# IMMERSEPRO: END-TO-END STEREO VIDEO SYN THESIS VIA IMPLICIT DISPARITY LEARNING

Anonymous authors

Paper under double-blind review

### ABSTRACT

We introduce *ImmersePro*, an innovative framework specifically designed to transform single-view videos into stereo videos. This framework utilizes a novel dualbranch architecture comprising a disparity branch and a context branch on video data by leveraging spatial-temporal attention mechanisms. *ImmersePro* employs implicit disparity guidance, enabling the generation of stereo pairs from video sequences without the need for explicit disparity maps, thus reducing potential errors associated with disparity estimation models. In addition to the technical advancements, we introduce the YouTube-SBS dataset, a comprehensive collection of 423 stereo videos sourced from YouTube. This dataset is unprecedented in its scale, featuring over 7 million stereo pairs, and is designed to facilitate training and benchmarking of stereo video generation models. Our experiments demonstrate the effectiveness of *ImmersePro* in producing high-quality stereo videos, offering significant improvements over existing methods. Compared to the best competitor stereo-from-mono we quantitatively improve the results by 11.76% (L1), 6.39% (SSIM), and 5.10% (PSNR).

025 026

027

004

010 011

012

013

014

015

016

017

018

019

021

023

### 1 INTRODUCTION

028 A stereo movie, also known as a 3D movie, provides three-dimensional visual effects by employing 029 stereoscopic techniques. By capturing or creating separate views for the left and right eyes, a 3D immersive experience can be achieved by using dedicated hardware such as head-mounted displays 031 or autostereoscopic displays. The disparity between the two views perceived by the viewer's brain 032 creates the illusion of depth, making the objects in the movie appear at varying distances, thereby 033 enhancing the immersive experience of the film. Shooting stereo movies in the film industry often 034 involves high costs due to the need for specialized equipment and meticulous post-production processes. Alternatively, the stereoscopic effect can be created through a post-production process for 035 videos that are shot with monocular cameras. This post-production process uses stereo conversion, 036 which adds the binocular disparity depth cue to digital images. It requires significant manual work 037 by artists since inaccurate depth mapping and misrepresentations of occluded areas can cause visual discomfort Devernay & Beardsley (2010). In this paper, we propose an automated system that can reduce the time and expense associated with the conversion process, making it more accessible and 040 economically feasible for more films. 041

Traditional stereo conversion involves creating disparity maps from single images or sequences and 042 then using them to generate the corresponding stereo pair for the other eye, creating the illusion 043 of depth for stereoscopic viewing. Recently, many deep learning-based methods (Xie et al., 2016; 044 Wang et al., 2019a; Shih et al., 2020; Watson et al., 2020; Ranftl et al., 2022) are primarily pro-045 posed for image-based stereo conversions, aiming to improve disparities and enhance inpainting 046 effectiveness on occluded areas. Unlike image data, video data provides additional temporal infor-047 mation, which can yield more detailed disparities and occlusion insights by leveraging information 048 across frames. To handle video inputs, Chen et al. (2019) synthesizes right-view video sequences by estimating a displacement map to move each pixel to a new location, with a 3D DenseNet. Temporal3D (Zhang & Wang, 2022) compromises to use three adjacent left-view frames to predict the 051 single right-view of the middle frame. Based on our analysis, current stereo conversion frameworks for video sequences are not robust and have several drawbacks. We believe the area is underex-052 plored and there is a large room for improvement. At the same time, we believe the topic will gain in importance due to recent efforts to manufacture stereo displays, e.g., from Apple and Magic Leap.



Figure 1: ImmersePro is a video method to convert a single-view video to a stereo video by predicting plausible right-view images for each input frame. Compared to previous work processing images frame by frame (3D Photo or Stereo from Mono), our method has the best visual quality.

We introduce *ImmersePro*, a novel approach designed specifically for video stereo conversion that utilizes the contextual information available across video frames to enhance stereo disparity consistency across the temporal dimension. For doing so, we collectively build a large-scale stereo movie dataset, Youtube-SBS, with over 7 million stereo pairs from a collection of stereo movies, game films, and music videos. Due to the absence of ground truth disparities, we propose to use implicit disparities to guide the generation of *layered disparities*, which outperforms the explicit disparity guidance (e.g. a depth estimation model) that was commonly used in previous work. We propose to use a *layered disparity* representation that refers to a stack of disparity maps corresponding to one image. Each pixel that appears in the image can be reused multiple times, avoiding creating black holes after the warping operation. This approach ensures that the generated stereo parts strictly adhere to the semantics of the input video, minimizing the need for improvisation and thus preserving the original narrative and visual intent. As a result, *ImmersePro* not only maintains the semantic integrity of the original video but also intelligently infers the geometry of occluded areas, enabling consistent right-view generation. As shown in Figure 1, previous methods may generate artifacts such as texture misalignment or object deformation, whereas our ImmersePro can keep the semantic integrity from the left-view image. Our main contributions are as follows:

