Light Field View Synthesis via Aperture Disparity and Warping Confidence Map

Nan Meng, Kai Li, Jianzhuang Liu, *Senior Member, IEEE*, and Edmund Y. Lam, *Fellow, IEEE*

***Abstract*— This paper presents a learning-based approach to synthesize the view from an arbitrary camera position given a sparse set of images. A key challenge for this novel view synthesis arises from the reconstruction process, when the views from different input images may not be consistent due to obstruction in the light path. We overcome this by jointly modeling the epipolar property and occlusion in designing a convolutional neural network. We start by defining and computing the aperture disparity map, which approximates the parallax and measures the pixel-wise shift between two views. While this relates to free-space rendering and can fail near the object boundaries, we further develop a warping confidence map to address pixel occlusion in these challenging regions. The proposed method is evaluated on diverse real-world and synthetic light field scenes, and it shows better performance over several state-of-the-art techniques.**

***Index Terms*— View synthesis, image-based rendering, light field, aperture flow, epipolar property, confidence map.**

I. Introduction

PERCEIVING the 3-dimensional (3D) nature of an object is a basic human instinct, but it can be very difficult for a computer. One reason is that the optical system cannot easily capture the geometric information of the scene, and therefore it often has difficulty recreating the visual perception [1]. Ordinarily, an image is not informative about the differences among light rays coming from various directions, as they are combined together to form the intensity at a pixel location. These differences, however, are crucial for us to perceive the world [2].

With advanced techniques and imaging setups, recording the 3D information of an object is becoming feasible [3]. One of the most promising techniques is light field photography, which uses a plenoptic camera to capture both the directions and radiance of the incident light rays [4]. The additional directional information allows a wider range of vision applications, such as depth estimation [5], [6], rendering [7], [8], refocusing [9], super-resolution [10], [11] etc. However, there is a limit to the density of the pixels that one can capture, necessitating a trade-off between spatial and angular resolution [12].

One way to relieve such a trade-off is to produce the intermediate views from the captured images. Ordinarily, approaches for view synthesis require geometric knowledge as priors in the recovery of a dense light field from the sparsely-sampled inputs [13], [14]. The new view can be reconstructed from any acquired images by projecting the pixels to proper 3D locations and re-projecting them onto the target picture [7]. However, such methods struggle with complex regions such as occlusions, as well as transparent and semi-transparent surfaces, where the depth information is difficult to calculate or estimate. Without accurate geometry information, the generated results often contain jarring artifacts [15], [16].

An alternative approach is image-based rendering (IBR), which directly reuses the pixels from the available images to produce the new views. For light field, the depth information is not necessary for the rendering process [17]. The spatio-angular redundancy makes it possible to infer the novel view from neighboring sub-aperture images (SAIs). One advantage of the IBR techniques is that they can avoid explicitly modeling the realistic geometry of the scene. However, such approaches usually require more samples to counter the undesirable aliasing effects in the outputs [7], [18].

These difficulties recently have led to investigations using the learning-based approach, which forgoes the explicit modeling of the problem and instead makes use of deep learning to approximate the ground truth with many training samples [15], [19]. The powerful representation ability of convolutional neural network (CNN) and its widespread success in vision tasks make this a promising direction [20], [21]. Generally, the CNN-based view synthesis algorithms are able to achieve relatively high-quality reconstruction, but often require the ground truth to acquire the accurate supervision signal [12], [22], which restricts the generalization or capacity of the algorithms. For instance, Meng *et al.* [23] and Yeung *et al.* [24] both adopt the learning framework to directly approximate the ground truth. However, the drawback is that they can only generate the views that are recorded in the labels.

Recently, multi-plane image (MPI), which is originally proposed as a scene representation for stereo imagery [25], has attracted increasing attention in light field imaging [26], [27]. One example is the pipeline proposed by Mildenhall *et al.* [28], which learns the MPI representations of each input image with a 3D CNN. They are subsequently warped and blended to

Manuscript received September 3, 2020; revised January 24, 2021 and March 2, 2021; accepted March 9, 2021. Date of publication March 22, 2021; date of current version March 29, 2021. This work was supported in part by the Research Grants Council of Hong Kong under Grant GRF 17201818, Grant 17200019, and Grant 17201620 and in part by The University of Hong Kong under Grant 104005438 and Grant 104005864. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiwen Lu. *(Corresponding author: Nan Meng.)*

Nan Meng and Edmund Y. Lam are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: nanmeng@hku.hk; elam@eee.hku.hk).

Kai Li is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: kai.li@sjtu.edu.cn).

Jianzhuang Liu is with Noah's Ark Lab, Huawei Technologies Company Ltd., Shenzhen 518129, China (e-mail: liu.jianzhuang@huawei.com).

Digital Object Identifier 10.1109/TIP.2021.3066293

reconstruct the target view. Meanwhile, a similar framework is developed by Flynn *et al.* [27], but they adopt learned gradient descent to extract the MPIs. A major benefit of such a layered representation is that it can be reused to reconstruct multiple views at arbitrary camera positions.

