Stereo Vision Conversion from Planar Videos Based on Temporal Multiplane Images

Shanding Diao[1], Yuan Chen[3], Yang Zhao[1,2*], Wei jia[1], Zhao Zhang[1], Ronggang Wang[2,4*]

[1]School of Computer and Information, Hefei University of Technology, Hefei 230009, China
[2]Peng Cheng National Laboratory, Shenzhen 518000, China
[3]School of Internet, Anhui University, Hefei 230039, China
[4]School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, Shenzhen 518055, China
yzhao@hfut.edu.cn, rgwang@pkusz.edu.cn

Abstract

With the rapid development of 3D movie and light-field displays, there is a growing demand for stereo videos. However, generating high-quality stereo videos from planar videos remains a challenging task. Traditional depth-image-based rendering techniques struggle to effectively handle the problem of occlusion exposure, which occurs when the occluded contents become visible in other views. Recently, the single-view multiplane images (MPI) representation has shown promising performance for planar video stereoscopy. However, the MPI still lacks real details that are occluded in the current frame, resulting in blurry artifacts in occlusion exposure regions. In fact, planar videos can leverage complementary information from adjacent frames to predict a more complete scene representation for the current frame. Therefore, this paper extends the MPI from still frames to the temporal domain, introducing the temporal MPI (TMPI). By extracting complementary information from adjacent frames based on optical flow guidance, obscured regions in the current frame can be effectively repaired. Additionally, a new module called masked optical flow warping (MOFW) is introduced to improve the propagation of pixels along optical flow trajectories. Experimental results demonstrate that the proposed method can generate high-quality stereoscopic or light-field videos from a single view and reproduce better occluded details than other state-of-the-art (SOTA) methods. https://github.com/Dio3ding/TMPI

Introduction

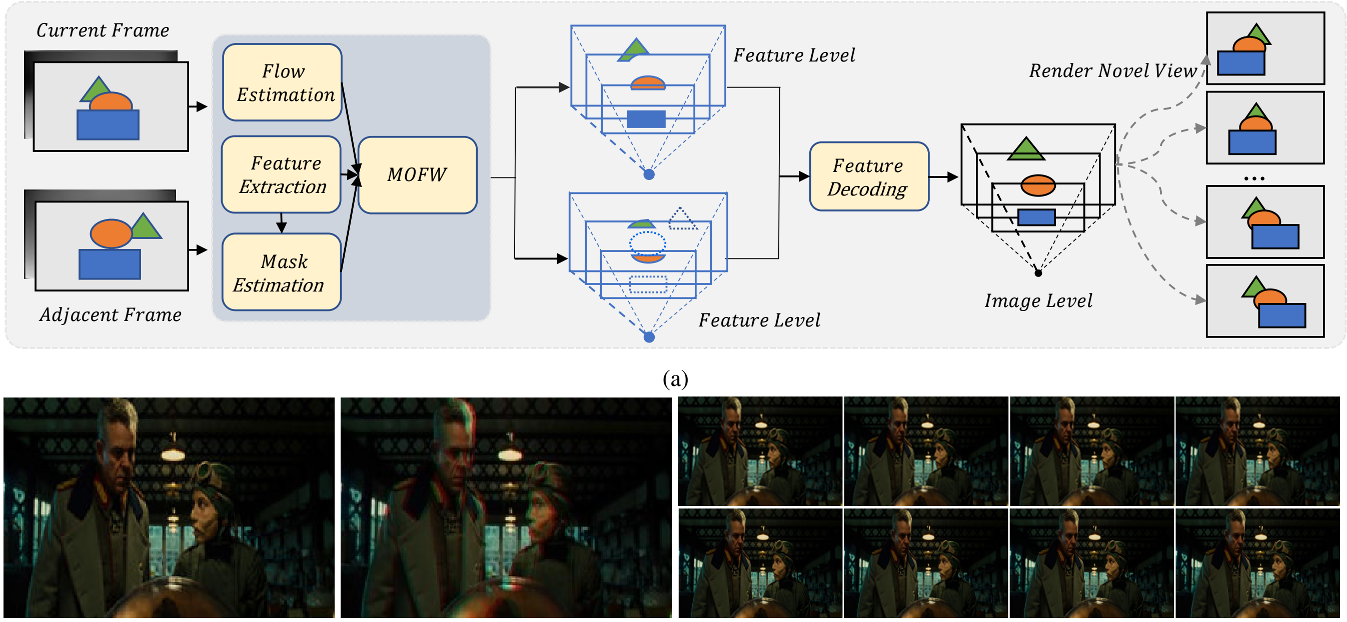
With the growing popularity of virtual reality, 3D movies and light-field display technology (Su et al. 2020), there is an increasing demand for stereoscopic videos. The most straightforward way to obtain stereo videos is multi-view capturing, which simultaneously records the same scene with multiple cameras. For 3D movies, two cameras can simulate human eyes and capture videos synchronously. In the case of light-field videos, more cameras are required to capture data from different viewpoints. However, after decades of accumulation, there are a huge number of precious planar video resources. At the same time, the cost and difficulty of single camera capturing are still much lower than that of multi-view capturing. Therefore, there is significant value in transforming classic planar videos into stereo vision using stereoscopic vision conversion techniques.

To generate stereo videos from monocular input, traditional methods usually estimate depth maps from single-view frame and then use depth-image-based rendering (DIBR) (Fehn 2004) algorithms to synthesize novel views. But DIBR process usually leaves holes in the occlusion exposure regions. Then, some deep neural network (DNN)-based 2D-to-3D video conversion methods have been proposed, e.g., Deep3D (Xie, Girshick, and Farhadi 2016) and depth estimation-based models (Lee et al. 2017). However, although recent monocular depth estimation methods (Ranftl, Bochkovskiy, and Koltun 2021; Godard et al. 2019; Zhang et al. 2023) have achieved promising performance, good depth maps can not guarantee satisfactory view synthesis. It is still difficult to acquire accurate dense 3D geometry or fill in the occluded contents of the scene. Implicit neural representation is another highly regarded approach to synthesize novel views (Wang et al. 2023; Peng et al. 2021). However, these methods require training specific representations for each scene, which greatly limits their application in the field of stereo video conversion, as each video contains a large number of constantly changing scenes.

