incorrectly in areas with significant reflections or refractions. For instance, it fails to correctly distinguish the objects in the scene when there is a reflection overlaying them , or a transparent object is in front of them. This can be observed in the top example, where a transparent umbrella is partially occluding a building. The single layer disparity

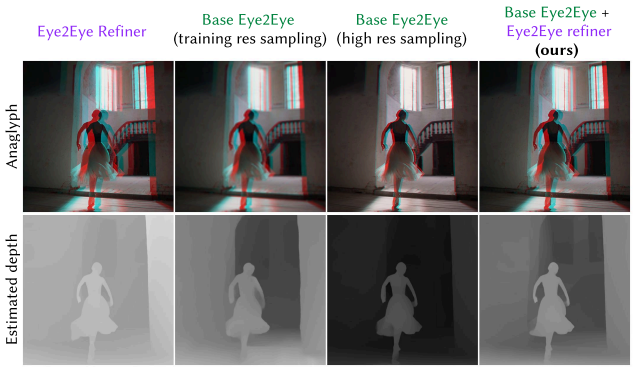

Figure 5. **Resolving training and inference gap.** We ablate the use of the two models in our pipeline, illustrating the training-inference gap of each of them. We visualize the resulting anaglyph and depth estimation (estimated using [34]) of their outputs. When sampling from the Eye2Eye-refiner model, that was trained on crops without any downsampling, far away content is still largely shifted (column 1). When sampling from the base Eye2Eye generator at a larger resolution than its training resolution, the scale of the disparity and novel content in the frame reduces, weakening the 3D effect compared to sampling at the training resolution (columns 2 and 3, in column 2 the outputs were upsampled). By upsampling the outputs of the base stereo model and noising and denoising it with the Eye2Eye-refiner model, we maintain both a good depth perception from the base model and the stereo refiner's ability to generate high quality frames (column 4).

model predicts a larger disparity for the areas of the umbrella, and thus the top part of the building is largely shifted while the bottom is not. This results in a distorted frame and in a wrong 3D effect when viewing in anaglyph glasses; the top part of the building appears as near as the umbrella. Or, the warp-in-inpaint can incorrectly shift, incorrectly shifts the reflected distant content by a large disparity: in the bottom example, the reflected distant pole is shifted largely, along with the woman's head. Our output (column a), on the other hand, manages to shift each of the different layers in the video, according to their own disparities: In the first example, the transparent umbrella is correctly shifted by a larger amount than the building behind it. In the bottom example, the reflected pole is almost not shifted, while the woman's head shifts by a large amount.

Dynamic Gaussian Marbles (DGM, column c) does not leverage a generative prior, but only uses the existing information in the video for the reconstruction. Thus, it can not unpaint missing content that does not exist in a different frame. This leads to white holes in the video frames in cases where the original camera motion in the video is small (see top example, along the borders of the people, and bottom example along the left edge of the frame). Additionally, as their pipeline uses single layer metric depth estimation for geometry regularization, its outputs suffer from similar distortions as the warp and inpaint baseline in the top example, and it fails to correctly model the geometry in the bottom example, where the metric depth predicts a flat depth map for the window. Deep3D (column d) fails to produce a sufficient 3D effect – the output videos are almost identical to the input ones in most examples.

5.4. Quantitative comparisons

5.4.1. User study.

We conducted a user study to evaluate our video-to-stereo model using a Two-alternative Forced Choice (2AFC) protocol [15, 27]. Participants wore a VR headset and viewed two videos side-by-side: our model's output and the baseline's output. Specifically, our model's and the *warp and inpaint* left view predictions were projected onto the participants' left eyes, while the input right video was projected onto their right eyes.

Prior to the main comparison, participants were shown a ground truth stereo pair featuring a large reflection alongside a warp-and-inpaint result that does not account for reflection. This preliminary step ensured that participants understood the task and excluded those with binocular vision dysfunctions (see the test examples in the supplemental material). During the main task, participants were asked to determine which video exhibited a more realistic 3D effect, included in areas with reflections or transparent surfaces. Overall, participants favored our videos 66% of the time based on 239 judgments.

5.4.2. iSQoE stereo perception metric

We evaluated the recently proposed stereo perception metric, iSQoE [33], which trained a model designed to assess the stereoscopic quality of experience (SQoE) by aligning closely with human perceptual preferences. The authors have demonstrated that iSQoE effectively evaluates different mono-to-stereo conversion techniques. This is an image metric, and it is meaningful only when comparing the same stereo pair generated through different processing or conversion methods. Thus, to obtain per-video preferences, we average the iSQoE scores across frames and compare the mean scores between methods. Our approach achieved higher average scores on 74% of the videos in our test set, supporting our approach's superior performance.

