Machine Intelligence Research
www.mi-research.net

20(3), June 2023, 408-420
DOI: 10.1007/s11633-022-1381-9

Robust Local Light Field Synthesis via Occlusion-aware Sampling and Deep Visual Feature Fusion

Wenpeng Xing[1] Jie Chen[1] Yike Guo[1, 2]

[1] Department of Computer Science, Hong Kong Baptist University, Hong Kong 999077, China

[2] Data Science Institute, Imperial College London, London SW7 2AZ, UK

Abstract: Novel view synthesis has attracted tremendous research attention recently for its applications in virtual reality and immersive telepresence. Rendering a locally immersive light field (LF) based on arbitrary large baseline RGB references is a challenging problem that lacks efficient solutions with existing novel view synthesis techniques. In this work, we aim at truthfully rendering local immersive novel views/LF images based on large baseline LF captures and a single RGB image in the target view. To fully explore the precious information from source LF captures, we propose a novel occlusion-aware source sampler (OSS) module which efficiently transfers the pixels of source views to the target view′s frustum in an occlusion-aware manner. An attention-based deep visual fusion module is proposed to fuse the revealed occluded background content with a preliminary LF into a final refined LF. The proposed source sampling and fusion mechanism not only helps to provide information for occluded regions from varying observation angles, but also proves to be able to effectively enhance the visual rendering quality. Experimental results show that our proposed method is able to render high-quality LF images/novel views with sparse RGB references and outperforms state-of-the-art LF rendering and novel view synthesis methods.

Keywords: Novel view synthesis, light field (LF) imaging, multi-view stereo, occlusion sampling, deep visual feature (DVF) fusion.

Citation: W. Xing, J. Chen, Y. Guo. Robust local light field synthesis via occlusion-aware sampling and deep visual feature fusion. *Machine Intelligence Research*, vol.20, no.3, pp.408–420, 2023. http://doi.org/10.1007/s11633-022-1381-9

1 Introduction

One critical assumption for a successful novel view synthesis is accurate depth/disparity estimation, which is also a fundamental application of local light field (LF) imaging. In this paper, we propose a novel LF synthesis algorithm in a global multi-view stereo framework that can take large baseline input reference LFs to estimate accurate depth in the novel/target view. The estimated accurate depth in the novel/target view is used for warping the target view image into a novel LF.

In order to predict an accurate disparity in the target view, we fully exploit the advantages of LF on an accurate disparity estimation. We first calculate disparity probability volume (DPV) from the slope of lines in Epipolar-plane images (EPI)[1] in each source LF. Then, we fuse these source DPVs from source LFs into the target view′s camera frustum. But the source DPVs are in different scales and cannot be fused directly. So, we use the DPV rescaling and fusion methods in [2] to align these DPVs before fusion. After the fusion process, a novel DPV in the target view is generated. An accurate disparity map can be estimated from the novel DPV. The rationality of fusing DPVs rather than directly using depth projections is that DPVs contain the probability of a spatial point being occupied, which is much more indicative than single-plane depth when fused from multi-views.

Then, a preliminary LF is first synthesized by backward warping pixels of the target view image according to the estimated disparity map. To fully exploit the known valuable color information from multiple input views, occlusion-aware source sampler (OSS) and deep visual fusion (DVF) modules are proposed. The OSS module takes plane sweep volumes (PSV)[3] as input and locates depth planes with maximum confidence for the image patches in a PSV. A global background image is composed by the inverse-over composition of pixels from further depth layers of the PSV. Then the global background image is warped into a background LF, which is fused with the preliminary LF to enhance visual quality by eliminating image noises and recovering the occluded contents. Then the fused LF is fed into a spatial-angular regularisation module for improving spatial-angular consistency and visual quality. As shown in the evaluation results, our method outperforms other state-of-the-art novel view synthesis methods in the Stanford Lytro multi-view light field dataset (MVLF)[4] that contains challenging large baseline and discrete views in each scene. The proposed OSS and DVF modules can greatly improve the visual quality by suppressing noise and revealing occluded con-

Research Article
Manuscript received on June 25, 2022; accepted on October 18, 2022; published online on March 17, 2023
Recommended by Associate Editor Hui-Yu Zhou
Colored figures are available in the online version at https://link.springer.com/journal/11633

tents.

Our contributions are summarised as follows:

1) The OSS module is proposed to sample visual clues from source LFs in the depth planes of a PSV with minimum pixel-matching errors.

2) The DVF module is proposed to fuse the deep visual features of a background LF and a preliminary LF.

2 Related work

Novel view synthesis techniques usually use depth-based warping operations[5–9] that warp image pixels to produce novel views. Therefore, accurate depth estimation is crucial for accurate warping, and occlusions further complicate the rendering. Learning-based LF/novel view synthesis methods can be classified into three categories according to their input images′ sampling patterns: sparse angular inputs, single RGB input, and small baseline multi-view inputs.

2.1 LF synthesis based on sparse angular references

Sparse-input LF synthesis takes a sparse set of sub-aperture images (SAIs) captured within a target LF′s aperture to synthesize novel neighboring SAIs by interpolation or extrapolation. Kalantari et al.[5] introduced the first learning-based LF synthesis solution. Their method takes SAIs in four corners as input to synthesize a 4D LF using two sequential convolution neural networks to estimate disparity and color. But explicit scene geometry is not a necessary condition for LF synthesis, Zhang et al.[10] proposed a phase-based LF synthesis method from a micro-baseline stereo pair. Yeung et al.[9] reconstructed dense-sampled SAIs from sparse-sampled SAIs using spatial-angular alternative convolutions to exploit dense spatial and angular clues. The sparse inputs within the LF′s aperture usually require fixed input sub-aperture positions, e.g., four corner views in [5]. So, FlexLF[11] was proposed for LF synthesis with sparse input SAIs in varying aperture positions. The angular correlations among SAIs are revealed by building a cost volume to calculate pixel intensity matching errors. After predicting depth by pixel intensity matching errors, depth discontinuity can help locate edges. Liu et al.[12] proposed an edge-aware painting network to complement the preliminary LF for LF angular super-resolution task.

2.2 Novel view synthesis based on single image as input

Single-input novel view synthesis takes a single RGB image as input to synthesize novel views. In the context of LF, Srinivasan et al.[6] made the first attempt to synthesize an LF from a single image by utilizing the image-based rendering (IBR) technique. However, the IBR methods are constrained to Lambertian surfaces and are unable to handle occlusions effectively. Given the high similarity between sub-aperture views, Ruan et al.[13] used a Wasserstein generative adversarial networks (GAN) with a gradient penalty to synthesize complete LF images. Couillaud and Ziou[14] synthesized LF from a single RGB image and depth map using optical geometry and light ray radiometry. Outside the context of LF, single image view synthesis (SynSin)[15] represents a scene by forming feature point clouds that are rotated and rendered at a novel angle. Shih et al.[16] separated a scene into different floating islands (objects with depth discontinuities around the edges), and the occluded regions around the edges of the floating island are painted to avoid showing blankness when rendered to novel angles.

2.3 Novel view synthesis based on multi-view references

Learning is an attractive tool for learning representations of scenes. Volume representation is highly differentiable and can learn complex shapes. Multi-plane images (MPI) is a volume-based approach but with discrete depth planes that help improve efficiency. A recent strand of learning-based research generates MPI for view synthesis in forward-facing scenes, either with single image input[17] or a set of images as input[18–21]. Each input view is expanded into a layered representation that can render high-quality local LF. Mildenhall et al.[18] proposed local light field fusion (LLFF), which can synthesize dense paths of novel views by blending adjacent layered representations together. In addition to the layered scene representation, the neural radiance fields (NeRF) proposed by Mildenhall et al.[22] learns a continuous volumetric scene function and encodes the inward-facing scene into a fully connected deep network. Moreover, Dai et al.[23] transforms point clouds into voxels, and the relative positions among voxelized points can be encoded as descriptors, which are learned and updated by gradients back-propagated from multi-plane rendering.