Recently, the multiplane images (MPI) has gained a lot of attention (Zhou et al. 2018; Mildenhall et al. 2019; Srinivasan et al. 2019; Li et al. 2021; Han, Wang, and Yang 2022), which consists of N fronto-parallel RGB-α planes in the frustum of the source viewpoint, arranged at depths $(z_1, \cdots, z_N)$ from nearest to farthest. Each plane in MPI contains both RGB values and α value, which denotes volume density or transparency. By means of MPI, planes at different depths can better represent the details and spatial structure of the scene. Recent MPI-based methods (Li et al. 2021; Tucker and Snavely 2020; Han, Wang, and Yang 2022) can render multiple views from planar video frames, achieving planar video stereoscopy. However, these methods predict MPI without considering the temporal information and still cannot regenerate obscured details.

To utilize the temporal information of videos, this paper extends MPI to temporal domain and proposes a temporal MPI (TMPI) representation, as illustrated in Fig.1 (a). TMPI combines the MPIs of neighbor frames and presents

[*]Corresponding authors.



(a)

(b)

(c)

(d)

Figure 1: Framework and performance of the proposed temporal multiplane images (TMPI) method, (a) overall framework of TMPI, (b) input planar frame, (c) reconstructed stereo video frames of TMPI (displayed in red-cyan format), (d) reconstructed light-field frames of TMPI (8 viewpoints).

a masked optical flow warping (MOFW) module to obtain fused TMPI based on optical flow and mask prediction. The proposed method leverages complementary features from adjacent frame to fill in the occluded regions in the fused TMPI representation. The generated TMPI can then be rendered to multiple viewpoints. Fig.1 (c) shows an example of stereo frame (red-cyan format) generated by means of TMPI. The proposed method can also reproduce more synthetic viewpoints for light-field video conversion, as illustrated in Fig.1 (d).

The main contributions can be summarized as follows:

- This paper proposes a temporal MPI to extract additional information from adjacent frames to fill in occluded regions during the synthesis of new viewpoints. Compared to normal MPI, the proposed TMPI can capture more complete structure description of scene in planar videos.
- To avoid introducing artifacts caused by temporal fusion, a novel flow warping module is designed to propagate and fuse features from adjacent frames to the current frame using masked flow. Compared to conventional optical flow warping, our approach avoids the negative impacts caused by inaccurate optical flow estimation.
- A stereo video dataset is built to train our model by extracting frames from 3D movies. Experimental results of stereo video and light field video conversion demonstrate the proposed method can reproduce high-quality results with significantly improved performance in occluded regions compared to SOTA methods.

Related Work

Single-View View Synthesis

Predicting new views from single-view is a challenging problem with high ambiguity. Srinivasan et al. (Srinivasan et al. 2017) synthesize a 4D RGBD light field from 2D RGB image. More approaches (Niklaus et al. 2019; Wiles et al. 2020) generate new views based on predicted depth maps. Layered depth image (LDI)-based methods (Shade et al. 1998; Tulsiani, Tucker, and Snavely 2018; Shih et al. 2020) store multiple RGBD pixels, which can use different numbers of layers at each pixel location. However, The LDI suffers from depth discontinuities and struggles with complex 3D scene structures. Recently, MPI representation (Zhou et al. 2018; Mildenhall et al. 2019; Srinivasan et al. 2019; Tucker and Snavely 2020; Li et al. 2021; Han, Wang, and Yang 2022) becomes popular as it can explicitly model occluded contents in the scene. Original MPI apporaches (Zhou et al. 2018; Mildenhall et al. 2019; Srinivasan et al. 2019) use multiple views as input to predict the scene representation, while some recent methods (Tucker and Snavely 2020; Li et al. 2021; Han, Wang, and Yang 2022) merely rely on a single input view. However, single-frame MPI neglects the temporal complementarity and content relevance of neighbor frames, and still cannot regenerate high-quality occluded regions.

Flow-Based Video Inpainting

Optical flow is commonly used for alignment and fusion of neighbor frames in many video processing tasks, such as video object segmentation (Cheng et al. 2017; Hu et al. 2018; Jampani, Gadde, and Gehler 2017; Li and Loy 2018), and video super-resolution (Liao et al. 2015; Haris, Shakhnarovich, and Ukita 2020; Xin et al. 2020). Additionally, optical flow has contributed to video inpainting, which tends to restore missing regions in corrupted videos. For instance, DFG (Xu et al. 2019) uses a flow-guided inpainting network that employed an adaptive fusion module to smooth the boundaries between different frames. FGVC (Gao et al. 2020) presents an end-to-end framework consisting of a flow completion module and a content completion module. Li et al. (Li et al. 2022) introduce a scene template to ensure consistency between the flow and the scene. These methods primarily rely on pixel propagation, where holes are filled by bidirectionally propagating pixels from visible areas guided by the optical flow.

The Proposed Method

Given a video sequence $\{\boldsymbol{I}^t \in \mathcal{R}^{H\times W\times 3} | t = 1\cdots T\}$ with sequence length T and corresponding frame-wise depth maps $\{\boldsymbol{D}^t \in \mathcal{R}^{H\times W\times 1} | t = 1\cdots T\}$, we aim at synthesizing a TMPI representation $\{\boldsymbol{P}^t \in \mathcal{R}^{H\times W\times N\times 4} | t = 1\cdots T\}$, where N denotes the number of planes. Then new view can be rendered from the TMPI representation.

TMPI Representation

The MPI representation uses N RGB-α planes that are fronto-parallel to a reference camera's frustum to represent the scene, with each plane placed at a fixed depth. Each plane in MPI is composed of 4 channels of values, i.e., 3 channels of RGB color values $\{\boldsymbol{C}_i | i = 1\cdots N\}$, and one channel of transparency values $\{\boldsymbol{\alpha}_i | i = 1\cdots N\}$. In the TMPI, density maps $\{\boldsymbol{\sigma}_i | i = 1\cdots N\}$ are predicted instead of transparency maps, because the density maps can produce clearer results as noted in (Li et al. 2021). The transparency map $\boldsymbol{\alpha}_i$ can be converted from density map $\boldsymbol{\sigma}_i$ as,

$$\boldsymbol{\alpha}_i = \exp(-\boldsymbol{\delta}_i \boldsymbol{\sigma}_i), \tag{1}$$

where $\boldsymbol{\delta}_i$ represents the distance map between two adjacent planes. The Cartesian coordinate conversion from the perspective 3D coordinate $[x,\ y,\ z]^T$ is defined as $\mathfrak{C}\left([x,\ y,\ z]^T\right)$. Then, the distance map between $(i+1)$-th plane and i-th plane can be computed as,

$$\boldsymbol{\delta}_i = \left\| \mathfrak{C}\left([x,\ y,\ z_{i+1}]^T\right) - \mathfrak{C}\left([x,\ y,\ z_i]^T\right) \right\|_2, \tag{2}$$