• We introduce the *YouTube-SBS* dataset, an extensive collection of stereo videos sourced from YouTube, featuring over 7 million stereo pairs. This dataset fills the gap to serve as a benchmark for training and evaluating stereo video generation models.

• We introduce *ImmersePro*, specifically tailored for converting single-view videos into stereo videos using *layered disparity* warping via implicit disparity guidance. Compared to the best competitor stereo-from-mono we quantitatively improve the results by 11.76% (L1), 6.39% (SSIM), and 5.10% (PSNR).

# 108 2 BACKGROUND

110

We discuss previous stereo conversion methods and stereo datasets in this section.

111 112

115

# 113 114 2.1 STEREO CONVERSION METHODS

**Image-Based Stereo Conversion.** Deep3D (Xie et al., 2016) relaxes the disparity map into a multi-116 layer probabilistic map and then multiplies it with several horizontally shifted copies of the input 117 image, which relaxes the non-differentiable warping operation. Watson et al. (2020) used a warping-118 and-inpainting framework, which creates stereo training pairs from single RGB images to improve 119 the modern monocular depth estimators. However, a non-differentiable strategy is used and the 120 inpainting randomly selects the texture from the training set. Apart from using pretrained depth 121 estimation models, Wang et al. (2019a); Ranftl et al. (2022) use FlowNet2.0 (Ilg et al., 2017) to 122 estimate optical flows as ground truth disparities. StereoDiffusion (Wang et al., 2024) proposes a 123 training-free approach to generate stereo pairs by directly warping the latent space of diffusion mod-124 els. It requires inversion methods to produce the latents to generate the stereo pair of a given image. The fine details of the resulting photo may vary due to the direct modification of the latent space. 125 Shih et al. (2020) proposed a layered depth inpainting method that generates a 3D representation by 126 intelligently estimating and filling depth information, particularly in areas where it is missing or un-127 certain. Our work does not rely on explicit disparity computation, with the additional consideration 128 of the context within video frames. 129

130 Video-Based Stereo Conversion. Chen et al. (2019) adopts a reconstruction-based approach by 131 using a 3D DenseNet to estimate the disparity map of an input sequence. Temporal3D (Zhang & Wang, 2022) estimates the middle frame using three adjacent frames, with the output being a 132 weighted sum of three disparity-warped images. Additionally, methods such as NVDS (Wang et al., 133 2023) may be adopted for consistent depth estimations across video frames. However, those methods 134 assume the pixels within the left image are adequate for the right image. Mehl et al. (2024) adopted 135 the warping-inpainting approach with a pretrained depth estimation method (*i.e.* MiDaS (Birkl et al., 136 2023)) for warping and inpainting with multiple adjacent frames. Still, this method relies on a single 137 frame depth estimation model that can likely break the temporal consistency between frames. In 138 this work, we propose an end-to-end video stereo conversion method based on implicit disparity 139 guidance across the temporal dimension.

140 141 142

143

## 2.2 STEREO DATASETS

144 There are limited resources on video-based stereo datasets. Sintel (Butler et al., 2012) contains 1064 145 synthetic stereo images with accurate disparities. KITTI (Menze & Geiger, 2015) offers 8.4K frames captured from the real world for autonomous driving. Wang et al. (2019a) introduces a WSVD dataset 146 and proposes to use optical flow as disparities as ground truth for supervision. Similarly, Ranftl 147 et al. (2022) collected a private 3D movie dataset and extracted ground truth disparities by estimated 148 optical flows to improve depth estimation. Since different levels of stereoscopic effects may exist 149 for different purposes of a dataset, a movie-specific benchmark dataset is preferable. Ranftl et al. 150 (2022) is the only relevant dataset but it is built on top of real movies with intellectual property right 151 issues. Therefore, we propose a benchmark stereo dataset that contains publicly available content. 152

Dataset	content	GT depth	available	No. frames
KITTI (Menze & Geiger, 2015)	autonomous driving	metric	Y	8.4K
WSVD (Wang et al., 2019a)	mixed	NA	Y	1.5M
3D Movies (Ranftl et al., 2022)	movies	NA	Ν	75K
Sintel (Butler et al., 2012)	synthetic	metric	Y	1064
Youtube-SBS	movies	NA	Y	7M

159 160 161

153 154

Table 1: Relevant datasets.