The idea of continuous view synthesis has also been explored in video frame interpolation [29]. Typically, one estimates the optical flow and uses it to warp input frames to produce the interpolation samples [30]. For light field, the warping is usually based on the estimated disparity map. A recent method utilizes fused-pixel and feature-based (a.k.a. FPFR) [31] information follows this idea, but the difference is that it warps both the input images and features. In doing so, the model consists of two branches, i.e. one for pixel-based reconstruction and the other for feature-based reconstruction, and the target image is synthesized by combining the intermediate outputs from both.

In this work, we focus on the inherent epipolar property of the light field, and explicitly model the relations among the disparity maps obtained from various SAIs. According to the structural property of the light field, we approximate the relationship between disparity value and changes in the angular viewpoint position with a linear relation. This assumption allows us to calculate the intermediate disparity map between the input and target views, which can be further exploited to warp the input images. We name the matrix measuring the pixel shift between different views of light field the *aperture disparity map* (ADM) to emphasize that there is a relationship between the value of ADM and the aperture position. However, in practice, such linearity assumption may not always hold for real-world light fields, due to many physical and environmental factors, such as lens distortion of the camera, chaotic environmental light rays, and non-Lambertian object surfaces. To compensate for the defects and address the pixel occlusion issue, we further estimate the warping confidence maps (WCMs), which equilibrate the radiance information from different input views, to produce the final image. Finally, the coarse results are further refined by a four-dimensional (4D) CNN with alternating filters [24].

In summary, the contributions of this paper are as follows:

- We propose a novel disparity model known as ADM, that is tailored to the light field images to measure the pixel-wise shift distance between a given pair of images. Experimental results show that it works well on both real-world and synthetic scenes.
- We introduce the WCM to combine the pixel values from different views for the target image. It can efficiently handle occluded pixels, and therefore reduce the artifacts near the object boundaries.

II. Related Work

View synthesis and IBR are closely related. Generally, the explicit geometry models are not necessary for IBR when generating new images in a light field, while view synthesis usually requires both geometry information and a few images to provide the virtual views. In this section, we review the literature of IBR algorithms, view synthesis, and recent learning-based approaches in the context of light field imaging.

Early IBR algorithms for light field rely on the characterization of the plenoptic function and treat the creation of new views as resampling [32], [33]. This approach ignores occlusion, and thus is only feasible for free space rendering or for producing the views reasonably close to the original ones. It gradually becomes clear that interpolation of plausible views in high quality requires either intensive sampling or knowledge about the scene [7], [34]. Therefore, a different set of approaches to light field rendering attempt to infer at least some geometry information, which include methods that rely on image registration [35], prior knowledge [36], or image correspondence and warping [37], to name a few.

View synthesis methods, on the other hand, focus on making use of explicit geometric knowledge that assists in the recovery of a dense light field. Some enforce explicit priors on the light field itself, such as sparsity in the Fourier domain [38] or shearlet transform domain [39], a patch-based Gaussian mixture model [40], or Lambertian surfaces with modest depth discontinuities [41]. However, these methods require either a specific sampling pattern or a large number of views, which limit their practical uses. Other techniques involve partial reconstruction of the scene geometry, such as a global 3D reconstruction [42] or a soft model of the geometrical relationships [43]. Some methods infer the geometry by estimating the disparity for a single view [44] or for each input view [14]. Given that an accurate depth estimation is hard to obtain, such approaches often struggle with complex scenes.

Another group of synthesis algorithms do not require explicit geometric models but rely on the feature correspondence between images. The classical approaches of this kind interpolate the intermediate views by exploiting the optical flow [45], [46]. However, given that non-Lambertian surfaces and occlusions are still challenging to the flow estimation methods [47], [48], the interpolated views tend to have artifacts near the object boundaries. Wang *et al.* [49] make use of the images taken from another standard camera as references to generate plausible frames for the light field video, which also increases the complexity of the imaging system.

More recently, learning-based methods come into the spotlight due to their effectiveness on vision tasks [20], [50]. According to the way learning is involved, such methods can generally be divided into two groups. The first one attempts to establish a direct mapping using learning frameworks between the sparsely-sampled inputs and their dense correspondence. The difference among various techniques lies in the reconstruction level. For instance, Gul and Gunturk [51] handle the pixel-wise reconstruction. LFCNN [52] and Wang *et al.* [53] are aperture-wise methods. Wu *et al.* [22] explore an approach to recover the epipolar plane image (EPI), while Meng *et al.* [12] and Yeung *et al.* [24] directly restore the entire light field. The second group embeds learning in the traditional rendering pipeline. Kalantari *et al.* [16] make one of the early attempts to adopt two CNNs for disparity estimation and color prediction. Meanwhile, Srinivasan *et al.* [54] propose a two-stage learning process to estimate scene geometry and eliminate occlusions.