We summarise the existing view synthesis methods in terms of their input sampling requirements and rendering capability in Table 1. Compared with other methods, ours is flexible in dealing with large baseline sparse inputs with various capturing angles rather than requiring fixed or optimal sampling patterns in conventional novel view synthesis methods.

3 Proposed method

Synthesizing novel views over a wide baseline is challenging and is important for virtual reality systems[28–30]. Accurate depth estimation is one of the most critical assumptions for image warping in novel view synthesis. In order to generate accurate depth in a target view, depth from multiple reference views can be transferred by projections/warping according to camera extrinsic and intrinsic parameters. In our framework shown in Fig. 1, we

Table 1 Comparison with other view synthesis methods. * denotes monocular method.

Method	Input images				Baseline		View rendering		
	Local	Global	Sparse	Dense	Small	Large	Interpolation	Extrapolation	Arbitrary
Synsin[15]	✓*	×	×	×	✓*	×	×	✓	✓
LBVS[6]	✓	×	✓	×	✓	×	✓	×	×
NPtsR[23]	×	✓	×	✓	✓	×	✓	✓	✓
DeepVoxels[24]	×	✓	×	✓	✓	×	✓	✓	✓
ExtViewSyn[25]	×	✓	✓	×	✓	×	✓	✓	×
FlexLF[11]	✓	×	✓	×	✓	×	✓	×	×
LLFF[18]	×	✓	×	✓	✓	×	✓	✓	✓
NeRF[22]	×	✓	×	✓	✓	×	✓	×	✓
Stereo radiance fields (SRF)[26]	×	✓	✓	×	×	✓	✓	✓	✓
MVSNeRF[27]	×	✓	✓	×	×	✓	✓	✓	×
Ours	×	✓	✓	×	×	✓	✓	✓	✓

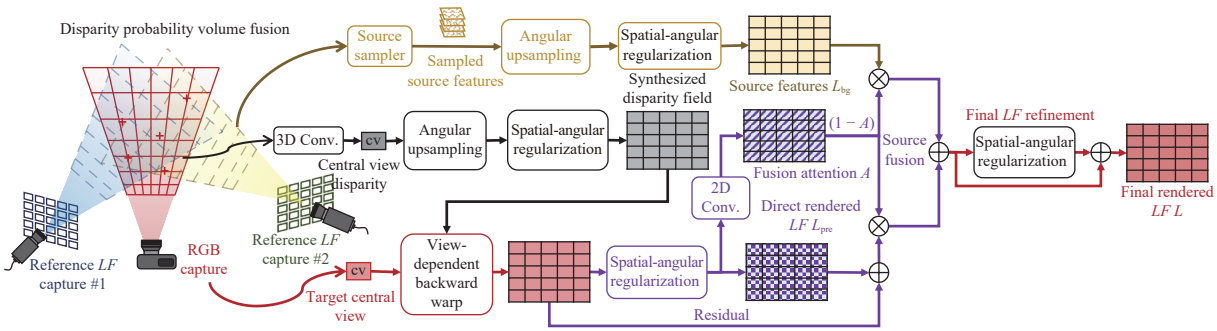


Fig. 1 The overall pipeline of our proposed method. The disparity probability volumes from source LFs are transferred and fused in a target camera. Then, a raw disparity map is generated by the 3D cost volume regularisation and the RGB-D fusion process proposed by [2]. A preliminary LF is synthesized by backward warping pixels of the target view image. The OSS module is proposed to sample and fetch RGB colors from varying source viewpoints to recover the background. Then the preliminary LF is fused with the recovered background weighted by fusion attentions $\mathcal{A}$ produced by the DVF module. The fused LF is further refined by a final spatial-angular regularization module that will render the final outputs.

take LF images as input and estimate the DPV in these source LFs. Then, the DPVs are warped to the target view for an accurate disparity/depth estimation. A preliminary LF is first synthesized by backward warping pixels of the target view image according to the estimated disparity map. The OSS and deep visual fusion (DVF) modules are proposed to fetch known valuable color information from multiple input views to complement the final rendering. Then, the spatial-angular regularisation module is adopted to improve spatial-angular consistency and visual quality.

Following [2, 31–33], we employ a 3D convolutional U-Net to improve the completeness and semantic correctness of the fused DPV $\boldsymbol{V}_{t_0}$. The fused DPV in the target view is converted to a disparity map by the probability weighted compositing methods used in [2, 31–33]. We adopt the multi-scale residual fusion module in [2] that combines visual features from a target view image I_{t_0} to restore the fine details and surface smoothness of the target view's disparity map. The refined disparity map in the target view is denoted as D.

3.1 DPV estimation

The DPV of LF can be estimated by comparing the pixel intensities around a given point O along different EPI lines, e.g., l_1, l_2, and l_3 in Fig. 2. The calculation of pixel intensity variance along the slopes of EPI lines for a candidate depth d' in a query point O is given as (1) in [34]:

$$\sigma_{d'}(O)^2 = \frac{1}{N_u - 1} \sum_{u'} \left[L\left(O + u' \left(1 - \frac{f_0}{d'} \right), u' \right) - L_{d'}(O) \right]^2 \tag{1}$$

where u is the index in angular domain, N_u is the total number of angular views, and f_0 is the focal length. As shown in Fig. 2, the candidate line l_2 that produces the least variance of pixel intensity is taken as the best response, and its slope is proportional to the query point's depth[35].

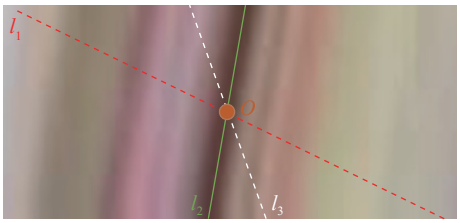


Fig. 2 Candidate EPI lines in point O

3.2 Depth estimation

We leverage the observation from [2] that multi-view disparity probability values $\boldsymbol{V}$ are very informative for accurate disparity estimation when fused together in a target viewpoint. The advantage of warping the DPV over directly warping the depth values is that the fusion of multi-view depth probabilities can generate more accurate and cross-view consistent depth estimations. This is a general approach as other cost volume based depth estimation methods in [32, 36]. In order to warp these DPVs, the scale consistency of DPV for multi-view projections are important. Thus, we adopt the DPV estimation method in [34] and the DPV rescaling and fusion pipeline proposed by [2] that involves rescaling the DPV $\boldsymbol{V}$ to a multi-view scale consistent $\hat{\boldsymbol{V}}$, projecting $\hat{\boldsymbol{V}}$ to a target camera's viewing frustum as $\boldsymbol{V}_{t\to t_0}$ and fusing as $\boldsymbol{V}_{t_0}$:

$$\boldsymbol{V}_{t_0} = \sum_{t=1}^{N_{\text{src}}} \hat{\boldsymbol{V}}_{t\to t_0} \times W_{\text{pos}} \times W_{\text{dir}} \tag{2}$$

where N_{src} is the number of source LF captures, W_{pos} and W_{dir} are separately calculated based on the Euclidean distance between all source and target cameras' positions and directions. The weights W_{pos} and W_{dir} are converted to [0, 1] by softmax operation.

3.3 Preliminary light field synthesis

Our local immersive novel view/LF synthesis starts from synthesizing a preliminary LF by backward warping pixels from the target view image I_{t_0} to novel sub-aperture positions. Then the preliminary LF $\boldsymbol{L}_{\text{pre}}$ is improved and refined by the proposed OSS and DVF modules to reveal occlusions and eliminate noise.

3.3.1 Generating disparity field

Based on the accurate disparity map $D \in \mathbf{R}^{1\times H\times W}$, a cross-view consistent disparity field $\mathcal{D} \in \mathbf{R}^{H\times W\times 2\times M\times N}$ is estimated. By using angular up-sampling layers followed by a pseudo-4D spatial-angular separable convolution network (spatial-angular regularization)[7, 9, 37] to process the disparity map, the spatial and angular consistency among local rendering instances are implicitly regularised:

$$\mathcal{D} = f_{\text{disp}}(D) \tag{3}$$

where $f_{\text{disp}}(\cdot)$ represents the spatial-angular regularization that has the following two advantages[37]: 1) It is memory-efficient compared with 4D convolutions; 2) It alleviates separable filtering in digital signal processing by performing separable 2D spatial and angular convolutions.