#### 162 YOUTUBE-SBS 3 163

164 We aim to set up a large-scale publicly accessible benchmark dataset. The direct collection of 165 3D movies often encounters legal challenges to publish as an open-source dataset. Therefore, we 166 present Youtube-SBS, an open-source dataset collected from YouTube. This dataset contains over 167 400 3D side-by-side (SBS) videos. With a particular interest in stereo movies, our dataset primarily consists of movie trailers, game films, and music videos. We explicitly excluded 360-degree virtual 168 reality videos and gameplay videos (that contain user interfaces). To ensure accessibility for future research, we select videos that (1) have existed for at least one year, and (2) from accounts that 170 have at least 500 followers. This curated selection includes 423 videos at a standard resolution of 171 1920x1080. During the frame extraction, as some videos include a non-stereo intro section, we skip 172 the first 600 frames to capture valid stereo pairs. 173

To measure the general stereo effects of our dataset, we propose to compute a metric that evaluates 174 the left-right consistency of the disparity. For a stereo pair with subtle stereo effects, the disparity 175 maps for the left and right images should be almost symmetrical with one another. That is, a point 176 in the left image should have a corresponding point in the right image at the same row but shifted 177 horizontally according to the disparity. For large stereo effects there is an increasing number of 178 occluded and disoccluded areas. In these regions, the right image can no longer be reconstructed 179 from the left image with simple warping (and vice versa). To compute our metric, we use the 180 optical flow method RAFT (Teed & Deng, 2020). We also evaluated STTR (Li et al., 2021) and 181 RAFT-Stereo (Lipson et al., 2021), but these two methods produced worse results. Note that high 182 consistency means that the left-to-right optical flow  $F_{l \to r}$  and right-to-left optical flow  $F_{r \to l}$  are the 183 negative of each other. We calculate the consistency  $\varepsilon$  as follows: 184

$$\mathcal{E}_p = ||F_{l \to r}(p)) + F_{r \to l}(p + F_{l \to r}(p))||, \tag{1}$$

where p is the pixel position of a frame. We provide a breakdown to demonstrate consistency metric in Table 2 to present the general stereo effects of the dataset. We compute occluded areas with  $\sum_{p} 1(\mathcal{E}_p > \epsilon)$ . We use  $\epsilon = 4$  for improved stability on RAFT-computed optical flows. We present a visual demonstration of different levels of stereo effects in Figure 6.

occluded area	< 10%	< 20%	< 30%	< 40%
Percentage	71.27%	84.60%	91.30%	94.71%

Table 2: Flow-based consistency check results. Most frames present subtle stereo effects in the dataset.

#### 4 METHOD

199 A stereo video sequence  $I = \{I^l, I^r\}$  contains left and right video sequences of  $I^l \in R^{T \times H \times W \times 3}$ and  $I^r \in R^{T \times H \times W \times 3}$ , respectively. We use T, H, W to denote the video sequence length, video 200 201 height, and video width, respectively. We aim to predict a right video sequence  $I^r$  based on the input 202 left video sequence  $I^l$  to make  $\hat{I} = \{I^l, \hat{I}^r\}$  presents similar stereo effects as I. 203

As shown in Figure 2, our method compromises six stages. First, we use a dual branch architec-204 ture (section 4.1) that consists of a disparity branch and a context branch, to extract disparity and 205 semantic features, respectively. Second, we apply spatial-temporal self-attention (section 4.2) on 206 each scale feature to achieve multi-frame awareness. Third, we fuse the multi-scale features to ob-207 tain implicit disparity features (section 4.3). Fourth, we then use a spatial-temporal cross-attention 208 module (section 4.2) to inject contextual information into the implicit disparity features to obtain 209 layered disparity features (section 4.4). Fifth, right-view video sequences can be estimated by warp-210 ing through layered disparities. Finally, we enrich the estimated right-view sequences with a context 211 fusion module.