In this paper, the TMPI is designed to leverage the complementary information of two neighbor frames $\boldsymbol{I}^t$ and $\boldsymbol{I}^{t+1}$. Each plane $\{\boldsymbol{P}_i^t | i = 1\cdots N\}$ is thus composed of fused RGB information $\mathcal{T}(\boldsymbol{C}_i^t, \boldsymbol{C}_i^{t+1})$ and fused density information $\mathcal{T}(\boldsymbol{\sigma}_i^t, \boldsymbol{\sigma}_i^{t+1})$, where $\mathcal{T}$ denotes the proposed TMPI calculation model which can extract complementary features from adjacent frame $\boldsymbol{I}^{t+1}$ and then fuse the warped features into the current frame $\boldsymbol{I}^t$. Then, the image $\boldsymbol{I}^t$ can be remapped back from TMPI representation by blending the N RGB-σ planes $\boldsymbol{P}_i^t$ from back to front in an iterative manner, as follows,

$$\boldsymbol{I}^t = \sum_{i=1}^{N} T_i \left(1 - \exp\left(\mathcal{T}(\boldsymbol{\sigma}_i^t, \boldsymbol{\sigma}_i^{t+1})\boldsymbol{\delta}_i\right)\ \mathcal{T}(\boldsymbol{C}_i^t, \boldsymbol{C}_i^{t+1})\right), \tag{3}$$

$$T_i = \exp\left(-\sum_{j=1}^{i-1} \mathcal{T}(\boldsymbol{\sigma}_j^t, \boldsymbol{\sigma}_j^{t+1})\boldsymbol{\delta}_j\right), \tag{4}$$

where T_i denotes the transmittance from the camera to the i-th plane. For synthesizing new views from TMPI, each plane $\boldsymbol{P}_i^t$ is firstly warped according to its depth and the target camera parameters as in (Hartley and Zisserman 2003; Zhou et al. 2018):

$$[x_s\ , y_s\ , 1]^T \sim K\left(R - \frac{v_T n^T}{z_i}\right) K^{-1} [x_t\ , y_t\ , 1]^T\ , \tag{5}$$

where x_s, y_s denote the pixel coordinates in the source camera, x_t, y_t are the pixel coordinates in the target camera, K represents the camera intrinsic matrix, R denotes the camera rotation matrix, v_T is the camera translation vector, $n = [0, 0, 1]^T$ is the plane normal vector, and z_i denotes the distance from the plane to the source camera.

Network Architecture

As shown in Fig.2, the proposed network is composed of two mask branches and a RGB-σ branch. The mask branch tends to predict assign masks, which determine which pixels should be placed in which plane in MPI. The two mask branches share the same structure and parameters for two adjacent frames $\boldsymbol{I}^t$ and $\boldsymbol{I}^{t+1}$. The RGB-σ branch then adopts masked flow warping to fuse intermediate features and produces the final TMPI representation. In each mask branch, the depth map $\boldsymbol{D}^t$ of frame $\boldsymbol{I}^t$ is firstly obtained by a monocular depth estimation model DPT (Ranftl, Bochkovskiy, and Koltun 2021). Then, $\boldsymbol{D}^t$ and $\boldsymbol{I}^t$ are simultaneously fed into a RGBD encoder $\mathcal{E}_{RGBD}$ to encode scene structure with depth information. The RGBD feature $\boldsymbol{E}^t$ is calculated as,

$$\boldsymbol{E}^t = \mathcal{E}_{RGBD}(\boldsymbol{I}^t, \boldsymbol{D}^t). \tag{6}$$

The RGBD feature $\boldsymbol{E}^t$ is then replicated by N times, and each copy is used to predict the mask and RGB-σ value for each plane. Motivated by ESPNetv2 (Bae, Moon, and Im 2023), the depth maps $\boldsymbol{D}^t$ is downsampled through an average pooling layer f_{avp} and then concatenated to $\boldsymbol{E}^t$, so that the spatial and depth information loss can be reduced. The mask decoder $\mathcal{D}_m$ takes the concatenated features and a single depth value $z_i (i = 1, \cdots, N)$ as input. Specifically, we sample the disparity value $d_i (i = 1, \cdots, N)$ in the perspective geometry and $z_i = 1/d_i$. Note that d_i is uniformly distributed from 1 to 0.001. Finally, N assign masks $\{\boldsymbol{M}_i^t | i = 1\cdots N\}$ are produced after a softmax layer:

$$\boldsymbol{M}_i^t = Softmax(\mathcal{D}_m(\boldsymbol{E}^t, z_i, f_{avp}(\boldsymbol{D}^t))). \tag{7}$$

Next, the optical flow $\boldsymbol{F}^{t+1\rightarrow t}$ from frame $\boldsymbol{I}^{t+1}$ to frame $\boldsymbol{I}^t$ is estimated using the flow estimation model RAFT (Teed