Especially, the disparity map D is first processed by an angular up-sampling convolution layer $\text{Conv}_{\text{up\_ang}}$ that increases the channel number of D from 1 to the number of angular dimensions $M \times N$, then followed by a ReLU activation function:

$$D_{\text{up\_ang}} = ReLU(Conv_{\text{up\_ang}}(D)) \tag{4}$$

then, $D_{\text{up\_ang}} \in \mathbf{R}^{M\times N\times 1\times H\times W}$ is processed by a channel up-sampling layer $Conv_{\text{up\_chan}}$ that increases the number of the channel from 1 to C, then followed by a ReLU activation function:

$$D_{\text{up\_chan}} = ReLU(Conv_{\text{up\_chan}}(D_{\text{up\_ang}})) \tag{5}$$

then $D_{\text{up\_chan}} \in \mathbf{R}^{M\times N\times C\times H\times W}$ is reshaped to $D_{\text{ang}} \in \mathbf{R}^{H\times W\times C\times MN}$ for 2D angular convolutions Conv_{ang}:

$$\hat{D}_{\text{ang}} = ReLU(Conv_{\text{ang}}(D_{\text{ang}})) \tag{6}$$

then, the angular filtered disparity $\hat{D}_{\text{ang}}$ is reshaped to $D_{\text{spa}} \in \mathbf{R}^{M\times N\times C\times HW}$ for 2D spatial convolutions $Conv_{\text{spa}}$:

$$\hat{D}_{\text{spa}} = ReLU(Conv_{\text{spa}}(D_{\text{spa}})). \tag{7}$$

The above 2D spatial and angular convolutions are repeated six times, and each layer's network parameters are separately learned. The final spatial and angular regularized output is processed by a residual convolutional layer $Conv_{\text{res}}$ that decreases the number of channels from C to 2 (disparity along x and y dimensions respectively):

$$\mathcal{D} = Conv_{\text{res}}(D_{\text{spa}}). \tag{8}$$

3.3.2 Disparity based pixel warping

A preliminary LF $\boldsymbol{L}_{\text{pre}} \in \mathbf{R}^{W\times H\times 3\times M\times N}$ is synthesized by backward warping pixels from the input target view image I_{t_0} according to the disparity field $\mathcal{D}$:

$$I_v(x) = I_0(x + \mathcal{D}_v) \tag{9}$$

where I_v represents the synthesized v-th sub-aperture view image, and $\mathcal{D}_v$ denotes the disparity map in the v-th sub-aperture view of the target preliminary LF $\boldsymbol{L}_{\text{pre}}$.

Although the disparity field $\mathcal{D}$ preserves geometry and intensity consistencies among angular views, it is still impossible to predict occluded contents based on target view image I_{t_0}. The image quality based on a single capture is also limited without reference to other source captures.

3.4 Occlusion-aware source sampler and deep visual fusion modules

In addition to the scene geometry $\mathcal{D}$, visual features from source LFs $\boldsymbol{L}_{t_s}$ are important for improving the rendering quality of a target LF $\boldsymbol{L}_{t_0}$ for two reasons: first, occluded regions are impossible to be correctly rendered only based on the target central view I_{t_0}. With references from different observation angles of $\boldsymbol{L}_{t_s}$, these occluded visual contents can be located and transferred to the off-center views in the target LF $\boldsymbol{L}_{t_0}$ by the proposed OSS module, as illustrated in Fig. 3; Second, single image capture can be visually noisy. With aligned references from the source captures $\boldsymbol{L}_{t_s}$, the visual quality of the target LF $\boldsymbol{L}_{t_0}$ can be greatly improved via the DVF module.

3.4.1 Plane-sweep volume generation

The PSV is first introduced by [3] to determine pixel correspondences and 3D locations across multiple images. To build a PSV, we first transfer the central views $\{I_{t_s}\}$ of the source LFs to the target view's camera frustum. In theory, the source image I_{t_s} is swept through the volume of the space along the principal axis of the source camera. In practice, the source image I_{t_s} is warped to the target view camera's frustum via homography warping $\mathcal{H}(d)$ according to (10):

$$\mathcal{H}(d) = K_{t_0} \times R_{t_0} \times \left(\boldsymbol{I} - \frac{(\tau_t - \tau_{t_s}) \times n_{t_s}^{\mathrm{T}}}{d} \right) \times R_{t_s}^{\mathrm{T}} \times K_{t_s}^{\mathrm{T}} \quad (10)$$

where $\{K, R\}$ represent camera intrinsic and extrinsic parameters, respectively; d represents the depth plane, $n_{t_s}^{\mathrm{T}}$ denotes the principle axis of the camera frustum; and the subscripts t_0 and t_s denote the index of target and source cameras.

For a better illustration, we separately draw the warped image planes from the source view and the target view in Fig. 3(a) and 3(b), respectively. Due to the different angles of the source and target cameras' principle axes, the aligning direction of warped sweeping image planes from the source view is different from that of the target view. So the image planes warped from the source view to the target view's frustum are oblique in Fig. 3(a). The image planes in the target view are sweeping along the target view camera's principle axis, so the target view's image planes are vertical in Fig. 3(b).

3.4.2 The cost volume estimation

We decompose the PSV into two parts, PSV_{t_s} warped from the source view t_s and PSV_{t_0} warped from the target view t_0. For the purpose of transferring pixels from the source view to the target view, finding the pixel's corresponding relationship between PSV_{t_0} and PSV_{t_s} is important. So, a pixel matching cost volume $\boldsymbol{V}_{\mathrm{cost}}$ is calculated by the difference of pixel intensity between PSV_{t_0} and PSV_{t_s} in their sweeping planes $d \in [1, P]$:

$$\boldsymbol{V}_{\mathrm{cost}} = \{PSV_{t_s}(d) - PSV_{t_0}(d)\}_{d=1}^{P} \quad (11)$$

where d is the index and P is the total number of the sweeping depth planes in PSV. The best matching depth layer for each pixel (x, y) in the PSV_{t_s} and PSV_{t_0} is found by calculating the minimum pixel intensity errors in $\boldsymbol{V}_{\mathrm{cost}}$:

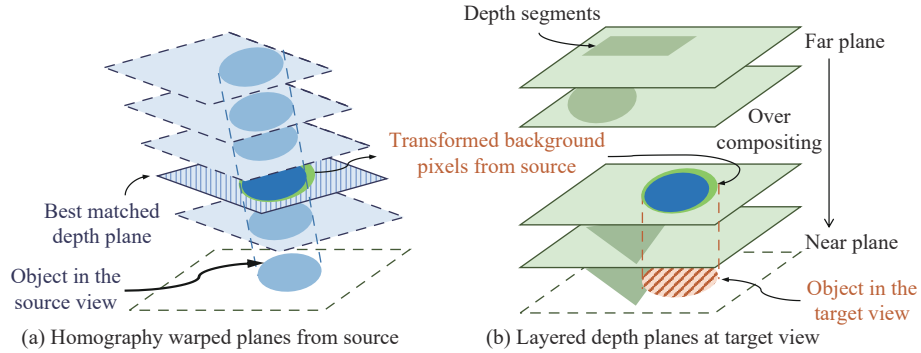
$$S(x, y) = \arg\min_{d} \{\boldsymbol{V}_{\mathrm{cost}}(x, y, d)\} \quad (12)$$

where $S(x, y) = d$ stores the index of depth plane d for pixel (x, y) that is with minimum matching errors between PSV_{t_0} and PSV_{t_s}.

3.4.3 Occlusion-aware source sampling module

We propose an OSS module to extract and transfer the occluded pixels (denoted green in Fig. 3) from PSV_{t_s} to PSV_{t_0}. After the transfer, our objective is to compose a global background image from PSV_{t_0}. The global background image can help improve the visual quality of the final rendered LF.