212

214

185

187

188

194

195 196 197

198

#### 213 4.1 DUAL BRANCH ARCHITECTURE

We use a dual-branch architecture to enhance stereo video conversion by separately processing 215 disparity and contextual information, as shown in Figure 2. We employ a pretrained DepthAny-



Figure 2: Illustration of ImmersePro framework. Our network contains six parts: (1) dual-branch feature extractors for extracting disparity features and context features (section 4.1), (2) multi-scale spatial-temporal self-attention to refine disparity features (section 4.2), (3) implicit disparity to generate stereo images without explicit disparities (section 4.3), (4) spatial-temporal cross attention block to inject contextual information into the implicit disparity features (section 4.2), (5) layered disparity to obtain the estimated right view video sequences (section 4.4), and (6) context fusion to enrich the estimated right view video sequences with detailed semantic information (section 4.5).

237 238 239

240

241

242

217 218

224

225

230

231

232

233

234

235

236

thing (Yang et al., 2024) model for the disparity branch to extract disparity-oriented feature maps, while a context feature extractor with the same architecture from Zhou et al. (2023); Li et al. (2022)'s encoder is used to extract contextual semantic features.

The disparity branch operates on multiple scales, extracting features at 1/2 and 1/4 resolutions of 243 the original input to capture detailed disparity information. The disparity features are from the de-244 coder of the model<sup>1</sup>. This branch utilizes spatial-temporal self-attention modules (section 4.2) to 245 prioritize relevant spatial and temporal details on different scales, ensuring that the model focuses 246 on areas with significant disparity changes or movement. After combining the multi-scale features 247 into 1/2 resolution with a fusion block, we apply softmax to these features to create a probabil-248 ity distribution that represents the implicit disparities. The implicit disparity is used to select the 249 appropriate pixels from a stack of the multiple horizontally shifted copies of the input image (sec-250 tion 4.3). By encouraging accurate selection, these features implicitly represent the disparity for 251 stereo conversion.

Concurrently, we use a stack of convolution layers as the context encoder. We experimented with 253 multiple encoder architectures and settled on the architecture without aggressive downsampling. The 254 details for the context encoder are presented in Appendix A.2. The context branch focuses solely on 255 capturing texture information. This branch processes texture at 1/2 the original resolution, aligning 256 with the disparity branch's output. Finally, with spatial-temporal cross-attention modules to fuse the implicit disparity and texture information, we apply a layered disparity warping (section 4.4) to 257 obtain the final predicted right-view. 258

259 260

261

#### 4.2 SPATIAL-TEMPORAL ATTENTION

262 Video transformers have demonstrated excellent performances in video-based tasks such as video 263 segmentation (Duke et al., 2021), video-text feature mapping (Li et al., 2023), and video inpaint-264 ing (Li et al., 2022; Zhou et al., 2023). This work builds sparse video transformers on top of the ProPainter's version, considering the highly redundant and repetitive textures in adjacent frames. We 265 remove the mask guidance in the original model and use a temporal stride of 2 to avoid redundant 266 key/value tokens within each transformer block and to improve the computational efficiency. Aside 267

<sup>&</sup>lt;sup>1</sup>We use the output from the neck of the model, as implemented by https://github.com/ huggingface/transformers.



Figure 3: Visual demonstration of the implicit disparity guidance. We can observe that (1) the implicit disparity module tries to resolve the disparity from the given image, and (2) our method can significantly rectify the error introduced by the implicit disparity estimations. Our method offers a significant improvement regarding clarity with less irregular texture deformation on the image. The implicit disparity map contains multiple channels and we apply argmax to obtain the visual output.

from spatial-temporal self-attention, we also use spatial-temporal cross-attention to fuse features from different sources.

Given a video feature sequence  $E_s \in \mathbb{R}^{T_s \times H_s \times W_s \times C}$ , we first perform soft split (Liu et al., 2021) to generate patch embeddings  $Z \in \mathbb{R}^{T_s \times M \times N \times C_z}$ . Subsequently, Z is partitioned into  $m \times n$ non-overlapping windows, yielding the partitioned embedding features  $Z_w \in \mathbb{R}^{T_s \times m \times n \times h \times w \times C_z}$ , where  $m \times n$  denotes the number of windows and  $h \times w$  denotes their size. For self-attention transformer blocks, we obtain the query Q, key K, and value V from  $Z_w$  through three linear layers, respectively. For cross-attention transformer blocks, we repeat the above process to obtain embeddings  $Z_c \in \mathbb{R}^{T_s \times m \times n \times h \times w \times C_z}$  from another feature sequence  $E_c \in \mathbb{R}^{T_s \times H_s \times W_s \times C}$ . Note that  $Z_c$  shares the same shape with  $Z_w$ . Then Q is extracted from  $Z_w$  whilst K and V are extracted from  $Z_c$ . For both self-attention and cross-attention mechanisms, the final embedding features are gathered using soft composition Liu et al. (2021) for further processing.