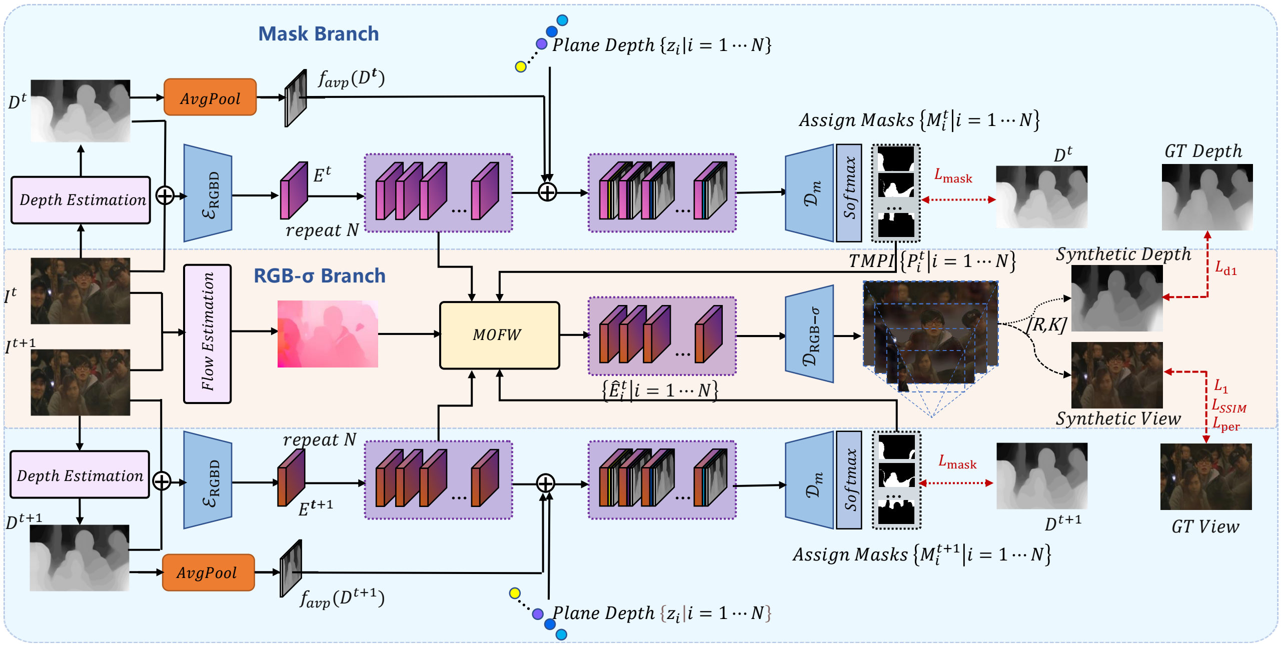


Figure 2: The network architecture of the proposed TMPI method.

and Deng 2020). The proposed MOFW module is then used to propagate the features of adjacent frames and fuse them with $\boldsymbol{E}^t$ to generate the fused feature $\hat{\boldsymbol{E}}_i^t$ as follows,

$$\hat{\boldsymbol{E}}_i^t = f_{MOFW}(\boldsymbol{E}^t, \boldsymbol{E}^{t+1}, \boldsymbol{M}_i^t, \boldsymbol{M}_i^{t+1}, \boldsymbol{F}^{t+1\to t}). \quad (8)$$

Finally, the RGB-σ decoder $\mathcal{D}_{RGB-\sigma}$ reconstructs the fused feature $\hat{\boldsymbol{E}}_i^t$ into the i-th plane in TMPI, as follows:

$$\boldsymbol{P}_i^t = \mathcal{D}_{RGB-\sigma}(\hat{\boldsymbol{E}}_i^t), \quad (9)$$

where $\boldsymbol{P}_i^t$ denotes the plane which contains 4 channels of R, G, B values $\boldsymbol{C}_i^t$ and density value $\boldsymbol{\sigma}_i^t$. After the calculation of each $\boldsymbol{P}_i^t$, the TMPI representation is obtained, which can be used to synthesize stereo viewpoints by using Eq.3 and Eq.5. In this paper, the RGBD encoder, mask decoder, and RGB-σ decoder adopt the same backbone structure as in (Zhang et al. 2023).

Masked Optical Flow Warping

Directly applying traditional optical flow warping to fuse features of two frames may lead to blurry, ringing or ghosting artifacts due to inaccurate flow estimation. Moreover, we do not require all information from the adjacent frame, but only the valid region where the areas occluded in $\boldsymbol{I}^t$ but visible in $\boldsymbol{I}^{t+1}$. Hence, a MOFW module is designed. First, the context feature $\{\boldsymbol{E}_i^t|i=1\cdots N\}$ and context flow $\left\{\boldsymbol{F}_i^{t+1\to t}\middle|i=1\cdots N\right\}$ are calculated for each plane of TMPI as follows,

$$\boldsymbol{E}_i^t = \boldsymbol{E}^t \odot \sum_{j=i}^{N} \boldsymbol{M}_j^t, \quad (10)$$

$$\boldsymbol{F}_i^{t+1\to t} = \boldsymbol{F}^{t+1\to t} \odot \sum_{j=i}^{N} \boldsymbol{M}_j^{t+1}, \quad (11)$$

where symbol $\odot$ represents element-wise multiplication, $\boldsymbol{M}_j^t$ denotes the estimated assign mask, and $\sum_{j=i}^N \boldsymbol{M}_j^t$ denotes the context regions for plane $\boldsymbol{P}_i^t$ that can fill in the occluded pixels. To avoid the influence of the occluding content in the front, we set the context mask to be the combination of the pixels on and behind the i-th plane. Similarly, $\sum_{j=i}^N \boldsymbol{M}_j^{t+1}$ denotes the context regions for calculating the context flow. For each plane, the contained information is the feature selected by the context mask $\sum_{j=i}^N \boldsymbol{M}_j^t$, while the other regions are holes (near 0 values). In other words, the mask of hole regions is the inversion of the context mask $(1-\sum_{j=i}^N \boldsymbol{M}_j^t)$, namely inpainting mask. Next, we hope to inpainting the missing hole regions by fusing the context features $\boldsymbol{E}_i^{t+1}$ extracted from the neighbor frame. We can intuitively obtain the warped features $\hat{\boldsymbol{E}}_i^{t+1}$ by forward warping the $\boldsymbol{E}_i^{t+1}$ to current frame with the context flow $\boldsymbol{F}_i^{t+1\to t}$. Then the warped features $\hat{\boldsymbol{E}}_i^{t+1}$ is concatenated with $\boldsymbol{E}_i^t$ to provide additional context features.