Because pixels around edges are the most likely to be occluded, an edge attention mask M_{mask} is first generated by calculating discontinuities in the target view's disparity map D. The edge attention mask works as an indicator to help select pixels only around edges for trans-



(a) Homography warped planes from source (b) Layered depth planes at target view

Fig. 3 The OSS module transfers occluded background pixel clues to the target view from the source view's best matched depth plane. The target object is homography warped from the source view's depth plane to the target view. The occluded contents around the object's boundary (depth discontinuities) in the target view are to be replaced by pixels from the source view's best matched depth plane. The final novel view is synthesized via inverse-over compositing.

ferring. The calculation of the attention mask is given in (13):

$$M_{\text{mask}} = \begin{cases} 1, & \text{if } \text{Grad}(D) \geq \mathcal{T} \\ 0, & \text{if } \text{Grad}(D) < \mathcal{T}. \end{cases} \tag{13}$$

Here $\text{Grad}(\cdot)$ is a gradient function to calculate the gradient of the disparity map D. $\mathcal{T}$ is a threshold of the gradients. If the gradient is larger than the given threshold (0.05 in our experiments), the pixel is taken as being around the edges.

To transfer edge pixels indicated by M_{mask} from PSV_{t_s} to PSV_{t_0}, we find the corresponding pixels in each depth layer of PSV_{t_0} and PSV_{t_s} by $S(x, y)$. Hence, the $PSV_{t_0}(x, y, d)$ can be updated. The updating process can be represented as

$$PSV_{t_0}(x, y, d) = PSV_{t_s}(x, y, d) \times M_{\text{mask}} \tag{14}$$

$$S(x, y) = d. \tag{15}$$

Using the updated PSV_{t_0}, a global background I_{bg} can be composed by inverse-over operation[38] of pixels on all depth planes in PSV_{t_0} that contains transferred pixels from the source view. The inverse-over operation can help generate a global background rather than a local background. Because the pixels from nearer layers $d-1$ of PSV_{t_0} can be overwritten by revealed occluded contents from layers further away d.

To implement the inverse-over operation, the compositing algorithm starts from the furthest plane $d = P$ to the nearest depth plane $d = 1$, and the composed output in the nearest/first depth plane is the final composed background image I_{bg}. More specifically, when compositing pixels in a nearer depth plane $d-1$ of PSV_{t_0}, a newer intermediate background image $I_{\text{bg}_{d-1}}$ is composed by fusing the selected pixels from the older background image I_{bg_d} and $PSV_{t_0}(d-1)$. The selection process and conditions are shown in (16),

$$I_{\text{bg}_{d-1}}(x, y) = \begin{cases} I_{\text{bg}_d}(x, y), & \text{if } I_{\text{bg}_d}(x, y) \neq 0 \ \& \ S(x, y) = d \\ \text{PSV}_{t_0}(d-1)(x, y), & \text{elif } (x, y) \in M_{\text{mask}} \ \& \ S(x, y) = d-1 \\ 0, & \text{else} \end{cases} \tag{16}$$

where the $I_{\text{bg}_{d-1}}(x, y)$ will be the pixel from $I_{\text{bg}_d}(x, y)$ or $PSV_{t_0}(x, y, d-1)$ if the (x, y) is with minimum matching errors in depth plane d and also meets other constraints. The finally composed I_{bg} contains the furthest edge pixels, also with minimum matching errors in PSV_{t_0}. The OSS module aims at preserving as many occluded visual features as possible to enable perspective rendering of the target view.

3.4.4 Deep visual fusion module

Subsequently, we have the sampled global background I_{bg} from the source LF captures $\boldsymbol{L}_{t_s}$, which will be first warped into a background LF $\boldsymbol{L}_{\text{bg}}$ using the same method as in Section 3.3. The background LF $\boldsymbol{L}_{\text{bg}}$ is first spatial and angular regularized, then fused with a preliminary LF $\boldsymbol{L}_{\text{pre}}$ that is also spatial-angular regularized. The DVF module learns fusion attention $\mathcal{A}$ between the contents of $\boldsymbol{L}_{\text{pre}}$ and $\boldsymbol{L}_{\text{bg}}$:

$$\boldsymbol{L}_{\text{fuse}} = \mathcal{A} \times f_{\text{LF}}(\boldsymbol{L}_{\text{pre}}) + (1 - \mathcal{A}) \times f_{\text{LF}}(\boldsymbol{L}_{\text{bg}}). \tag{17}$$

The attention mechanism[39–41] is also proven to be able to extract representative features from ambiguous regions. Finally, the fused light field features $\boldsymbol{L}_{\text{fuse}}$ will go through another spatial-angular regularisation module to implicitly regularise the structure of LF contents:

$$\boldsymbol{L} = f_{\text{LF}}(\boldsymbol{L}_{\text{fuse}}) \tag{18}$$

where $f_{\text{LF}}(\cdot)$ represents the spatial angular regularizations, which has the same network structure as $f_{\text{disp}}(\cdot)$.

4 Implementation details

4.1 Training setup

The proposed framework has been implemented with PyTorch. The disparity estimation model and the LF synthesis model were separately trained in two stages. In the first stage, the disparity probability volumes are precalculated and fused in a target camera′s frustum for efficient training. The training of the disparity estimation model initiates the learning rate as 1E–2 and decays by 0.1 since the second epoch. The training of the disparity estimation model needs 128 epochs that take 50 hours to finish on two NVIDIA Tesla V100S GPUs. In the second stage, the OSS and DVF modules are trained, and the learning rate is initialized to 1E–5 and decays by 0.5 since the second epoch. The patch size is set to 128, and number of depth planes P is set to 128 across the disparity range of [–4, 4]. All network parameters are initialized as normal, and the momentum term of the Adam optimizer[42] is set to 0.5. The training needs 148 epochs that take 20 hours on one NVIDIA Tesla V100S GPU.

4.2 The dataset

The Stanford Lytro multi-view light field dataset (MVLF)[4] was used for training and evaluating models. In each scene, there are 3 to 5 LF captures, but without camera parameters and good ground truth disparity maps. Hence, we estimated the camera parameters K, R, and τ by COLMAP[43]. The ground truth of the disparity

maps was estimated by the state-of-the-art LF disparity estimation method introduced in [34]. The proposed pipeline relies on the accuracy of the large-baseline disparity estimation method from [2] that involves volume rescaling, homography warping, and fusion. Due to the limitations of computing memory and the resolution of the disparity estimated from the slope of the EPI line, the number of planes of the DPV is limited. As introduced in LLFF[18], the number of planes in MPI determines the extrapolation boundary. This also applies to homography warped DPV. So, the outdoor scenes with large-baseline views require more planes in DPV than indoor scenes, thus pixels of source views can be accurately allocated into equal-disparity-distant planes of DPV. Therefore, both the disparity estimation algorithm in [2] and our source sampler fail in outdoor scenes that extend out to infinity. After filtering out the outdoor scenes by a threshold of disparity estimation error, 133 scenes are finally reserved, and most are indoor scenes, as expected. We randomly selected 123 scenes for training and ten scenes for model evaluation and comparison. For each rendering instance, we selected two LF captures from each scene and selected the central view of the target LF as I_{t_0}.