Generally, the learning-based methods produce more plausible visual results, but they also require a large amount of

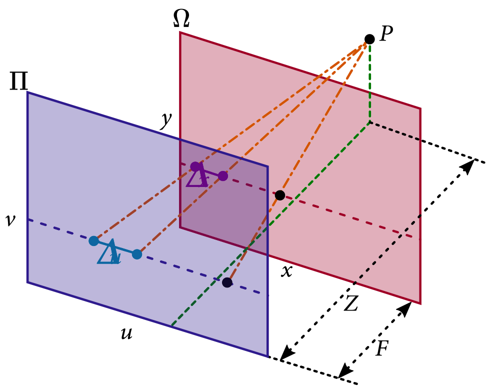


Fig. 1. Illustration of pixel shift from different viewpoints in the two-plane parametrization of light field imaging.

training data paired with labels. To overcome such a problem, Chen *et al.* [55] come up with a self-supervised approach by fine-tuning a video interpolation framework based on cycle consistency. Gao *et al.* [56] employ a CNN to restore EPI coefficients in the shearlet domain. In addition to the data volume, for many prevailing learning methods, the rigid learning strategy heavily relies on the training data, consequently restricting generalization of the model. The end-to-end training pattern makes many models hard to reconstruct an image from a viewpoint that has not been recorded in the ground truth [23], [24].

III. Method

We adopt the two-plane parametrization [57] to represent the 4D light field. Each light ray is represented by the intersections with two parallel planes transmitting from the spatial coordinate $\boldsymbol{x} = (x, y)$ to the angular coordinate $\boldsymbol{u} = (u, v)$, and thus denoted by $L := L(\boldsymbol{x}, \boldsymbol{u})$. We assume that all the coordinate variables in the plenoptic function $L(\boldsymbol{x}, \boldsymbol{u})$ are continuous. Therefore, the goal of IBR is to reconstruct such a function based on a set of discrete samples $L(\boldsymbol{x}_i, \boldsymbol{u}_j)$ $(i, j \in \mathbb{N})$,

$$L(\boldsymbol{x}_i, \boldsymbol{u}_j) \xrightarrow{g} L(\boldsymbol{x}, \boldsymbol{u}), \tag{1}$$

where $g(\cdot)$ denotes the reconstruction algorithm.

A. Spatio-Angular Relationship

For any given point P in the free space, the two-plane parametrization is depicted in Fig. 1, where F is the orthogonal distance between the two parallel planes, and Z is the depth of the focused point. The pixel shifts in different views can be inferred using similar triangles. If we vary $\boldsymbol{x}$ with a distance $\Delta\boldsymbol{x}$, the angular coordinate has to change according to

$$\Delta\boldsymbol{x} = \frac{Z - F}{Z}\Delta\boldsymbol{u} = \frac{\alpha F - F}{\alpha F}\Delta\boldsymbol{u} = \left(1 - \frac{1}{\alpha}\right)\Delta\boldsymbol{u}, \tag{2}$$

where $\alpha = \frac{Z}{F}$ denotes the disparity ratio. $\Delta\boldsymbol{u}$ is the distance between two viewpoints located at the camera plane Π. In the following, we pay more attention to the appearance of the view, and for the sake of illustration, we use the symbol $\mathcal{L}_{\boldsymbol{u}}(\cdot)$ to denote the plenoptic function obtained from the viewpoint at location $\boldsymbol{u}$ as

$$\mathcal{L}_{\boldsymbol{u}}(\boldsymbol{x}) := L(\boldsymbol{x}, \boldsymbol{u}). \tag{3}$$

An equivalent expression of Eq. (3) is

$$\mathcal{L}_{\boldsymbol{u}}(\boldsymbol{x}) = \mathcal{L}_{\boldsymbol{u}+\boldsymbol{\Delta u}}(\boldsymbol{x} + \boldsymbol{\Delta x}), \tag{4}$$

if we assume the pixels in an image shift by $\boldsymbol{\Delta x}$ when the viewpoint changes by $\boldsymbol{\Delta u}$.

Substituting Eq. (2) into Eq. (4), we obtain

$$\begin{aligned}\mathcal{L}_{\boldsymbol{u}}(\boldsymbol{x}) &= \mathcal{L}_{\boldsymbol{u}+\boldsymbol{\Delta u}}(\boldsymbol{x} + \boldsymbol{\Delta x}) \\ &= \mathcal{L}_{\boldsymbol{u}+\boldsymbol{\Delta u}}\left(\boldsymbol{x} + \left(1 - \frac{1}{\alpha}\right)\boldsymbol{\Delta u}\right).\end{aligned} \tag{5}$$