As mentioned before, most of the information in $\hat{\boldsymbol{E}}_i^{t+1}$ is redundant with $\boldsymbol{E}_i^t$. What we need is only the information that is missing from the $\boldsymbol{E}_i^t$, i.e., $\hat{\boldsymbol{E}}_i^{t+1} \odot \left(1-\sum_{j=i}^N \boldsymbol{M}_j^t\right)$. As a result, the final warped feature can be further merged with current context feature $\boldsymbol{E}_i^t$ as follows,

$$\hat{\boldsymbol{E}}_i^t = \mathcal{P}\left(\boldsymbol{E}_i^t, \mathcal{W}(\boldsymbol{E}_i^{t+1}, \boldsymbol{F}_i^{t+1\to t}) \odot \left(1-\sum_{j=i}^{N} \boldsymbol{M}_j^t\right)\right), \quad (12)$$

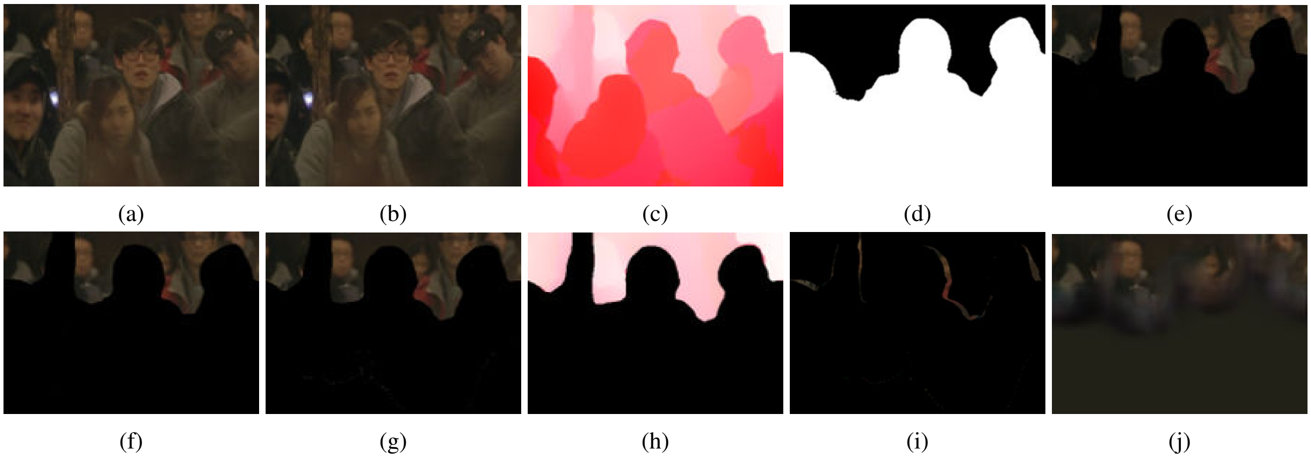


Figure 3: Visualization of MOFW process, (a) RGBD feature $\boldsymbol{E}^t$, (b) RGBD feature $\boldsymbol{E}^{t+1}$, (c) Optical Flow $\boldsymbol{F}^{t+1\to t}$, (d) Inpainting Mask $(1-\sum_{j=9}^{N}\boldsymbol{M}_j^t)$, (e) Fusion Features $\hat{\boldsymbol{E}}_9^t$, (f) Context Feature $\boldsymbol{E}_9^t$, (g) Context Feature $\boldsymbol{E}_9^{t+1}$, (h) Context Flow $\boldsymbol{F}_9^{t+1\to t}$, (i) $\mathcal{W}\left(\boldsymbol{E}_9^{t+1},\boldsymbol{F}_9^{t+1\to t}\right)\odot\left(1-\sum_{j=9}^{N}\boldsymbol{M}_j^t\right)$, (j) 4-channel plane $\boldsymbol{P}_9^t$ in TMPI. For the sake of intuition, the frame is used in these figures instead of feature maps.

where $\mathcal{W}(\cdot)$ denotes the forward warping operation based on optical flow, and the propagation function $\mathcal{P}(\cdot)$ consists of two 3×3 convolutional layers.

Fig.3 visualizes the process of how our MOFW module generates the fused feature $\hat{\boldsymbol{E}}_9^t$ of the 9-th plane. It utilizes the context feature $\boldsymbol{E}_9^t$ in Fig.3 (f), context feature $\boldsymbol{E}_9^{t+1}$ in Fig.3 (g), context flow $\boldsymbol{F}_9^{t+1\to t}$ in Fig.3 (h), and inpainting mask $\left(1-\sum_{j=i}^{N}\boldsymbol{M}_j^t\right)$ in Fig.3 (d) to propagate complementary features from adjacent frame, as shown in Fig.3 (i). Then the features of current frame and the complementary features are fused to generate the final feature $\hat{\boldsymbol{E}}_9^t$. The fused features are then fed into the aforementioned RGB-σ decoder to produce the 4-channel plane $\boldsymbol{P}_9^t$.

Loss Functions

There are total five terms in the loss function, i.e., RGB L1 loss $\mathcal{L}_1$, RGB SSIM loss $\mathcal{L}_{SSIM}$, RGB perceptual loss (Liu et al. 2018) $\mathcal{L}_{per}$, L1 loss of depth map $\mathcal{L}_{d1}$, and the mask assign loss $\mathcal{L}_{mask}$. The total loss is given by:

$$\mathcal{L}=\lambda_1\mathcal{L}_1+\lambda_2\mathcal{L}_{SSIM}+\lambda_3\mathcal{L}_{per}+\lambda_4\mathcal{L}_{d1}+\lambda_5\mathcal{L}_{mask}, \tag{13}$$

where the weights $\lambda_1,\lambda_2,\lambda_3,\lambda_4$, and λ_5 are experimentally set to $1,1,1,1$, and 10, respectively.

Inspired by (Han, Wang, and Yang 2022), $\mathcal{L}_{mask}$ can measure the error to represent the depth map $\boldsymbol{D}^t$ using N discrete planes at depth $\{z_i\}_{i=1}^N$ with masks $\{\boldsymbol{M}_i^t\}_{i=1}^N$, as follows,

$$\mathcal{L}_{mask}=\frac{1}{HW}\sum_{i=1}^{N}\sum_{(x,y)}\boldsymbol{M}_i^t\odot\left|\boldsymbol{D}^t-z_i\right|. \tag{14}$$

Experiments

Datasets and Implementation Details

Training set We introduce a training dataset contained 12388 pairs of 720×480 stereo frames collected from 20 recent 3D movies. As 3D movies provide binocular perspectives, the binocular depth estimation network DPSNet (Im et al. 2019) is applied to obtain depth maps for each frame.

In addition, in order to further improve the scale and diversity of training samples, we adopt the approach proposed in (Watson et al. 2020) to transform a collection of single-view RGB videos into stereo training data. In our experiments, this method is applied to the widely used Vimeo90K video dataset, which consists of $73,171$ 3-frame sequences with a fixed resolution of 448×256.

2D-to-3D video conversion test set For evaluating the performance of 3D video conversion, 10 3D movies different from that in training set are used to create the test set contained 3323 five-frame sequences. The left view of the 3D movies is used as input and the right view is reconstructed with different methods. The differences between the ground-truth (GT) right view and the reconstructed right view are measured for quantitative and qualitative evaluation.