4.3 Loss function

Mean square errors (MSE) between the predicted disparity maps D and the ground truth of disparity maps D_g were used to supervise the training of the disparity estimation model given by (19). Mean absolute errors (MAE) between the ground truth LF $\boldsymbol{L}_g$ and the final rendered output LF $\boldsymbol{L}_{t_0}$, the preliminary rendering output based on central view warping $\boldsymbol{L}_{\text{pre}}$, and the features of DVF module $\boldsymbol{L}_{\text{fuse}}$ were calculated to supervise the learning of network parameters according to (20). λ_1 and λ_2 are weight coefficients for the losses of the direct rendering, OSS, and DVF modules, respectively. In our experiments, these two weights are set as 0.2 and 0.1.

$$\mathcal{L}_{\text{dp}} = \|D - D_g\|_2 \tag{19}$$

$$\mathcal{L}_{\text{view}} = \|\boldsymbol{L}_{t_0} - \boldsymbol{L}_g\|_1 + \lambda_1\|\boldsymbol{L}_{\text{pre}} - \boldsymbol{L}_g\|_1 + \lambda_2\|\boldsymbol{L}_{\text{fuse}} - \boldsymbol{L}_g\|_1. \tag{20}$$

5 Experimental results

5.1 Evaluation of view synthesis quality

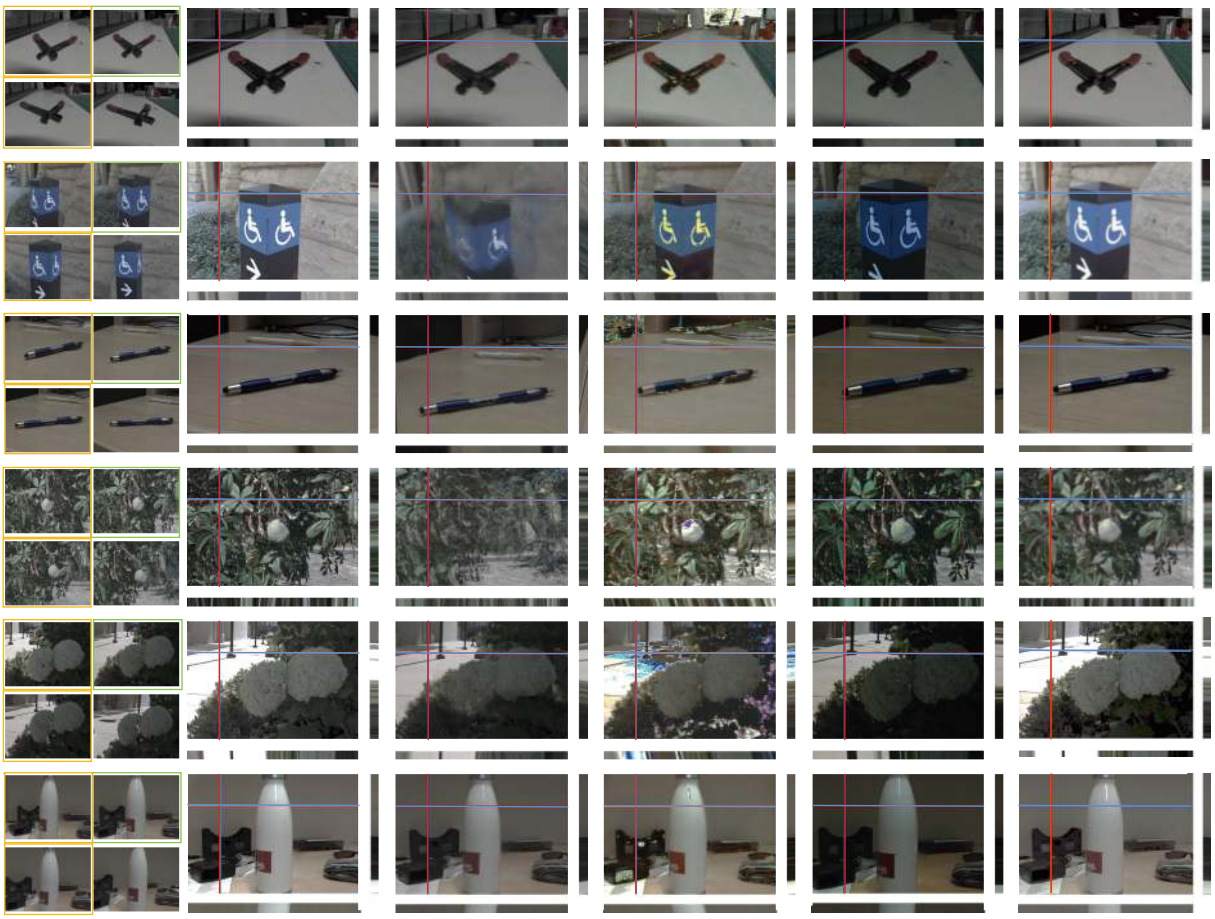
The proposed method is evaluated against the state-of-the-art novel view synthesis methods, including local light field fusion (LLFF)[18], learning based view synthesis (LBVS)[5], and single image view synthesis (SynSin)[15]. Qualitative comparisons are shown in Fig. 4. The LBVS are trained using the MVLF dataset, and the evaluations of SynSin and LLFF use pre-trained models from their official repository. Forty nine novel virtual viewpoints are arranged in parallel planes to a target camera's focal plane, so a set of novel view positions are arranged in a 7×7 array neighbouring the target view. The number of Planes P in PSV is 128. Metrics of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are calculated to evaluate the quantitative quality of view synthesis. As can be seen from Table 2, our method produces higher PSNR results than others and shows competing SSIM results. Note that LBVS uses four corner SAIs of target LF as input, which is a much less challenging scenario for LF view synthesis in terms of angular variations. Our experiments demonstrated that, even when LBVS is doing a much simpler task than ours, our method still produces similar, and most of the time, better results than LBVS, which further validates the efficiency of our model. LLFF generates and fuses neighbouring multiplane images (MPIs) to render novel views, which can adapt to large-baseline parallax inputs. But, it cannot handle the camera rotations well that largely exist in the MVLF dataset, which will directly affect the MPI fusion process. Thus, the LLFF's image rendering quality degraded. One of the approaches most closely related to ours is SRF[26] which is designed for large-baseline spherical-surrounding views. Because our inputs are configured as two source LFs that have tens of SAIs in the micro-baseline, which are too close to each other, thus making it almost uninformative for the multi-view correspondence searching method used in SRF. The correspondences among SAIs can only be effectively established by searching for the minimal pixel intensity variance along the slope of EPI lines, as shown in Fig. 2. Thus, the SRF will have consequently degraded performance on the LF dataset. So, we did not make comparisons with SRF due to unfair inputs.

5.2 Ablation study

We carry out experiments to validate the contributions of the OSS and DVF modules. Table 3 shows quantitative ablations of LF synthesis without the OSS and DVF modules. Full model ours in Table 3 has the best novel view synthesis quality.

In the experiments w/o OSS module, the source pixels sampling is disabled. So was the synthesis of the background LF $\boldsymbol{L}_{\text{bg}}$. This proves that our source pixels sampling approach is important for completing the final rendering results with revealed occluded contents. Fig. 5 shows that occluded contents around depth discontinuities have been successfully recovered. Fig. 6 shows qualitative ablations of the OSS module.

In the experiments w/o DVF module, the fusion attention $\mathcal{A}$ is removed. We can find that the performance



(a) (b) (c) (d) (e) (f)

Fig. 4 Visual quality comparison. From column (a) to (f) are input views (two source views in the left column and target view in the top right), ground truth LF SAI (the most top left, 1st.), results from LLFF, SynSin, LBVS and ours. From the visual results in LLFF (c), the ghost effect is obvious, and the object boundary is unclear; the color output of SynSin (d) has artifacts in the dark region, and its output is blurred; the scene boundary of LBVS (e) is incorrect, many scenes around boundaries are lost.