#### 4.3 IMPLICIT DISPARITY

For stereo vision, different from common generative tasks, the generated right view requires a pre-cise match to the input view with as little improvisation as possible. The stereo pair of an image is commonly constructed by obtaining the disparity map to find the shifting distances of each pixel within the input view. Assuming  $d_{i,j}$  is the disparity value at pixel location (i, j) in the left image, the corresponding pixel in the right image is:

$$I_{i,j}^r = I_{i,j+d_{i,j}}^l.$$
 (2)

It is typically a non-differentiable operation due to its piecewise nature. Jaderberg et al. (2015) propose to use sub-gradients for backpropagation through spatial transformations to handle such non-smooth operations, enabling differentiable warping.



Figure 4: Visual demonstration of our layered disparity representation. We show the  $1^{st}$  and  $3^{rd}$  disparity layers in Figures 4b and 4c. We denote darker colors as moving to the right and lighter colors as moving to the left. We use 7 layers in total while we found the  $1^{st}$  and  $3^{rd}$  layers contribute to the right-view generation most. We observe that the  $1^{st}$  layer aims to warp the majority of the pixels to the right to their correct right-view location while the  $3^{rd}$  layer moves pixels to the left to fill the resulting holes, e.g. near the left border.

Xie et al. (2016) proposed another approach to use a depth selection layer to align the output right view to the source input view structure. Subsequent works such as Zhang et al. (2019) follow a similar idea. We employ it as auxiliary supervision. We found this method to be suitable for guidance only. Directly using it to compute the output leads to blurry results. *Implicit disparity* predicts a probability distribution across possible disparity values d at each pixel location.  $p_{i,j}^d$ , with  $\sum_d p_{i,j}^d = 1$ , denotes the probability of pixel (i, j) having disparity d. We denote an image that is shifted by d pixels horizontally as  $I_{i,j}^d = I_{i,j-d}$ . We then obtain the right-view pixel values as:

$$\hat{I}_{i,j}^{aux} \sum_{d} = I_{i,j}^{d} p_{i,j}^{d}.$$
(3)

where  $\hat{I}^{aux}$  is the auxiliary predicted right view. We use  $V_{i,j}^d = I_{i,j}^d D_{i,j}^d$  for subsequent computations. This approach estimates the stereo pair of a given image without an explicit disparity map, serving as a relaxation of the warping operation in Equation (2). Without implicit disparity, our model can hardly converge as shown in Section 5.1.

### ALGORITHM 1: Synthesis from layered disparities.

```
the number of disparity layers.
357
        number_layered_disparity:
       # warped_output: 'BDTCHW'. A stack of images warped by layered
358
       disparities. D is the number of disparity layers.
359
       # warped_mask: `BDTCHW`. A stack of masks warped by layered disparities.
360
       D is the number of disparity layers.
361
      layered_mask = zeros_like(output_mask)
362
       total_mask = zeros_like(output_mask)
       for i in range(number_layered_disparity):
         if i == 0:
364
             layered_mask[:, i] = warped_mask[:, i]
365
            total_mask[:, i] = warped_mask[:, i]
366
         else:
             total_mask[:, i] = logical_or(warped_mask[:, i], layered_mask[:, i -
367
             11)
368
             layered_mask[:, i] = torch.logical_and((1 - total_mask[:, i]),
369
             warped_mask[:, i - 1])
370
       output = layered_mask * warped_output
```

371 372

333

334

335

336

337 338

339 340

341

342

343

344

345

346

351

352

353

354 355

356

# 373 4.4 LAYERED DISPARITY374

The *implicit disparity* is a summation-based approach that computes pixel colors as a blend of other pixel colors, weighted by the estimated probabilities. This may produce good results with a correct estimation, but it may introduce artifacts such as blurring if the estimation is inaccurate. The final output visually improves if each pixel location is selected from a set of candidate disparity 378 layers, rather than blending all the layers. The proposed Layered Disparity uses a smaller stack of 379 candidate layers, and each layer represents disparity information. Therefore, our layered disparity 380 representation is a stack of disparity maps. We use a differentiable warping (Jaderberg et al., 2015) 381 operation to warp the input image to an output image. While a single disparity map already defines 382 a solution to the problem, there may be problems due to occlusion and disocclusion artifacts. These problems can then be fixed by other layers. Our approach avoids the mentioned blending problem. 383 Meanwhile, we maximize the reuse of pixel information within the image while at the same time 384 avoiding generating image holes. 385