B. Photo-Consistency

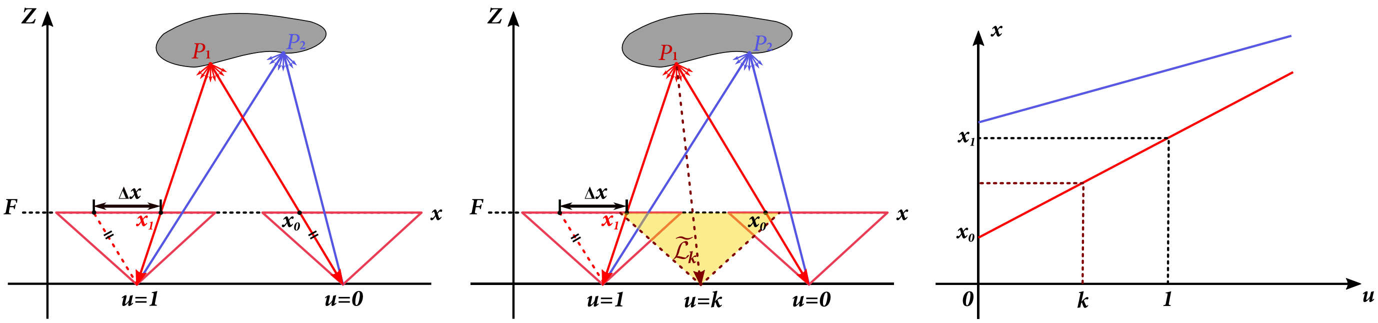
The photo-consistency assumption is commonly adopted in the multi-view vision tasks. It assumes that all light rays coming from the same focus point in the scene should result in the same photometric values. Since the rays from different directions are recorded separately in a light field, this assumption means that the value of the recorded pixels in different SAIs corresponding to the same point of the scene should be identical. Nevertheless, it does not always hold, and usually fails when there are occlusions along the path of light ray transmission.

C. Free-Space Intermediate Radiance Inference

We first establish a model to describe the radiance of a light ray impinging on an arbitrary position in the camera plane. Assume that all the light rays are transmitted in free space, i.e., the space free of occluders, as illustrated in Fig. 2a. For simplicity, in the following we will retain only one angular coordinate u, and derive $L(\boldsymbol{x}, u)$ for a collection of light rays traveling from the position u to the position $\boldsymbol{x}$ in the respective planes. Such collection of light rays form the view image from the viewpoint u. We consider the continuous change of viewpoint u and set its range to be $[0, 1]$. The two boundary viewpoint positions are $u = 0$ and $u = 1$.

Given two collections of light rays $\mathcal{L}_0(\boldsymbol{x})$ and $\mathcal{L}_1(\boldsymbol{x})$ from two different viewpoints, and a factor $k \in (0, 1)$, our goal is to infer the intermediate rays $\widetilde{\mathcal{L}}_k(\boldsymbol{x})$, as is demonstrated in Fig. 2b. According to Eq. (2), one can approximate the relationship between coordinates u and x using a linear mapping. The EPI is a map obtained by gathering the light field samples with a fixed spatial (x or y) and a fixed angular (u or v) coordinate. It can reflect the relationship between a pair of spatial and angular coordinates (x and u, or y and v). Given a certain point on the Lambertian surface with depth Z, when changing the viewpoint u, the spatial position x will also change according to Eq. (2), forming a line on the EPI. The slope of the line is related to the depth of the point.

Fig. 2c shows the EPI pattern of two points at different depths. The red line corresponds to the point P_1 while the blue line corresponds to P_2. The photo-consistency assumption ensures that each line should have a uniform color, i.e. projections of the same point in different views should have the same intensity value. This allows us to approximate



(a) Illustration of light ray transmission in free space.

(b) Illustration of intermediate radiance inference model in free space.

(c) EPI pattern corresponding to (b). Each line has a uniform color due to the photo-consistency assumption.

Fig. 2. Intermediate radiance inference model in space free of occluders in light field.

the radiance $\mathcal{L}_k(\boldsymbol{x})$ by shifting the pixels corresponding to $\mathcal{L}_0(\boldsymbol{x})$ and $\mathcal{L}_1(\boldsymbol{x})$ properly. Also, in terms of Eq. (5), we have

$$\begin{aligned}\widetilde{\mathcal{L}}_k(\boldsymbol{x}) &\cong (1-k)\cdot\mathcal{L}_k(\boldsymbol{x}) + k\cdot\mathcal{L}_k(\boldsymbol{x})\\ &= (1-k)\cdot\mathcal{L}_0\left(\boldsymbol{x}-\left(1-\frac{1}{\alpha}\right)k\right)\\ &\quad +k\cdot\mathcal{L}_1\left(\boldsymbol{x}+\left(1-\frac{1}{\alpha}\right)(1-k)\right), \end{aligned}\tag{6}$$

where $\widetilde{\mathcal{L}}_k(\boldsymbol{x})$ denotes the estimate of $\mathcal{L}_k(\boldsymbol{x})$. Eq. (6) provides a more general expression. Practically, $\mathcal{L}_0(\boldsymbol{x})$ and $\mathcal{L}_1(\boldsymbol{x})$ are usually close but not the same, due to the noise and illumination. Therefore, the coefficient k can also be regarded as a weighting factor to balance the information from $\mathcal{L}_0(\boldsymbol{x})$ and $\mathcal{L}_1(\boldsymbol{x})$ for the estimate. However, since the depth information of the radiance is unknown, the ratio α cannot be computed.

To mitigate such a problem, we develop a way to estimate the disparity ($\boldsymbol{\Delta x}$) directly. An ADM is learned by a dense CNN, which provides the pixel-wise shift information between a pair of view images. Here, we use it to obtain the radiance inference. A more detailed demonstration will be presented in Section III-E.