Table 2 Quantitative evaluation on novel view synthesis quality measured by PSNR and SSIM. Best performance is highlighted in bold; the second best results are underlined

Method scene	LLFF		LBVS		SynSin		Ours	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Batteries	23.50	0.44	<u>38.31</u>	<u>0.95</u>	30.89	0.89	**39.10**	**0.99**
Bottles	23.24	0.47	<u>39.23</u>	<u>0.96</u>	31.03	0.87	**40.13**	**0.99**
Boxes	23.59	0.37	<u>38.23</u>	<u>0.96</u>	29.96	0.83	**39.20**	**0.99**
Cables	23.14	0.39	<u>35.37</u>	<u>0.93</u>	32.50	0.91	**39.81**	**0.98**
Cups	23.50	0.43	<u>38.65</u>	0.96	37.89	<u>0.98</u>	**39.02**	**0.99**
Flowers	28.48	0.77	<u>37.09</u>	<u>0.97</u>	28.47	0.77	**38.95**	**0.99**
Leaves	22.93	0.12	<u>34.84</u>	<u>0.96</u>	27.60	0.78	**35.10**	**0.97**
Pens	23.70	0.48	<u>38.89</u>	<u>0.96</u>	32.10	0.91	**39.71**	**0.99**
Signs	23.00	0.36	<u>38.14</u>	<u>0.97</u>	30.80	0.90	**39.52**	**0.99**
Tools	23.06	0.50	<u>38.29</u>	<u>0.96</u>	31.63	0.90	**40.53**	**0.99**
Average	23.81	0.47	<u>37.70</u>	<u>0.96</u>	30.61	0.86	**39.10**	**0.99**

drops without the DVF module. The degraded performance proves that the attention mechanism in the DVF module is important for accurately fusing background LF $\boldsymbol{L}_{\text{bg}}$ and preliminary LF $\boldsymbol{L}_{\text{pre}}$. We visually compare the

Table 3 Ablation study. The best performance is highlighted in bold; the second best results are underlined.

Method	PSNR↑	SSIM↑
w/o OSS	37.53	0.96
w/o DVF	<u>38.27</u>	<u>0.98</u>
Full model ours	**39.10**	**0.99**

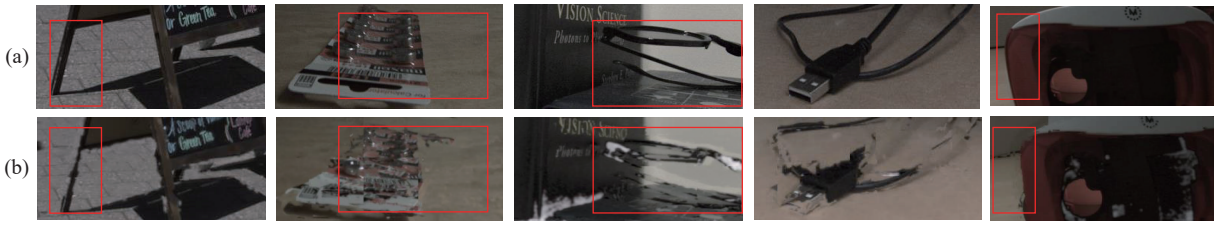


Fig. 5 Examples of revealed background content. The images in row (a) are ground truth images, the images in row (b) are with occlusion removed. We can observe that occlusions around depth discontinuities are successfully replaced by background contents.

Fig. 6 Images from (a) to (c) are ground truth, output images from the model without OSS module, and with the OSS module.

output from direct rendering based on the backward warping of central views in Fig. 7(b) and the final output LF image in Fig. 7(c). Fig. 7(c) has much less noise than Fig. 7(b). Therefore, the visualization in Fig. 7 proves that the DVF module can further suppress noise in output images, the effectiveness of attention-guided CNN for image denoising was also validated in a previous study[44]. Fig. 8 further proves that the DVF module is important in complementing the background LF.

6 Limitation

Our method is configured as a multi-view LF framework that adopts methodologies from multi-view stereo techniques, such as homography warping and pixel-intensity-based cost volume estimation. Therefore, our method has inherited limitations just like other multi-view stereo algorithms[32, 33, 45].

First, our method suffers from low-texture regions. Compared to other cost volume estimation methods in [32, 33, 45] that use convolution neural networks to extract high-level features, our pixel-consistency based cost volume estimation method in the OSS module could be less robust in low-texture regions.

Second, our method cannot handle large-baseline images of outdoor scenes well; our method uses DPV, which has a limited number of planes to transfer disparity from source views to the target view. The limited number of

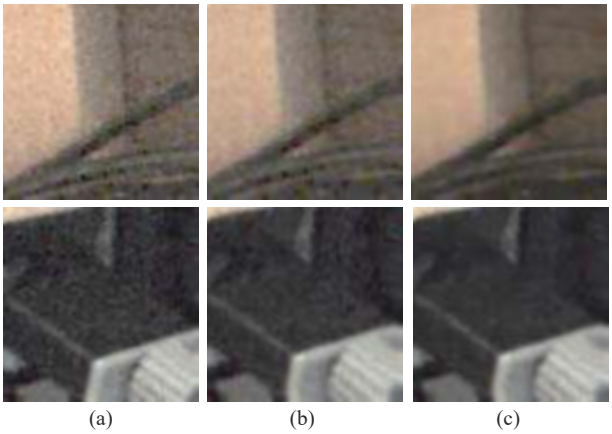


Fig. 7 Visualizations of outputs in different stages. This figure visualizes a close-up of the (0, 0)-th SAI in the target view. (a) shows ground truth LF image; (b) shows a warped initial LF image; (c) shows the output LF image refined by DVF.

disparity planes in DPV is inefficient in resolution for large-baseline outdoor scenes with a large depth range. Additionally, the performance of the OSS module is also affected by the inefficient number of planes limited by computing memory. Therefore, our method does not perform well on the scene flowers and leaves.

Additionally, the noise suppression brought by the DVF module may further result in the loss of high-frequency details in output images.

7 Conclusions

Rendering a local immersive LF based on arbitrary large baseline references is a challenging task. Our method takes large baseline LF captures as input and can synthesize immersive novel views in a novel target viewpoint. Conventional view synthesis methods require a small baseline or hundreds of dense input views, while ours only requires two LF captures, which are convenient with existing commercial LF cameras. Furthermore, the OSS and

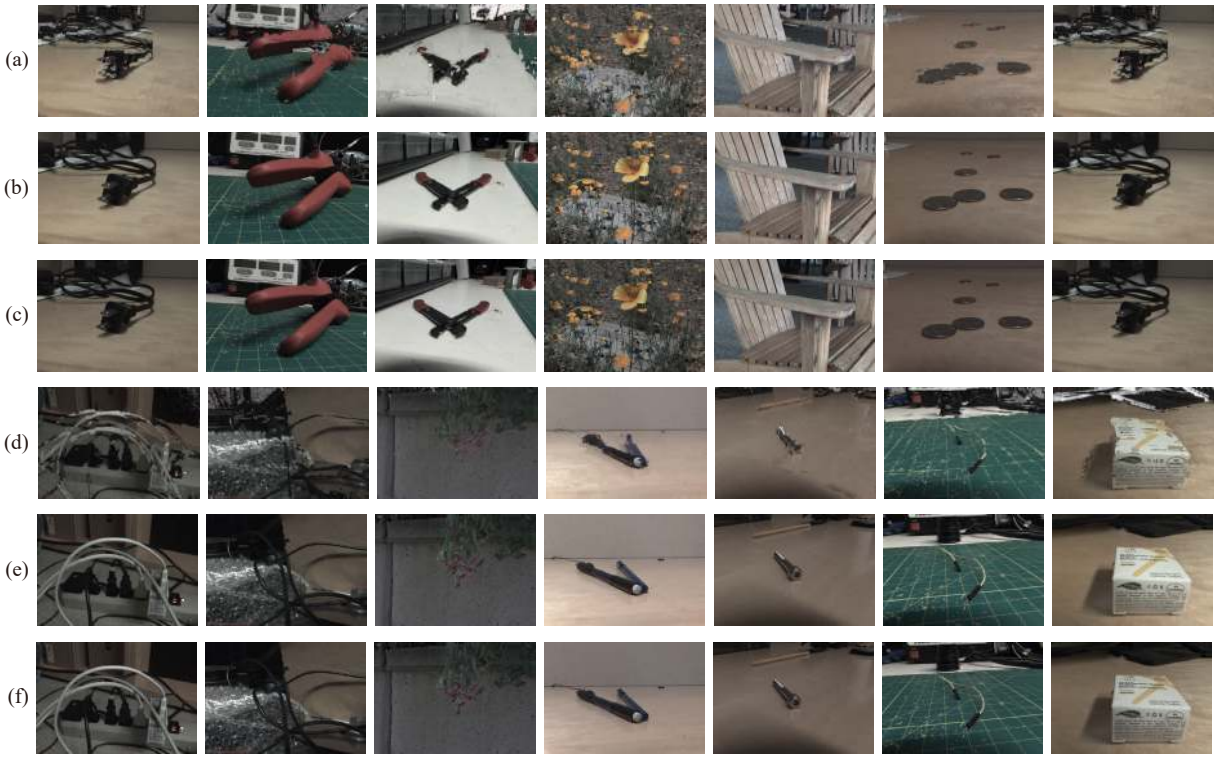


Fig. 8 Ablation study of DVF module. The images in rows (c) and (f) are ground truth images, and the images in row (a) and (d) are without the DVF module. The images in rows (b) and (e) are outputs with the DVF module.