Light-field video conversion test set Due to the lack of light field video datasets, a test set is built to simulate sparse light field data based on Middlebury stereo dataset (Scharstein and Pal 2007; Hirschmuller and Scharstein 2007), which consists of high-resolution stereo sequences with complex geometry. Each sequence contains 7 views, with views 1 and 5 providing pixel-accurate ground-truth disparity data. However, this dataset lacks temporal variations between frames. Therefore, we use view 1 as the current frame and view 5 as the adjacent frame, and then compare three synthesized views with the GT views.

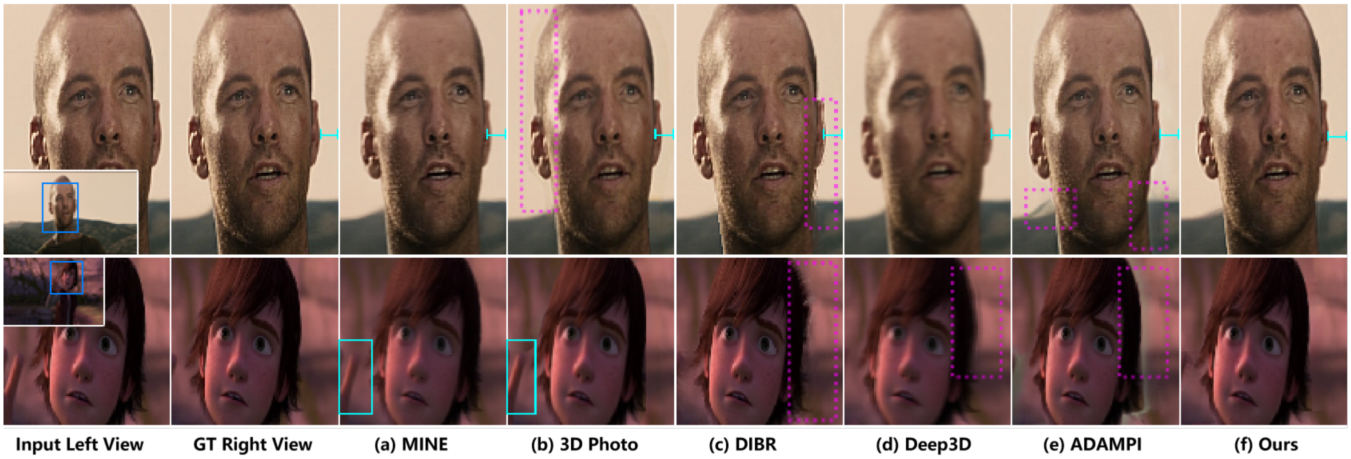


Figure 4: The 2D-to-3D conversion results of different methods on 3D movie test set.

Method	LPIPS↓	PSNR↑	SSIM↑	Param (M)
ADAMPI	0.036	35.44	0.956	57.12
MINE	0.057	30.78	0.877	38.06
3D Photo	0.068	29.15	0.859	114.56
Deep3d	0.070	31.41	0.898	84.01
DIBR	0.036	35.14	0.928	**4.19**
Ours	**0.027**	**36.33**	**0.960**	37.55

Table 1: PSNR, SSIM, LPIPS scores and Parameters of different methods on 3D movie test set

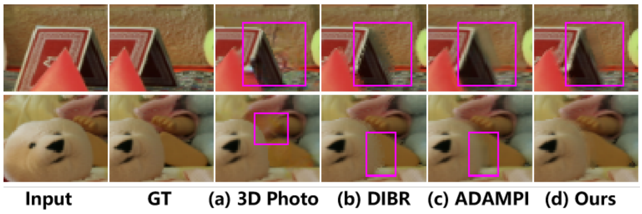


Figure 5: Visual comparisons of synthetic views of different methods on Middlebury stereo dataset.

In addition, to verify the generalization and robustness of the light field synthesis models, we select 10 planar video clips from the Youku2K video set (Youku 2019), and then convert them to light field videos with 8 views using different methods.

It is noted that these simulated light field videos are still different from the real dense light field data, and mainly focus on horizontal disparity, thus they are only used to verify the potential capability of different methods when synthesizing multiview light-field-like data.

Implementation We first train our network on stereo training data generated from Vimeo90K for 800,000 steps with an initial learning rate of 0.0002 for the encoder, and 0.001 for the decoder. Then we use 3D movies dataset to finetune the model for 200,000 steps. Our model is optimized by Adam with weight decay $1e-4$ in training stage. The number N of planes in TMPI is set to 64.

Experimental Results

To verify the effectiveness of the proposed method, several stereo vision conversion methods are used for comparison, such as Deep3d (Xie, Girshick, and Farhadi 2016), DPSNet+DIBR, 3D-Photo (Shih et al. 2020), MINE (Li et al. 2021), and ADAMPI (Han, Wang, and Yang 2022). We augment Deep3d with depth map, as it does not utilize depth information.

3D Video Conversion Results Fig.4 illustrates 2D-to-3D conversion results of different methods. Firstly, DIBR causes holes artifacts and Deep3D tends to produce blurry results. Secondly, 3D-Photo and MINE produce smaller disparities than GT, which leads to weaker stereoscopic perception. Thirdly, ADAMPI still cannot recover occluded contents and generates unsatisfying halos. Overall, these methods cannot well restore the difficult occlusion exposure regions, while the proposed method can obtain better subjective results than other methods.

For objective testing, we adopt two commonly used distortion metrics, i.e., PSNR and SSIM, and one perceptual similarity measure LPIPS (Zhang et al. 2018). Table 1 lists the quality scores of these methods, it can be found that our method outperforms other methods on all metrics, which indicates that the proposed TMPI can achieve less distortion and higher fidelity. In addition, the parameters of the proposed method do not exceed recent SOTA models.