Mathematically, $A_{k\leftarrow 0}(\boldsymbol{x})$ and $A_{k\leftarrow 1}(\boldsymbol{x})$ denote the ADMs from $\mathcal{L}_0(\boldsymbol{x})$ to $\mathcal{L}_k(\boldsymbol{x})$ and $\mathcal{L}_1(\boldsymbol{x})$ to $\mathcal{L}_k(\boldsymbol{x})$, respectively. Following Eq. (6), we have

$$\begin{aligned}\widetilde{\mathcal{L}}_k(\boldsymbol{x}) = (1-k)P\Big(\mathcal{L}_0(\boldsymbol{x}), A_{k\leftarrow 0}(\boldsymbol{x})\Big)\\ +kP\Big(\mathcal{L}_1(\boldsymbol{x}), A_{k\leftarrow 1}(\boldsymbol{x})\Big),\end{aligned}\tag{7}$$

where $P(\cdot,\cdot)$ is the pixel-wise warping operation that translates each pixel of the input image to its correspondence in the target image in terms of the disparity map between the two images. Given the input image $\mathcal{L}_i(\boldsymbol{x})$ ($i\in[0,1]$) and the ADM $A_{o\leftarrow i}(\boldsymbol{x})$ ($o\in[0,1]$), the output image of the function $P(\cdot,\cdot)$ is

$$P\Big(\mathcal{L}_i(\boldsymbol{x}), A_{o\leftarrow i}(\boldsymbol{x})\Big) = \mathcal{L}_o(\boldsymbol{x}) = \mathcal{L}_i\Big(\boldsymbol{x}+A_{o\leftarrow i}(\boldsymbol{x})\Big). \tag{8}$$

Eq. (7) implies that the closer the viewpoint position k is to 0, the more contribution $\mathcal{L}_0(\boldsymbol{x})$ will make to $\widetilde{\mathcal{L}}_k(\boldsymbol{x})$.

D. Inference With Occlusions

We next discuss the situation where there exist occlusions along the path of the light ray, violating the

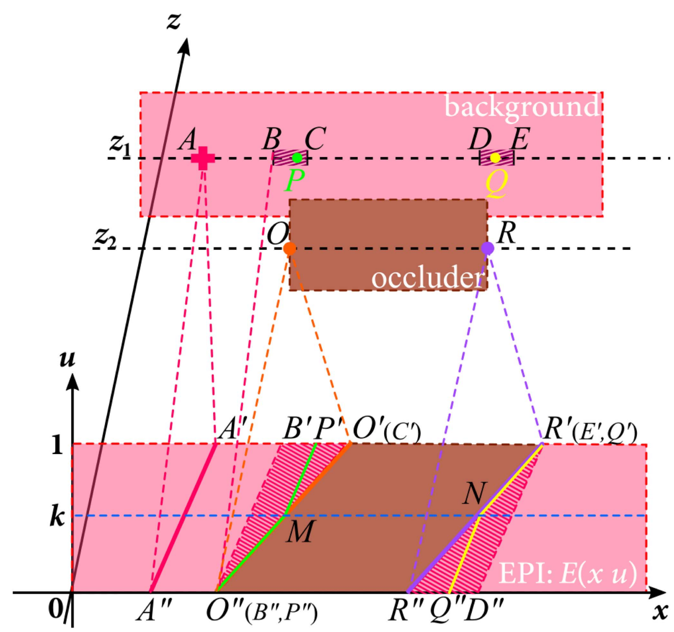


Fig. 3. EPI pattern with occlusions between the observer and the object.

photo-consistency assumption. Given the direction of light rays recorded by a plenoptic camera, an important property of the light field is that the *occluded pixel in the synthesized image always appears in either the leftmost SAI or the rightmost SAI (i.e., the boundary images)*. This is illustrated in Fig. 3, which presents the EPI pattern of the dashed line region of two objects placed at different distances from the camera. The near object (named "occluder") with depth z_2 partially occludes the further object (named "background") with distance z_1. On the background, the region CD is totally occluded in all views. The regions BC and DE are partially occluded and the other places are totally exposed. In the EPI, each slope line corresponds to a point. For example, the line $A'A''$ with a small slope on EPI corresponds to the point A of the background, while the line $O'O''$ is projected from point O on the occluder with a larger slope on the EPI pattern.

With such a configuration, we now illustrate how to modify the free-space inference model to fit scenarios with occlusions. As the occlusion appears near the boundary of the object, we shade the corresponding EPI in Fig. 3 to highlight these regions with occlusions, i.e. $B'O'O''$ and $R'D''R''$. We first focus on a certain pixel located at point P on the background that is also at the boundary of the occluder when $u = k$. Its EPI pattern is marked with a green line $P'MP''$. This point is occluded when $u \in (0, k)$, but is exposed when $u \in (k, 1)$. On the other hand, another point Q is projected to the line $Q'NQ''$, which is occluded when $u \in (k, 1)$ but is exposed when $u \in (0, k)$.