6. Discussion

We presented a simple approach for video mono-to-stereo conversion, highlighting complexities that were often overlooked in prior and concurrent work. As video models continue to grow in size and training data, they not only produce higher-quality outputs but also appear to implicitly approximate certain aspects of our physical world; our results highlight these emerging capabilities. Our user-study suggests

Figure 6. **Qualitative Comparison.** Our method successfully generates left from right views in complex scenarios where light is both reflected on a transparent material and refracted through it. The warp-and-inpaint baseline, relying on a single-layer disparity prediction, fails in such cases. For instance, in the top example, the top of the building appears as near as the transparent umbrella overlaying it (see anaglyph), and the building is distorted (see output left eye). Our method, in contrast, successfully shifts the umbrella without shifting the building behind it. In the bottom example, the pole reflected on the glass appears as near as the woman behind the window (see anaglyph); in our result, the pole is almost not shifted, as it is far away. Dynamic Gaussian Marbles (DGM, c), a 4D reconstruction method, lacks generative capabilities. Thus, their output left eye has white regions of missing content (see top example along the borders of the people, and in the bottom example along the left edge of the frame). In addition, since DGM relies on metric depth estimation as a regularization, it often fails to correctly model the geometry in complex scenarios—producing distortions similar to those of the warp-and-inpaint baseline in the top example, and a "flat" output in the bottom example. Finally, Deep3D (d) generally fails to generate a sufficient 3D effect, as seen in the anaglyph visualizations.

handling reflections in modern VR headsets would help increase the realism of immersive experiences, encouraging VR development and research to consider these nuances. Nonetheless, an inherent limitation of our current approach is that we do not control the baseline between the cameras, constraining the extent of the 3D effect. Future work could explore methods for dynamically adjusting this baseline, thereby offering more flexibility in creating immersive stereo content.

7. Acknowledgments

We thank Shir Amir for her valuable assistance in running the iSQoE model for our evaluation.