DVF modules are proposed to fuse sampled occluded source features into a final refined LF. Such source sampling and fusion mechanisms not only help provide occlusion information from varying observation angles, but also prove to be able to effectively enhance the visual quality by suppressing sensor noise. Experimental results show that our proposed method is able to render high-quality LF images with sparse LF references and significantly outperforms the other state-of-the-art LF rendering and novel view synthesis methods.

Acknowledgements

The research was supported by the Theme-based Research Scheme, Research Grants Council of Hong Kong (No. T45-205/21-N). The authors would like to thank NVIDIA AI Technology Center (NVAITC) for the GPU computing resources.

Open Access

References

[1] R. C. Bolles, H. H. Baker, D. H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, vol. 1, no. 1, pp. 7–55, 1987. DOI: 10.1007/BF00128525.

[2] W. P. Xing, J. Chen, Z. F. Yang, Q. Wang, Y. K. Guo. Scale-consistent fusion: From heterogeneous local sampling to global immersive rendering. *IEEE Transactions on Image Processing*, vol. 31, pp. 6109–6123, 2022. DOI: 10.1109/TIP.2022.3205745.

[3] R. T. Collins. A space-sweep approach to true multi-image matching. In *Proceedings of CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, pp. 358–363, 1996. DOI: 10.1109/CVPR.1996.517097.

[4] D. G. Dansereau, B. Girod, G. Wetzstein. LiFF: Light field features in scale and depth. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 8034–8043, 2019. DOI: 10.1109/CVPR.2019.00823.

[5] N. K. Kalantari, T. C. Wang, R. Ramamoorthi. Learning-

based view synthesis for light field cameras. *ACM Transactions on Graphics*, vol. 35, no. 6, Article number 193, 2016. DOI: 10.1145/2980179.2980251.

[6] P. P. Srinivasan, T. Z. Wang, A. Sreelal, R. Ramamoorthi, R. Ng. Learning to synthesize a 4D RGBD light field from a single image. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 2262–2270, 2017. DOI: 10.1109/ICCV.2017.246.

[7] Y. L. Wang, F. Liu, Z. L. Wang, G. Q. Hou, Z. A. Sun, T. N. Tan. End-to-end view synthesis for light field imaging with pseudo 4DCNN. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 340–355, 2018. DOI: 10.1007/978-3-030-01216-8_21.

[8] G. C. Wu, M. D. Zhao, L. Y. Wang, Q. H. Dai, T. Y. Chai, Y. B. Liu. Light field reconstruction using deep convolutional network on EPI. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 1638–1646, 2017. DOI: 10.1109/CVPR.2017.178.

[9] H. W. F. Yeung, J. H. Hou, J. Chen, Y. Y. Chung, X. M. Chen. Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 138–154, 2018. DOI: 10.1007/978-3-030-01231-1_9.

[10] Z. T. Zhang, Y. B. Liu, Q. H. Dai. Light field from micro-baseline image pair. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 3800–3809, 2015. DOI: 10.1109/CVPR.2015.7299004.

[11] J. Jin, J. H. Hou, J. Chen, H. Q. Zeng, S. Kwong, J. Y. Yu. Deep coarse-to-fine dense light field reconstruction with flexible sampling and geometry-aware fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 1819–1836, 2022. DOI: 10.1109/TPAMI.2020.3026039.

[12] X. Liu, M. H. Wang, A. Z. Wang, X. Y. Hua, S. S. Liu. Depth-guided learning light field angular super-resolution with edge-aware inpainting. *The Visual Computer*, vol. 38, no. 8, pp. 2839–2851, 2022. DOI: 10.1007/s00371-021-02159-6.

[13] L. Y. Ruan, B. Chen, M. L. Lam. Light field synthesis from a single image using improved Wasserstein generative adversarial network. In *Proceedings of the 39th Annual European Association for Computer Graphics Conference: Posters*, Delft, The Netherlands, pp. 19–20, 2018.

[14] J. Couillaud, D. Ziou. Light field variational estimation using a light field formation model. *The Visual Computer*, vol. 36, no. 2, pp. 237–251, 2020. DOI: 10.1007/s00371-018-1599-2.

[15] O. Wiles, G. Gkioxari, R. Szeliski, J. Johnson. SynSin: End-to-end view synthesis from a single image. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 7465–7475, 2020. DOI: 10.1109/CVPR42600.2020.00749.

[16] M. L. Shih, S. Y. Su, J. Kopf, J. B. Huang. 3D photography using context-aware layered depth inpainting. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 8025–8035, 2020. DOI: 10.1109/CVPR42600.2020.00805.

[17] R. Tucker, N. Snavely. Single-view view synthesis with multiplane images. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 548–557, 2020. DOI: 10.1109/CVPR42600.2020.00063.

[18] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, A. Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics*, vol. 38, no. 4, Article number 29, 2019. DOI: 10.1145/3306346.3322980.

[19] A. Jain, M. Tancik, P. Abbeel. Putting NeRF on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 5865–5874, 2021. DOI: 10.1109/ICCV48922.2021.00583.

[20] W. P. Xing, J. Chen. NEX.+: Novel view synthesis with neural regularisation over multi-plane images. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, pp. 1581–1585, 2022. DOI: 10.1109/ICASSP43922.2022.9746938.

[21] W. P. Xing, J. Chen. Temporal-MPI: Enabling multi-plane images for dynamic scene modelling via temporal basis learning. In *Proceedings of the 17th European Conference on Computer Vision*, Springer, Tel Aviv, Israel, pp. 323–338, 2022. DOI: 10.1007/978-3-031-19784-0_19.

[22] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 405–421, 2020. DOI: 10.1007/978-3-030-58452-8_24.

[23] P. Dai, Y. D. Zhang, Z. W. Li, S. C. Liu, B. Zeng. Neural point cloud rendering via multi-plane projection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 7827–7836, 2020. DOI: 10.1109/CVPR42600.2020.00785.

[24] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, M. Zollhöfer. DeepVoxels: Learning persistent 3D feature embeddings. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 2437–2446, 2019. DOI: 10.1109/CVPR.2019.00254.

[25] I. Choi, O. Gallo, A. Troccoli, M. H. Kim, J. Kautz. Extreme view synthesis. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 7780–7789, 2019. DOI: 10.1109/ICCV.2019.00787.

[26] J. Chibane, A. Bansal, V. Lazova, G. Pons-Moll. Stereo radiance fields (SRF): Learning view synthesis for sparse views of novel scenes. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 7907–7916, 2021. DOI: 10.1109/CVPR46437.2021.00782.

[27] A. P. Chen, Z. X. Xu, F. Q. Zhao, X. S. Zhang, F. B. Xiang, J. Y. Yu, H. Su. MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of IEEE/CVF International Conference on Com-*

puter Vision, IEEE, Montreal, Canada, pp. 14104–14113, 2021. DOI: 10.1109/ICCV48922.2021.01386.

[28] L. Liu, Z. Y. Wang, Y. Liu, C. Xu. An immersive virtual reality system for rodents in behavioral and neural research. *International Journal of Automation and Computing*, vol. 18, no. 5, pp. 838–848, 2021. DOI: 10.1007/s11633-021-1307-y.