Nevertheless, both P and Q can be inferred from one of the boundary images, i.e. $\mathcal{L}_0(\boldsymbol{x})$ for Q and $\mathcal{L}_1(\boldsymbol{x})$ for P. As a consequence, when there are occlusions along the light ray, the missing pixel information can be inferred from one of the boundary images. Based on this observation, we design another network to estimate the warping confidence maps, known as WCMs, to assist in the inference of the pixel value. The estimated WCMs, $O_{k\leftarrow 0}(\boldsymbol{x})$ and $O_{k\leftarrow 1}(\boldsymbol{x})$, denote the confidence level that a pixel value of $\mathcal{L}_k(\boldsymbol{x})$ can be inferred from $\mathcal{L}_0(\boldsymbol{x})$ and $\mathcal{L}_1(\boldsymbol{x})$, respectively.

As a result, Eq. 7 can be modified as

$$\widetilde{\mathcal{L}}_k(\boldsymbol{x}) = \Phi^{-1} \odot \Big[(1-k)O_{k\leftarrow 0}(\boldsymbol{x}) \odot P\Big(\mathcal{L}_0(\boldsymbol{x}), A_{k\leftarrow 0}(\boldsymbol{x})\Big) + kO_{k\leftarrow 1}(\boldsymbol{x}) \odot P\Big(\mathcal{L}_1(\boldsymbol{x}), A_{k\leftarrow 1}(\boldsymbol{x})\Big)\Big], \quad (9)$$

where $\Phi = kO_{k\leftarrow 0}(\boldsymbol{x}) + (1-k)O_{k\leftarrow 1}(\boldsymbol{x})$ denotes a normalization factor. The symbol $\odot$ denotes the Hadamard product between two matrices. The values of the confidence map should fall within $[0, 1]$. Take, as an example, the point P (with coordinate $\boldsymbol{p}$), which is occluded when $u = 0$ and exposed when $u = 1$. We should have $O_{k\leftarrow 0}(\boldsymbol{p}) = 0$, which means that $\mathcal{L}_0(\boldsymbol{p})$ has no contribution to $\widetilde{\mathcal{L}}_k(\boldsymbol{p})$, and $O_{k\leftarrow 1}(\boldsymbol{p}) = 1$, such that the value is fully contributed by $\mathcal{L}_1(\boldsymbol{p})$. Similarly, for point Q (with coordinate $\boldsymbol{q}$), the value of two confidence maps should satisfy $O_{k\leftarrow 0}(\boldsymbol{q}) = 1$ and $O_{k\leftarrow 1}(\boldsymbol{q}) = 0$. For the points not in the partially occluded region, the value is contributed from both boundary light rays. Consequently, these two maps should satisfy the constraint

$$O_{k\leftarrow 0}(\boldsymbol{x}) + O_{k\leftarrow 1}(\boldsymbol{x}) = 1. \quad (10)$$

E. Aperture Disparity Map Estimation

After we obtain the occlusion-aware inference expression in Eq. (9), the next crucial problem is how to estimate the intermediate ADM $A_{k\leftarrow 1}(\boldsymbol{x})$ and $A_{k\leftarrow 0}(\boldsymbol{x})$ for target view synthesis. As discussed earlier in Section III-C, each value of the ADM represents the spatial shift of the corresponding pixel (radiance). It approximates the relationship between shift distance and viewpoint changes using a linear mapping. To demonstrate this, we define an auxiliary variable $A(\boldsymbol{x}, u)$, which denotes the distance $\boldsymbol{x}$ has shifted from the original viewpoint (0) to viewpoint u. Consequently, ADM can also be expressed using the auxiliary variable, i.e. $A(\boldsymbol{x}, u) = A_{u\leftarrow 0}(\boldsymbol{x})$.

According to Eq. (2), one can easily obtain the proportional relation between $A(\cdot,\cdot)$ and coordinate u, as demonstrated in Fig. 4. Therefore, for any $\boldsymbol{x}$ and k, the disparity value satisfies

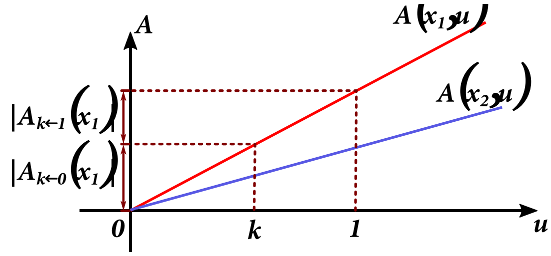


Fig. 4. Illustration of a linear relation between $A(\cdot,\cdot)$ and the angular coordinate u. The values x_1 and x_2 denote the pixels projected from points P_1 and P_2 (Fig. 2), respectively. The EPI patterns corresponding to the two points are highlighted in different colors.