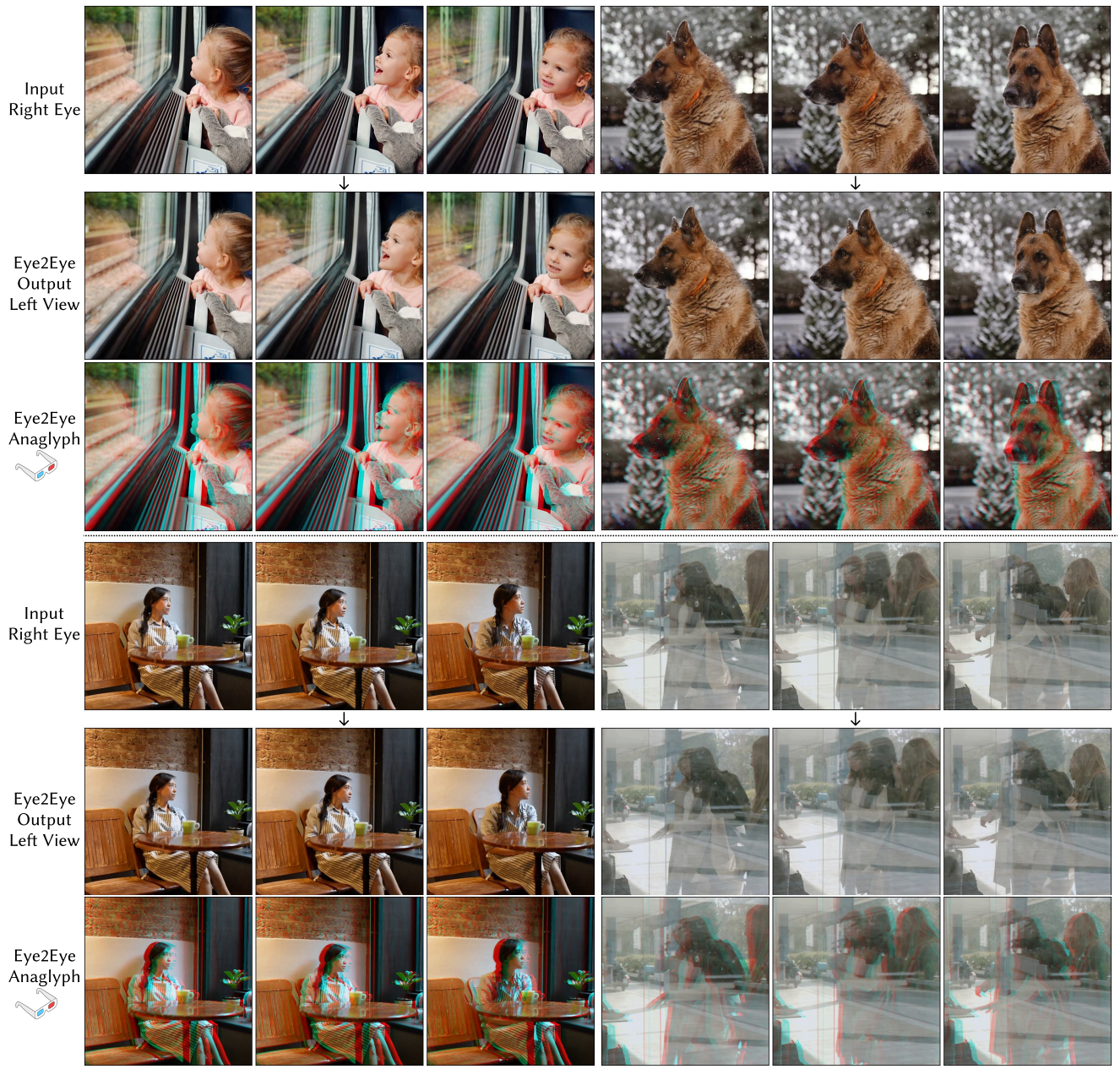

Figure 7. **Results.** A sample of resulting stereo views using our method. Our approach is particularly successful in complex scenarios involving reflective objects such as glass doors or specular tables, where traditional methods often produce distortions. See videos in the supplementary material

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 18187–18197. IEEE, 2022. 5

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, 2024. 2, 4

[3] Sai Bi, Zexiang Xu, Pratul P. Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Milos Hasan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv*, 2020. 3

[4] Andrew Blake and Heinrich Bülthoff. Does the brain know

the physics of specular reflection? *Nature*, 1990. 2

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable Video Diffusion: Scaling latent video diffusion models to large datasets, 2023. 2

[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. *arXiv*, 2024. 2, 4

[7] 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, 2024. 2, 5, 10

[8] Cheng-De Fan, Chen-Wei Chang, Yi-Ruei Liu, Jie-Ying Lee, Jiun-Long Huang, Yu-Chee Tseng, and Yu-Lun Liu. Spectromotion: Dynamic 3d reconstruction of specular scenes. *arXiv*, 2024. 3

[9] Ruiqi Gao*, Aleksander Holynski*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole*. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv*, 2024. 2

[10] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models, 2023. 4

[11] 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. 3, 11

[12] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. TensoIR: Tensorial Inverse Rendering. *CVPR*, 2023. 3

[13] Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Holynski. Stereo4d: Learning how things move in 3d from internet stereo videos. *arXiv preprint*, 2024. 3, 4

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2

[15] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019. 6

[16] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation, 2024. 4

[17] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon. Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. In *arXiv*, 2024. 2

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 2023. 3, 5

[19] Ruofan Liang, Huiting Chen, Chunlin Li, Fan Chen, Selvakumar Panneer, and Nandita Vijaykumar. ENVIDR: Implicit Differentiable Renderer with Neural Environment Lighting. *ICCV*, 2023. 3

[20] Li Ma, Vasu Agrawal, Haithem Turki, Changil Kim, Chen Gao, Pedro Sander, Michael Zollhöfer, and Christian Richardt. Specnerf: Gaussian directional encoding for specular reflections. *arXiv 2312.13102*, 2023. 3

[21] Alexander Mai, Dor Verbin, Falko Kuester, and Sara Fridovich-Keil. Neural microfacet fields for inverse rendering. *ICCV*, 2023. 3

[22] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *CoRR*, abs/2108.01073, 2021. 4

[23] 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. In *ECCV*, 2020. 2

[24] Martin Mišiak, Arnulph Fuhrmann, and Marc Erich Latoschik. The impact of reflection approximations on visual quality in virtual reality. In *ACM Symposium on Applied Perception 2023*, New York, NY, USA, 2023. Association for Computing Machinery. 2