386 We use implicit disparity  $V_{i,j}^d$  as a guidance to generate layered disparities. First, we employ three

*Conv-ReLU* blocks to refine the  $V_{i,j}^d$  to shrink them from  $\mathbb{R}^{T_s \times H \times W \times D}$  into  $\mathbb{R}^{T_s \times \frac{H}{2} \times \frac{W}{2} \times D}$ , where 387 388 D is the number of stacked disparities. We then apply the spatial-temporal cross-attention process, 389 as mentioned in Section 4.2. With the attention-applied features, a deconvolution operation and 390 three Conv-ReLU blocks are used to obtain the final layered disparity  $LD_{i,j}^d$ . Here, d = 7 since 391 we use 7 disparity layers in our work. We then apply the differentiable warping operation with 392 the layered disparity to obtain layered warped images  $\hat{I}_{i,j}$  and masks  $\hat{M}_{i,j}$ , respectively. We select 393 pixel values according to the layered masks as in Algorithm 1. As shown in Figure 3, our proposed 394 approach significantly improved the visual quality compared to the output from the implicit disparity layer. Figure 4 visualizes an example of learned disparity maps from the proposed layered disparity 395 representation. 396

4.5 CONTEXT FUSION 398

399 The final stage of our network focuses on enriching semantic details while maintaining the learned 400 right-view structure. The context fusion module integrates semantic and disparity features from a 401 video sequence by concatenating the encoder feature map with layered disparity features to form 402 a fused representation. These fused features are then processed through spatial-temporal attention 403 (section 4.2), enabling global context awareness. We apply spatial-temporal attention modules at 404 1/2 the original resolution, as mentioned in section 4.1. To retain structural integrity, a residual 405 connection reintroduces the refined transformer output into the original fused feature map. We then 406 apply a deconvolution to obtain a texture map in the original resolution, then enrich the texture map 407 by three *Conv-ReLU* blocks. Next, the module supplements the layered disparity-warped images from section 4.4 with the enriched feature map. To be specific, a median blur with  $3 \times 3$  kernels is first 408 applied to the warped images to reduce noise and improve local smoothness before concatenating 409 them with the enriched feature map. A semantic residual is then derived by passing this combined 410 map through three *Conv-ReLU* blocks. The final output is produced by combining the blurred image 411 with the semantic residual. This approach ensures that the final result maintains sharp textures while 412 preserving structural consistency, achieving a balance between local detail and global coherence. 413

- 5 RESULTS
- 415 416

414

397

We implement our method using Pytorch and train on four NVIDIA A100 (80G) GPUs for 50,000 417 iterations (approx. 3 days). Models are trained for 40,000 iterations for our ablations. At training 418 time, we first resize the input sequence to  $422 \times 422$  and then randomly crop the resized video 419 sequence to  $384 \times 384$ . Each input sequence contains 8 frames. We use L1 loss during training 420 to encourage an accurate reconstruction of the right-view images using both implicit and layered 421 disparities. In addition, an LPIPS (Zhang et al., 2018) loss is used for better reconstruction results. 422 An AdamW (Loshchilov & Hutter, 2017) optimizer is used. We use 3e-5 learning rate while image 423 losses are computed within the range of (-127.5, 127.5). We evaluated our method on our test set 424 which includes 43 video sequences with 558K frames.

425

426 427

5.1 COMPARISON WITH STATE-OF-ART MODELS

428 Benchmark methods. We compare our method with three state-of-the-art methods including Stereo-429 from-mono (Watson et al., 2020), 3D Photography (Shih et al., 2020), and StereoDiffusion (Wang et al., 2024). Note that those methods are designed for image-based stereo conversion purposes. 430 We are not aware of any open-source implementations for video stereo conversion. We use official 431 implementations for the selected methods.

	L1↓	SSIM $\uparrow$	PSNR $\uparrow$
Deep3D	0.2215	0.1935	11.9089
3D Photo	0.1069	0.3463	16.3658
Stereo Diffusion	0.0816	0.4651	18.6684
stereo-from-mono	0.0646	0.5685	20.7788
Ours w/o implicit disparity *	n/a	n/a	n/a
Ours w/o layered disparity	0.0885	0.4717	19.0523
Ours w/o attention blocks	0.0593	0.5894	21.4162
Ours w/o context fusion	0.0588	0.5959	21.6649
Ours	0.0570	0.6048	21.838

Table 3: Benchmark results. The best and second-best results are highlighted in green and yellow, respectively. \* indicates the model is not converged.