Light Field Video Synthesis Results Compared with stereo videos, multiview light-field-like videos have much larger disparities. Fig.5 shows synthesized results on Middlebury stereo dataset. By comparing the reconstructed occluded contents, we can get the following observations. Firstly, DIBR remains holes and 3D-Photo may inpaint error contents. Secondly, ADAMPI still cannot handle unknown occluded contents and produce blur halos. Benefit from the temporal complementary information in TMPI, the

	View0			View2			View3		
Method	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑
ADAMPI	0.036	29.71	0.950	0.034	32.22	0.956	0.052	28.71	0.916
3D Photo	0.052	27.21	0.916	0.053	26.97	0.890	0.065	25.50	0.853
DIBR	0.060	27.34	0.922	0.036	31.66	0.954	0.061	28.38	0.916
Ours	**0.032**	**30.51**	**0.957**	**0.030**	**32.46**	**0.959**	**0.050**	**28.75**	**0.919**

Table 2: Comparison results of different views on Middlebury stereo dataset

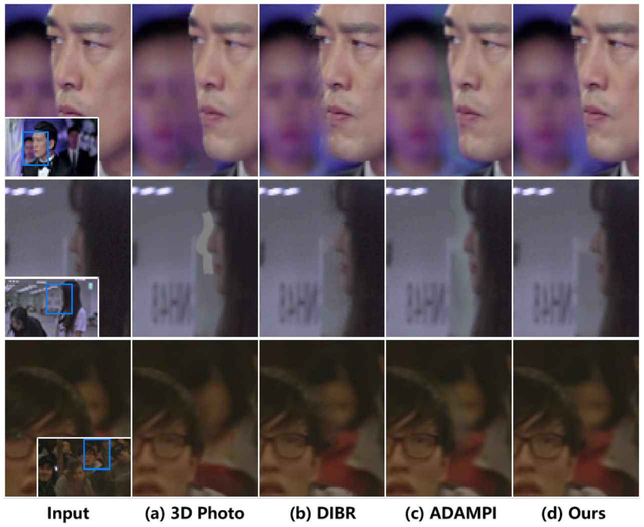


Figure 6: Comparison of light field synthesis results from planar video clips.

Method	paq2piq↑	ava↑	spaq↑	MOS↑
ADAMPI	72.70	4.54	61.71	3.33
3D Photo	72.40	4.42	59.84	2.65
DIBR	72.13	4.51	61.29	2.92
Ours	**72.91**	**4.59**	**62.26**	**3.96**

Table 3: BIQA and MOS results of light field video conversion from planar clips

proposed method can synthesize occluded regions that are more consistent with GT. Related quantitative results are listed in Table 2, from which we can find the proposed method sill outperforms other SOTA methods.

To verify the generalization ability, a planar video set is used to synthesize light field videos with different methods. As illustrated in Fig.6, DIBR, 3D-Photo, and ADAMPI methods suffer from the missing information of occluded areas, and thus cause blur, holes or halos. The proposed method can reproduce natural and high quality results. Because GT are unavailable in this test set, mean opinion score (MOS) and some blind image quality assessments (BIQA) are listed in Table 3, i.e., paqpiq (Ying et al. 2020), ava (Ke et al. 2021) and spaq (Ke et al. 2021). To obtain MOS, 16 observers are invited to score the visual quality of synthetic results from 1 (the worst) to 5 (the best). The proposed method still achieves the highest BIQA and MOS values, which verifies the robustness of the proposed TMPI methods.

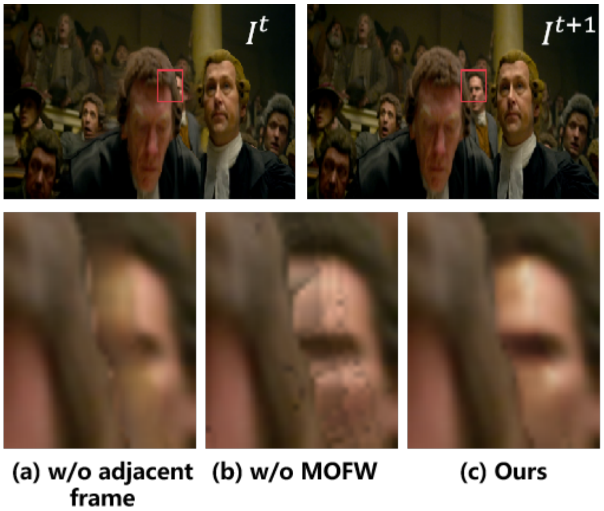


Figure 7: Subjective results for ablation study on 3D video conversion, (a) without using adjacent frame, (b) without MOFW module, (c) Ours.

	LPIPS↓	PSNR↑	SSIM↑
w/o adjacent frame	0.029	36.28	0.954
w/o MOFW	0.049	34.72	0.944
Ours	**0.027**	**36.33**	**0.960**

Table 4: Ablation study results on 3D movie test set

Ablation Study

Ablation experiments are conducted on the 3D movie test set to evaluate the effects of adjacent frames and the MOFW module. The results of the ablation tests are shown in Table 4. When adjacent frame is not used, there is a slight decrease in these metrics. It is worth noting that the limited decrease in numerical values is due to the fact that our method only utilizes complementary information from the exposed regions in the adjacent frames. Although these regions are visually noticeable, their area is relatively small compared to the entire image, resulting in only a small decrease in the metrics computed over the entire image. On the other hand, when the MOFW module is removed, traditional flow warping may introduce ghosting artifacts, thus leading to a significant drop of quality scores.

Fig.7 illustrates subjective comparisons of the ablation study. In the current frame, the face in background is occluded by the foreground character, resulting in an unrealistic blurred face when he is visible in other viewpoints, as shown in Fig. 7 (a). By using the adjacent frame, the missing facial details can be effectively reconstructed. As in Fig. 7 (b), without the MOFW module, traditional warping process may introduce severe artifacts, which confirms the effectiveness of the proposed MOFW module.

Limitation

While the TMPI model exhibits robust performance, its current inference speed is inadequate for real-time applications on devices with limited computational capabilities. In our future work, we aim to enhance the inference speed by optimizing the network architecture and simplifying the plane representation based on the scene structure. Furthermore, we plan to utilize the TMPI model to distill a lightweight model.

Conclusion

This paper proposes a novel Temporal Multiplane Images (TMPI) representation for stereo vision conversion from planar videos. In comparison to the single-frame MPI, the TMPI can extract complementary information from adjacent frames, effectively addressing the challenging task of recovering details in the occluded regions. Specifically, a TMPI network is designed with two mask branches and one plane reconstruction branch. To overcome fusion artifacts arising from inaccurate flow estimation, a masked optical flow warping module is introduced. This module can refine the occluded regions by fusing specific masked context features of neighboring frames. Experimental results demonstrate that the proposed TMPI can reproduce high-quality occlusion exposure areas and outperforms SOTA methods.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities (No. JZ2022HGPA0309), and the National Natural Science Foundation of China (Nos. 62272142, 62076086, 62072013, 62072151).