[25] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 11

[26] Masakazu Ohara, Juno Kim, and Kowa Koida. The role of specular reflections and illumination in the perception of thickness in solid transparent objects. *Frontiers in Psychology*, 13, 2022. 2

[27] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A. Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In *Advances in Neural Information Processing Systems*, 2020. 6

[28] Abhishek Sharma, Adams Yu, Ali Razavi, Andeep Toor, Andrew Pierson, Ankush Gupta, Austin Waters, Aäron van den Oord, Daniel Tanis, Dumitru Erhan, Eric Lau, Eleni Shaw, Gabe Barth-Maron, Greg Shaw, Han Zhang, Henna Nandwani, Hernan Moraldo, Hyunjik Kim, Irina Blok, Jakob Bauer, Jeff Donahue, Junyoung Chung, Kory Mathewson, Kurtis David, Lasse Espeholt, Marc van Zee, Matt McGill, Medhini Narasimhan, Miaosen Wang, Mikołaj Bińkowski, Mohammad Babaeizadeh, Mohammad Taghi Saffar, Nando de Freitas, Nick Pezzotti, Pieter-Jan Kindermans, Poorva Rane, Rachel Hornung, Robert Riachi, Ruben Villegas, Rui Qian, Sander Dieleman, Serena Zhang, Serkan Cabi, Shixin Luo, Shlomi Fruchter, Signe Nørly, Srivatsan Srinivasan, Tobias Pfaff, Tom Hume, Vikas Verma, Weizhe Hua, William

Zhu, Xinchen Yan, Xinyu Wang, Yelin Kim, Yuqing Du, and Yutian Chen. Veo. *arXiv*, 2024. 4

[29] Sudipta N. Sinha, Johannes Kopf, Michael Goesele, Daniel Scharstein, and Richard Szeliski. Image-based rendering for scenes with reflections. *ACM Transactions on Graphics (TOG)*, 31:1 – 10, 2012. 2

[30] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. NeRV: Neural reflectance and visibility fields for relighting and view synthesis. *CVPR*, 2021. 3

[31] Colton Stearns, Adam Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In *SIGGRAPH Asia 2024 Conference Papers*, page 1–11. ACM, 2024. 5

[32] Deqing Sun, Charles Herrmann, Fitsum Reda, Michael Rubinstein, David J Fleet, and William T Freeman. Disentangling architecture and training for optical flow. In *ECCV*, 2022. 5

[33] Netanel Y. Tamir, Shir Amir, Ranel Itzhaky, Noam Atia, Shobhita Sundaram, Stephanie Fu, Ron Sokolovsky, Phillip Isola, Tali Dekel, Richard Zhang, and Miriam Farber. What makes for a good stereoscopic image?, 2024. 6

[34] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 3, 5, 6, 10, 11

[35] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *CVPR*, 2022. 3

[36] Dor Verbin, Pratul P. Srinivasan, Peter Hedman, Ben Mildenhall, Benjamin Attal, Richard Szeliski, and Jonathan T. Barron. Nerf-casting: Improved view-dependent appearance with consistent reflections, 2024. 3

[37] Fangjinhua Wang, Marie-Julie Rakotosaona, Michael Niemeyer, Richard Szeliski, Marc Pollefeys, and Federico Tombari. UniSDF: Unifying Neural Representations for High-Fidelity 3D Reconstruction of Complex Scenes with Reflections. *arXiv:2312.13285*, 2023. 3

[38] Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J. Fleet. Controlling space and time with diffusion models, 2024. 2

[39] Hongyu Wen, Erich Liang, and Jia Deng. Layeredflow: A real-world benchmark for non-lambertian multi-layer optical flow. *arXiv preprint arXiv:2409.05688*, 2024. 2

[40] Gunnar Wendt, Franz Faul, and Rainer Mausfeld. Highlight disparity contributes to the authenticity and strength of perceived glossiness. *Journal of Vision*, 8(1):14–14, 2008. 2

[41] Gunnar Wendt, Franz Faul, Vebjørn Ekroll, and Rainer Mausfeld. Disparity, motion, and color information improve gloss constancy performance. *Journal of Vision*, 10(9):7–7, 2010. 2

[42] Junyuan Xie, Ross B. Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*, 2016. 2, 5

[43] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency, 2024. 2

[44] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T. Freeman. A computational approach for obstruction-free photography. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 34(4), 2015. 5

[45] Sijie Zhao, Wenbo Hu, Xiaodong Cun, Yong Zhang, Xiaoyu Li, Zhe Kong, Xiangjun Gao, Muyao Niu, and Ying Shan. Stereocrafter: Diffusion-based generation of long and high-fidelity stereoscopic 3d from monocular videos, 2024. 2, 10