Benchmark settings. Due to the high runtime of those methods (especially for StereoDiffusion which is required to perform inversion (Mokady et al., 2023) for each image), we compare those methods with a subsampled dataset every 3 seconds (72 frames). At test time, we process 8 frames as input where the last 2 frames are taken as reference frames. We use widely employed L1, SSIM, and PSNR to evaluate the quality of the generated stereo pairs.

Benchmark results. Our qualitative and quantitative results are shown in Figure 1 and Table 3, re-452 spectively. The visual results show that other methods tend to generate right-view images with tex-453 ture deformations. To be specific, 3D photo struggles to find accurate depth cues with MiDaS (Birkl et al., 2023) depth estimation model, resulting in inaccurate warping on given images. Stereo-from-455 mono can generate images well but often comes with unpleasant black dots around the warping 456 shapes. StereoDiffusion requires using null-text inversion Mokady et al. (2023) to convert a given 457 image to the latent space and then warp the latent features to create the right-view image. It highly 458 depends on the performance of the inversion, which creates unstable performances. As shown in 459 our table, our method yields better numerical results. This finding aligns with the visual results. 460 In addition, our accompanying videos demonstrate better stability in terms of jittering and shaking. 461 Please watch the accompanying videos with 0.5 speed to see the artifacts generated by the different 462 methods.

463 Ablation results. Table 3 shows our ablation results. We show that our method is not going to con-464 verge without using implicit disparity guidance, while a significant performance drop may occur 465 when removing our proposed layered disparity. We show that our layered disparity generates better 466 visual quality in Figure 3 compared to the outputs from implicit disparities. Though not significant, 467 the attention blocks can slightly improve the overall performance, while the context fusion module contributes significantly. Additional experiments including alternative masking strategies, the inclu-468 sion of the context fusion module, flow-guided feature propagation, and different backbone choices 469 are included in our supplementary material. Lastly, we show our method may generate different 470 levels of stereo effects in Figure 5 compared to the ground truth, but this is expected due to the 471 underdetermined nature of the problem, and we consider our solution also as reasonable. 472

473

444

445 446

## 6 DISCUSSION AND LIMITATION

474 475

476 To enhance the viewing experience, films sometimes employ a stronger stereoscopic effect at the 477 start and end, while moderating it in the middle to ensure viewer comfort Neuman (2009); Ranftl et al. (2022). Thus, the stereo parameters such as focal length, are hard to retrieve even for the 478 same film. Theoretically, the precise reproduction of the right view is impossible without knowing 479 the stereo parameters in advance. By learning through a large-scale dataset, ImmersePro estimates 480 its average disparity, then tries to create an average-level stereo effect for input videos rather than 481 reproduce the precise right pair. Therefore, as shown in Figure 5, our model may produce reasonable 482 but "inaccurate" stereo effects if compared with the ground truth. 483

A reasonable stereo conversion pipeline involves a warping-and-inpainting process, where the in painting operation fills the black holes created by the warping operation. One sample work is stereo from-mono (Watson et al., 2020) that performs inpainting with a randomly sampled image from the



Figure 5: Visual demonstration of the disparity analysis results. Our network predicts reasonable stereo effects but may be stronger or weaker if compared to the ground truth. The L2R disparity computes the left-to-right disparity using RAFT-Stereo (Lipson et al., 2021).

training dataset. In a way, our method can be seen as an improvement to stereo-from-mono by intelligently selecting the correct regions for inpainting. However, this strategy works for creating stereo movies with "subtle" stereo effects without the need for significant inpainting. As we observed in most 3D movie examples, very few movies contain strong stereo effects. Notably, our method cannot produce strong stereo effects due to the limited dataset and limited inpainting capabilities. In future work, we would like to investigate how Nerf (Mildenhall et al., 2021)-based inpainting can be used for stereo-movie generation.

7 CONCLUSION

534 535

521

522

523 524 525

526

527

528

529

530

531 532 533

This work presents an end-to-end video-based stereo conversion method that generates right-view
 video sequences according to the input video. Our method automatically utilizes layered disparity
 maps on top of implicit disparities. Additionally, we propose *Youtube-SBS*, a large-scale stereo
 dataset that is publicly available for benchmarking purposes. Extensive qualitative and quantitative
 evaluations demonstrated the robustness of our approach against previous works.