References

Bae, J.; Moon, S.; and Im, S. 2023. Deep digging into the generalization of self-supervised monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1, 187–196.

Cheng, J.; Tsai, Y.-H.; Wang, S.; and Yang, M.-H. 2017. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*, 686–695.

Fehn, C. 2004. Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In *Stereoscopic displays and virtual reality systems XI*, volume 5291, 93–104. SPIE.

Gao, C.; Saraf, A.; Huang, J.-B.; and Kopf, J. 2020. Flow-edge guided video completion. In *Computer Vision–ECCV 2020: 16th European Conference*, 713–729. Springer.

Godard, C.; Mac Aodha, O.; Firman, M.; and Brostow, G. J. 2019. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3828–3838.

Han, Y.; Wang, R.; and Yang, J. 2022. Single-view view synthesis in the wild with learned adaptive multiplane images. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–8.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2020. Space-time-aware multi-resolution video enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2859–2868.

Hartley, R.; and Zisserman, A., eds. 2003. *Multiple view geometry in computer vision*. Cambridge university press.

Hirschmuller, H.; and Scharstein, D. 2007. Evaluation of cost functions for stereo matching. In *2007 IEEE conference on computer vision and pattern recognition*, 1–8. IEEE.

Hu, P.; Wang, G.; Kong, X.; Kuen, J.; and Tan, Y.-P. 2018. Motion-guided cascaded refinement network for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1400–1409.

Im, S.; Jeon, H.-G.; Lin, S.; and Kweon, I. S. 2019. Dpsnet: End-to-end deep plane sweep stereo. arXiv:1905.00538.

Jampani, V.; Gadde, R.; and Gehler, P. V. 2017. Video propagation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 451–461.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157.

Lee, J.; Jung, H.; Kim, Y.; and Sohn, K. 2017. Automatic 2d-to-3d conversion using multi-scale deep neural network. In *2017 IEEE International Conference on Image Processing*, 730–734. IEEE.

Li, J.; Feng, Z.; She, Q.; Ding, H.; Wang, C.; and Lee, G. H. 2021. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12578–12588.

Li, X.; and Loy, C. C. 2018. Video object segmentation with joint re-identification and attention-aware mask propagation. In *Proceedings of the European conference on computer vision*, 90–105.

Li, Z.; Lu, C.-Z.; Qin, J.; Guo, C.-L.; and Cheng, M.-M. 2022. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17562–17571.

Liao, R.; Tao, X.; Li, R.; Ma, Z.; and Jia, J. 2015. Video super-resolution via deep draft-ensemble learning. In *Proceedings of the IEEE international conference on computer vision*, 531–539.

Liu, G.; Si, J.; Hu, Y.; and Li, S. 2018. Photographic image synthesis with improved U-net. In *2018 Tenth International*

Conference on Advanced Computational Intelligence, 402–407. IEEE.

Mildenhall, B.; Srinivasan, P. P.; Ortiz-Cayon, R.; Kalantari, N. K.; Ramamoorthi, R.; Ng, R.; and Kar, A. 2019. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics*, 38(4): 1–14.

Niklaus, S.; Mai, L.; Yang, J.; and Liu, F. 2019. 3d ken burns effect from a single image. *ACM Transactions on Graphics*, 38(6): 1–15.

Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9054–9063.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12179–12188.

Scharstein, D.; and Pal, C. 2007. Learning conditional random fields for stereo. In *2007 IEEE conference on computer vision and pattern recognition*, 1–8. IEEE.

Shade, J.; Gortler, S.; He, L.-w.; and Szeliski, R. 1998. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 231–242.

Shih, M.-L.; Su, S.-Y.; Kopf, J.; and Huang, J.-B. 2020. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8028–8038.

Srinivasan, P. P.; Tucker, R.; Barron, J. T.; Ramamoorthi, R.; Ng, R.; and Snavely, N. 2019. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 175–184.

Srinivasan, P. P.; Wang, T.; Sreelal, A.; Ramamoorthi, R.; and Ng, R. 2017. Learning to synthesize a 4D RGBD light field from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, 2243–2251.

Su, Y.; Tang, X.; Cai, Z.; Wu, J.; Chen, Y.; Hua, M.; and Wan, W. 2020. Performance improvement of projection-type multiview holographic three-dimensional display using spatial light modulators. *Optics and Lasers in Engineering*, 129: 106079.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference*, 402–419. Springer.

Tucker, R.; and Snavely, N. 2020. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 551–560.

Tulsiani, S.; Tucker, R.; and Snavely, N. 2018. Layer-structured 3d scene inference via view synthesis. In *Proceedings of the European Conference on Computer Vision*, 302–317.

Wang, T.; Sarafianos, N.; Yang, M.-H.; and Tung, T. 2023. Neural Rendering of Humans in Novel View and Pose from Monocular Video. arXiv:2204.01218.

Watson, J.; Aodha, O. M.; Turmukhambetov, D.; Brostow, G. J.; and Firman, M. 2020. Learning stereo from single images. In *Computer Vision–ECCV 2020: 16th European Conference*, 722–740. Springer.

Wiles, O.; Gkioxari, G.; Szeliski, R.; and Johnson, J. 2020. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7467–7477.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference*, 842–857. Springer.

Xin, J.; Wang, N.; Li, J.; Gao, X.; and Li, Z. 2020. Video face super-resolution with motion-adaptive feedback cell. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 07, 12468–12475.

Xu, R.; Li, X.; Zhou, B.; and Loy, C. C. 2019. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3723–3732.

Ying, Z.; Niu, H.; Gupta, P.; Mahajan, D.; Ghadiyaram, D.; and Bovik, A. 2020. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3575–3585.

Youku. 2019. Youku Video Super-Resolution and Enhancement Challenge dataset (Youku-VSRE2019). https://tianchi.aliyun.com/dataset/39568. Accessed: 2019-09-24.

Zhang, N.; Nex, F.; Vosselman, G.; and Kerle, N. 2023. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18537–18546.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhou, T.; Tucker, R.; Flynn, J.; Fyffe, G.; and Snavely, N. 2018. Stereo magnification. *ACM Transactions on Graphics*, 37(4): 1–12.