[46] Frederik Zilly, Marcus Müller, Peter Eisert, and Peter Kauff. The stereoscopic analyzer—an image-based assistance tool for stereo shooting and 3d production. In *2010 IEEE International Conference on Image Processing*, pages 4029–4032. IEEE, 2010. 5

8. Additional Details.

8.1. Training details.

We fine-tune Lumiere on a dataset of 100K clips from Stereo4D as mentioned in section 4.3 of the main paper. We temporally subsample the videos into 80 frames at 16 fps to match Lumiere's pre-training temporal resolution. We train the model for 120K steps with batch size 32 and learning rate $2e^{-5}$. The original clips resolution is 512x512 pixels. To train the Eye2Eye base model, we additionally downsample the frames spatially to 128x128 pixels. For the Eye2Eye refiner, we randomly sample crops of 128 pixels.

8.2. Sampling hyper-parameters for our method.

8.2.1. Base Eye2Eye sampling.

We sample with 50 diffusion timesteps and without classifier free guidance. We sample from this model at a resolution of 256 pixels, as we found that this resolution best mitigates visual quality and 3D effect.

8.2.2. Eye2Eye refiner.

We upsample the output of the base Eye2Eye model 512x512 pixels resolution and noise it to diffusion timestep $t = 0.9$. We then denoise it with 48 diffusion timesteps and without classifier free guidance

9. Baselines.

9.1. Warp-and-inpaint implementation.

As warp-and-inpaint baselines [7, 45] do not have open source implementations, we implement and train this baseline. We use the same dataset described in 4.3 to fine tune the base Lumiere inpainting model to inpaint left-right disocclusion masks. We use [34] to estimate disparity of each pair of stereo frames, V_{left}, V_{right} and obtain the disocclusion mask by computing left-right consistency of the disparity prediciton. At training, the model is conditioned on the right video warped according to the estimated disparity, V_{right}^{warped}, and the corresponding disocclusion mask M,

to denoise the left frame, with the standard diffusion objective:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, V_{right}^{warped}, M)\|_2^2\right] \quad (2)$$

Here $x_t = V_{left} \cdot \sqrt{\alpha_t} + \sqrt{1-\alpha_t}\epsilon$, and $\epsilon \sim \mathcal{N}(0, I)$. At inference time, given a video V, we use SOTA monocular disparity estimation [11] to estimate video disparity D^V. As this estimation is scale and shift invariant, we fit a scale and shift parameter to the disparity map to align it with the disparity of our outputs (we first estimate the disparity of our outputs using [34]). We then forward-warp the frames using depth ordered softmax splatting [25]. The inpainting mask here are the pixel locations that were not mapped onto by D^V. We open and dilate the mask to reduce temporal inconsistencies before feeding it to the base inpainitng model. For spatial super resolution, we use the pretrained Lumiere SSR model and take a blended diffusion approach for maintaining faithfulness to the original video. Specifically, given the low resolution base model video output $V_{128\times128}^{inpainted}$, the warped right video $V^{\text{warped}} = \text{softmax\_z\_splatting}(V, D^v)$, we blend the predicted clean super-resolved output $\tilde{x_0}_t(x_{t+1}, t+1, V_{128\times128}^{inpainted})$ with V^{warped}: and at each diffusion step t $x_{0t} = M \cdot \tilde{x_0}_t(x_{t+1}, t+1, V_{128\times128}^{inpainted}) + (1-M) \cdot V^{warped}$. We use a the standard lumiere sampling of 256 and 32 diffusion timesteps for the base model and the SSR model, respectively, and a classifier free guidance of 8.

9.2. Deep3D.

As the original paper implementation uses a deprecated codebase, we turn to a more recent implementation found in the link: https://github.com/HypoX64/Deep3D. Their training data consists of 3D movies, which are typically processed in a differnt manner then our data - the zero disparity plane is usually shited to increase human comfort, making the RGB comparison difficult. We thus encourage the viewer to use anaglyph glasses for these results.

9.3. Dynamic Gaussian marbles.

We optimize the Dynamic Gaussian Marbles using the official paper implementation https://github.com/coltonstearns/dynamic-gaussian-marbles, using their default real-world videos configuration. We observed the optimizing the representation for the full number of steps (100K) in this configuration diverges, and thus synthesize stereo views from it after 40K steps.