[29] N. N. Zhou, Y. L. Deng. Virtual reality: A state-of-the-art survey. *International Journal of Automation and Computing*, vol. 6, no. 4, pp. 319–325, 2009. DOI: 10.1007/s11633-009-0319-9.

[30] W. P. Xing, J. Chen. MVSPlenOctree: Fast and generic reconstruction of radiance fields in PlenOctree from multi-view stereo. In *Proceedings of the 30th ACM International Conference on Multimedia*, Lisboa, Portugal, pp. 5114–5122, 2022. DOI: 10.1145/3503161.3547795.

[31] Y. Yao, Z. X. Luo, S. W. Li, T. W. Shen, T. Fang, L. Quan. Recurrent MVSNet for high-resolution multi-view stereo depth inference. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 5520–5529, 2019. DOI: 10.1109/CVPR.2019.00567.

[32] Y. Yao, Z. X. Luo, S. W. Li, T. Fang, L. Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision*, Springer, Munich, Germany, pp. 785–801, 2018. DOI: 10.1007/978-3-030-01237-3_47.

[33] R. Chen, S. F. Han, J. Xu, H. Su. Point-based multi-view stereo network. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 1538–1547, 2019. DOI: 10.1109/ICCV.2019.00162.

[34] J. Chen, J. H. Hou, Y. Ni, L. P. Chau. Accurate light field depth estimation with superpixel regularization over partially occluded regions. *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4889–4900, 2018. DOI: 10.1109/TIP.2018.2839524.

[35] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, P. Hanrahan. Light Field Photography with A Hand-Held Plenoptic Camera, Ph.D. dissertation, Department of Computer Science, Stanford University, USA, 2005.

[36] Z. H. Yu, S. H. Gao. Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 1946–1955, 2020. DOI: 10.1109/CVPR42600.2020.00202.

[37] H. W. F. Yeung, J. H. Hou, X. M. Chen, J. Chen, Z. B. Chen, Y. Y. Chung. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2319–2330, 2019. DOI: 10.1109/TIP.2018.2885236.

[38] T. Porter, T. Duff. Compositing digital images. *ACM SIGGRAPH Computer Graphics*, vol. 18, no. 3, pp. 253–259, 1984. DOI: 10.1145/964965.808606.

[39] K. Y. Luo, T. Guan, L. L. Ju, Y. S. Wang, Z. Chen, Y. W. Luo. Attention-aware multi-view stereo. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 1587–1596, 2020. DOI: 10.1109/CVPR42600.2020.00166.

[40] P. H. Chen, H. C. Yang, K. W. Chen, Y. S. Chen. MVS-Net++: Learning depth-based attention pyramid features for multi-view stereo. *IEEE Transactions on Image Processing*, vol. 29, pp. 7261–7273, 2020. DOI: 10.1109/TIP.2020.3000611.

[41] X. D. Zhang, Y. T. Hu, H. C. Wang, X. B. Cao, B. C. Zhang. Long-range attention network for multi-view stereo. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, USA, pp. 3781–3790, 2021. DOI: 10.1109/WACV48630.2021.00383.

[42] D. P. Kingma, J. Ba. Adam: A method for stochastic optimization. [Online], Available: https://arxiv.org/abs/1412.6980, 2014.

[43] J. L. Schönberger, J. M. Frahm. Structure-from-motion revisited. In *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 4104–4113, 2016. DOI: 10.1109/CVPR.2016.445.

[44] C. W. Tian, Y. Xu, Z. Y. Li, W. M. Zuo, L. K. Fei, H. Liu. Attention-guided CNN for image denoising. *Neural Networks*, vol. 124, pp. 117–129, 2020. DOI: 10.1016/j.neunet.2019.12.024.

[45] Y. Yao, Z. X. Luo, S. W. Li, J. Y. Zhang, Y. F. Ren, L. Zhou, T. Fang, L. Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of IEEE/CVF Conference Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 1787–1796, 2020. DOI: 10.1109/CVPR42600.2020.00186.

Wenpeng Xing received the B. Eng. degree in civil engineering from Harbin Institute of Technology, China in 2017, the G. Dip. in computer engineering from University of Limerick, Ireland in 2019. He is currently a Ph. D. degree candidate in computer science at Department of Computer Science, Hong Kong Baptist University, China.

His research interests include computational imaging, 3D content capture, modelling and rendering.

E-mail: cswpxing@comp.hkbu.edu.hk

ORCID iD: 0000-0001-5848-9417

Jie Chen received the B. Sc. degree in Opto-Information science and engineering and the M. Eng. degree in optoelectronic information engineering both from School of Optical and Electronic Information, Huazhong University of Science and Technology, China in 2008 and 2011, respectly, and the Ph. D. degree in information engineering from School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore in 2016. He is currently an assistant professor at Department of Computer Science, Hong Kong Baptist University, China. He worked as a post-doctoral research fellow at ST Engineering-NTU Corporate Laboratory, Singapore, and then as a senior algorithm engineer at OmniVision Technologies Inc. He currently serves as an Associate Editor for The *Visual Computer*, Springer.

His research interests include computational photography (light fields, high dynamic range imaging, hyperspectral ima-

ging and computational tomography), multimedia signal capture, reconstruction and content generation (3D vision, motion and music), AI for art-tech and humanities.

E-mail: chenjie@comp.hkbu.edu.hk (Corresponding author)
ORCID iD: 0000-0001-8419-4620

Yike Guo received the B.Sc. degree (Hons.) in computing science from Tsinghua University, China in 1985, and the Ph.D. degree in computational logic from the Imperial College London, UK in 1994. He is currently a Chair professor in Department of Computer Science and Engineering, the Hong Kong University of Science and Technology, where he also serves as the provost since December, 2022. He is professor of computing science in Department of Computing at Imperial College London (2002 – now, on leave since 2020 January). Between 2020 and 2022, he was a professor in Department of Computer Science, Hong Kong Baptist University, China, where he served as the vice-president (research and development).

He has been working on technology and platforms for scientific data analysis since the mid-1990s, where his research focuses on knowledge discovery, data mining, and large-scale data management. He founded InforSense, a software company for life science and health care data analysis, and served as the CEO for several years before the company′s merger with IDBS, a global advanced R&D software provider, in 2009. He is also the founding director of the Data Science Institute, Imperial College, UK, where he also led Discovery Science Group. He also holds the position of CTO of the tranSMART Foundation, a global open-source community using and developing data sharing and analytics technology for translational medicine. He has contributed to numerous major research projects, including the UK EPSRC platform project, Discovery Net; the Wellcome Trust-funded Biological Atlas of Insulin Resistance (BAIR); and the European Commission U-BIOPRED project. He is also the principal investigator of the European Innovative Medicines Initiative eTRIKS project, a 23 Million Euros project that is building a cloud-based informatics platform, in which tranSMART is a core component for clinicogenomic medical research, and the co-investigator of Digital City Exchange, a 5.9 Million GBP research program, exploring ways to digitally link utilities and services within smart cities. Since 2021, he is the principal coordinator of the 52.8 Million HKD project funded by Hong Kong Research Grants Council which investigates AI-based symbiotic creativity for Art-Tech. He has published over 250 articles, papers, and reports. Projects he has contributed to have been internationally recognized, including winning the "Most Innovative Data Intensive Application Award" at the Supercomputing 2002 conference for Discovery Net, the Bio-IT World " Best Practices Award" for U-BIOPRED in 2014 and the " Best Open Source Software Award" from ACM SIGMM in 2017. He is a Fellow of Royal Academy of Engineering (FREng), Member of Academia Europaea (MAE), Fellow of British Computer Society (FBCS), Fellow of Hong Kong Academy of Engineering Sciences (FHKEng).

His research interests include AI for healthcare, data mining and art.

E-mail: yikeguo@ust.hk; yg@doc.ic.ac.uk
ORCID iD: 0000-0002-3075-2161