$$A(\boldsymbol{x}, 0) = 0 \quad (11)$$
$$A_{k\leftarrow 0}(\boldsymbol{x}) = A(\boldsymbol{x}, k) - A(\boldsymbol{x}, 0) = -A_{0\leftarrow k}(\boldsymbol{x}) \quad (12)$$
$$A_{k\leftarrow 1}(\boldsymbol{x}) = A(\boldsymbol{x}, k) - A(x, 1) = -A_{1\leftarrow k}(\boldsymbol{x}) \quad (13)$$
$$A_{k\leftarrow 0}(\boldsymbol{x}) = k \cdot A_{1\leftarrow 0}(\boldsymbol{x}) = -k \cdot A_{0\leftarrow 1}(\boldsymbol{x}) \quad (14)$$
$$A_{k\leftarrow 1}(\boldsymbol{x}) = (k-1) \cdot A_{1\leftarrow 0}(\boldsymbol{x}) = (1-k) \cdot A_{0\leftarrow 1}(\boldsymbol{x}) \quad (15)$$
$$A_{1\leftarrow 0}(\boldsymbol{x}) = A(\boldsymbol{x}, 1) - A(\boldsymbol{x}, 0) = A_{1\leftarrow k}(\boldsymbol{x}) + A_{k\leftarrow 0}(\boldsymbol{x}). \quad (16)$$

Eqs. (11) to (15) are derived directly from the epipolar property of light field described by Eq. (2) and Eq. (5). In addition, Eq. (16) can be deduced by combining Eq. (12) and Eq. (13). Therefore, ADMs $A_{k\leftarrow 1}(\boldsymbol{x})$ and $A_{k\leftarrow 0}(\boldsymbol{x})$ can be calculated in terms of $A_{1\leftarrow 0}(\boldsymbol{x})$ and $A_{0\leftarrow 1}(\boldsymbol{x})$ as

$$A_{k\leftarrow 0}(\boldsymbol{x}) = \frac{k+1}{2}A_{1\leftarrow 0}(\boldsymbol{x}) + \frac{1-k}{2}A_{0\leftarrow 1}(\boldsymbol{x}) \quad (17)$$
$$A_{k\leftarrow 1}(\boldsymbol{x}) = \frac{k}{2}A_{1\leftarrow 0}(\boldsymbol{x}) + \left(1 - \frac{k}{2}\right)A_{0\leftarrow 1}(\boldsymbol{x}), \quad (18)$$

if we assume the summation of the weights in Eq. 17 and Eq. 18 to be 1. The intermediate radiance inference model derived in Sections III-C and III-D can be easily extended to the other angular (v) coordinate.

The expressions in Eq. 17 and Eq. 18 allow us to calculate two intermediate ADMs based on the boundary ADMs, $A_{1\leftarrow 0}(\boldsymbol{x})$ and $A_{0\leftarrow 1}(\boldsymbol{x})$. Because the boundary images are given, the two boundary ADMs can be generated using deep learning methods. In our framework, ADMs and WCMs are estimated sequentially using a dense network. As shown in [58], the dense residual network (RDN) has an effective skip connection pattern, which encourages feature reuse and makes the model more compact and less prone to overfitting. In addition, each individual layer can directly receive the supervision from the loss function through the shortcut path, which provides implicit deep supervision. Considering such desirable properties of RDN, we adopt it for ADM and WCM estimation. We use two dense residual blocks, and each includes 4 convolutional layers and 3 ReLU layers. At the end of each block, the learned dense features are concatenated and fed into a convolutional layer with a 1×1 kernel to learn more effective features adaptively. In addition, each block allows direct connections from the previous blocks to extract the hierarchical features for disparity and confidence map estimation.

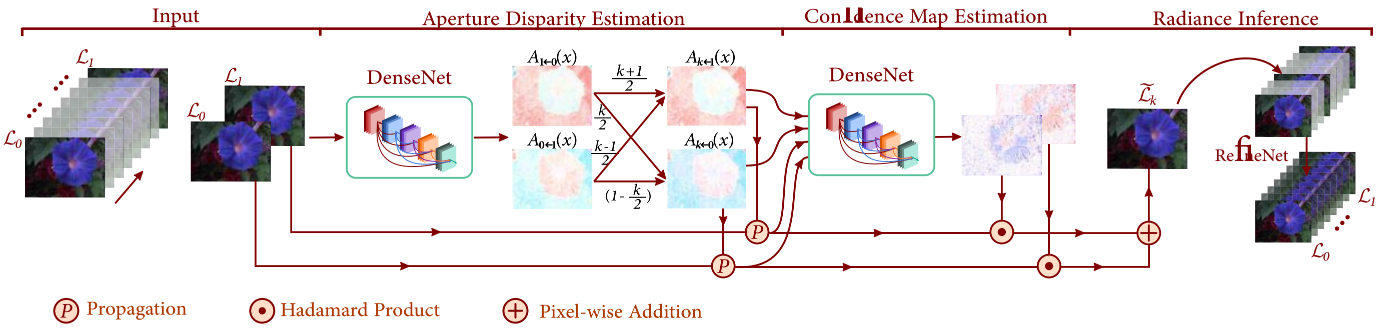


Fig. 5. Overview of our proposed framework. Symbol Ⓟ is the pixel-wise warping operation, $\odot$ is the Hadamard product, and $\oplus$ refers to the pixel-wise addition.