# 540 REFERENCES

547

555

556

565

566

567

568

583

- Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 a model zoo for robust monocular
   relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.
- 544 D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical
  545 flow evaluation. In A. Fitzgibbon et al. (Eds.) (ed.), *European Conf. on Computer Vision (ECCV)*,
  546 Part IV, LNCS 7577, pp. 611–625. Springer-Verlag, oct 2012.
- Bei Chen, Jiabin Yuan, and Xiuping Bao. Automatic 2d-to-3d video conversion using 3d densely connected convolutional networks. In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, November 2019. doi: 10.1109/ictai.2019.00058. URL http://dx.doi.org/10.1109/ICTAI.2019.00058.
- Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable
   convolutional networks. In *Proceedings of the IEEE international conference on computer vision*,
   pp. 764–773, 2017.
  - Frédéric Devernay and Paul Beardsley. Stereoscopic Cinema, pp. 11–51. Springer Berlin Heidelberg, 2010. ISBN 9783642123924. doi: 10.1007/978-3-642-12392-4\_2. URL http: //dx.doi.org/10.1007/978-3-642-12392-4\_2.
- Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W Taylor. Sstvos:
  Sparse spatiotemporal transformers for video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5912–5921, 2021.
- Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection net work for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3897–3906, 2019.
  - E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. URL http://lmb.informatik.uni-freiburg.de/ /Publications/2017/IMKDB17.
- Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. Advances in neural information processing systems, 28, 2015.
- Yi Li, Kyle Min, Subarna Tripathi, and Nuno Vasconcelos. Svitt: Temporal learning of sparse video text transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18919–18929, 2023.
- Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6197–6206, 2021.
- Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-toend framework for flow-guided video inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
  - Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*, 2021.
- 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*, pp. 14040–14049, 2021.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.
- Lukas Mehl, Andrés Bruhn, Markus Gross, and Christopher Schroers. Stereo conversion with
   disparity-aware warping, compositing and inpainting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4260–4269, 2024.

603

604

605

628

629

- Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- 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.
- Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for
   editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6038–6047, 2023.
  - Robert Neuman. Bolt 3d: a case study. In *Stereoscopic Displays and Applications XX*, volume 7237, pp. 133–142. SPIE, 2009.
- Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, March 2022. ISSN 1939-3539. doi: 10.1109/tpami.2020.3019967. URL http://dx.doi.org/10.1109/TPAMI. 2020.3019967.
- Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using contextaware layered depth inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2020.
- Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 402–419. Springer, 2020.
- Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In 2019 International Conference on 3D Vision (3DV), pp. 348–357. IEEE, 2019a.
- Lezhong Wang, Jeppe Revall Frisvad, Mark Bo Jensen, and Siavash Arjomand Bigdeli. Stereod iffusion: Training-free stereo image generation using latent diffusion models. *arXiv preprint arXiv:2403.04965*, 2024.
- Kintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration
   with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019b.
  - Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. *arXiv preprint arXiv:2307.08695*, 2023.
- Jamie Watson, Oisin Mac Aodha, Daniyar Turmukhambetov, Gabriel J Brostow, and Michael Firman. Learning stereo from single images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 722–740. Springer, 2020.
- 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*, pp. 842–857. Springer, 2016.
- Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement
  with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127(8):1106–1125, 2019.
- Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024.
- 647 Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

648 649 650	Yu Zhang, Dongqing Zou, Jimmy S Ren, Zhe Jiang, and Xiaohao Chen. Structure-preserving stereo- scopic view synthesis with multi-scale adversarial correlation matching. In <i>Proceedings of the</i> <i>IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pp. 5860–5869, 2019.
652 653	Zheyu Zhang and Ronggang Wang. Temporal3d: 2d-to-3d video conversion network with multi- frame fusion. In 2022 4th International Conference on Advances in Computer Technology, Infor- mation Science and Communications (CTISC), pp. 1–5. IEEE, 2022.
655 656 657	Shangchen Zhou, Chongyi Li, Kelvin C.K Chan, and Chen Change Loy. ProPainter: Improving propagation and transformer for video inpainting. In <i>Proceedings of IEEE International Conference on Computer Vision (ICCV)</i> , 2023.
658 659 660	Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In <i>Proceedings of the IEEE/CVF conference on computer vision and pattern recognition</i> , pp. 9308–9316, 2019.
662 663	
664 665	
666 667	
668 669 670	
671 672	
673 674	
675 676	
677 678 679	
680 681	
682 683	
684 685	
686 687 688	
689 690	
691 692	
693 694	
695 696	
698 699	
700 701	