The overview of the proposed framework is shown in Fig. 5, which mainly consists of three stages (excluding the input). In aperture disparity estimation, our model adopts a dense network to estimate the ADMs between two boundary images. They are further used to approximate the intermediate ADMs, $A_{k\leftarrow 1}(\cdot)$ and $A_{k\leftarrow 0}(\cdot)$. Then, in confidence map estimation, both the intermediate disparity maps and images are fed into a subsequent dense network to obtain the confidence maps that indicate the pixel-wise contribution from the two input SAIs. In radiance inference, the target image at k is computed according to Eq. (9). It is further refined by exploiting the parallax information of all the SAIs with the RefineNet. Recently, some studies have demonstrated the effectiveness of 4D convolution filter for spatial details reconstruction. Meng *et al.* [12] adopted multiple 4D convolutional layers to recover the high-frequency spatial details. While the results are satisfactory, the framework requires a large amount of computation. To deal with such a problem, we make use of the alternating convolution filter [24] in our RefineNet. It is a type of 4D convolution filter, which can fully exploit the parallax information of all the views and reduce a significant amount of computation. This fully-convolutional module can then efficiently handle the input light field with changeable angular size. RefineNet contains two 4D filtering steps, and each one is approximated with two alternating filters with kernel size $3 \times 3 \times 1 \times 1$ and $1 \times 1 \times 3 \times 3$, respectively. The learned features are concatenated and then fed into a 4D filter with kernel size $1 \times 1 \times 1 \times 1$ to obtain the refined results.

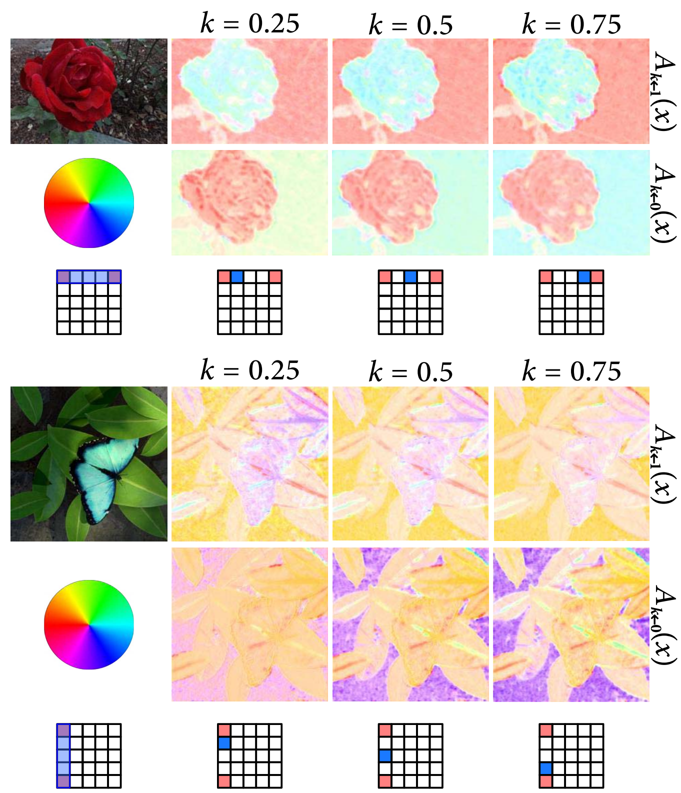


Fig. 6. Illustration of the ADMs from boundary images (denoted by red squares) to the target image (denoted by blue square). Actual ADMs are visualized using the color coding shown in the subfigure in the bottom-left corner of each scene. Colors represent the directions of vectors, and lighter colors mean smaller vectors.

F. Loss Function

All modules and calculations in our framework are differentiable, which enables us to train different parts of our model synchronously. Given the input images $\mathcal{L}_0(\boldsymbol{x})$ and $\mathcal{L}_1(\boldsymbol{x})$, and a set of intermediate images $\{\mathcal{L}_{k_i}(\boldsymbol{x})\}_{i=1}^N$, our loss function consists of four parts. First is a reconstruction loss that directly provides a supervision signal for the synthesized results by calculating the absolute residual error ℓ_r between the intermediate images and the corresponding labels, i.e.,

$$\ell_r = \frac{1}{N}\sum_{i=1}^{N} \left\| \mathcal{L}_{k_i}(\boldsymbol{x}) - \widetilde{\mathcal{L}}_{k_i}(\boldsymbol{x}) \right\|_1. \tag{19}$$

To reconstruct the high-frequency spatial details [50], [59], we also add the content perceptual loss component ℓ_c given by

$$\ell_c = \frac{1}{N}\sum_{i=1}^{N} \left\| \phi\left(\mathcal{L}_{k_i}(\boldsymbol{x})\right) - \phi\left(\widetilde{\mathcal{L}}_{k_i}(\boldsymbol{x})\right) \right\|^2, \tag{20}$$

where $\phi(\cdot)$ maps the input images into high-level feature vectors extracted from the ImageNet pretrained VGG16 model (`conv4_3` layer) [60].

The third part is a warping loss ℓ_ω, which models the quality of the four estimated AFMs as shown in Fig. 5. It includes four terms, and each measures the absolute difference between

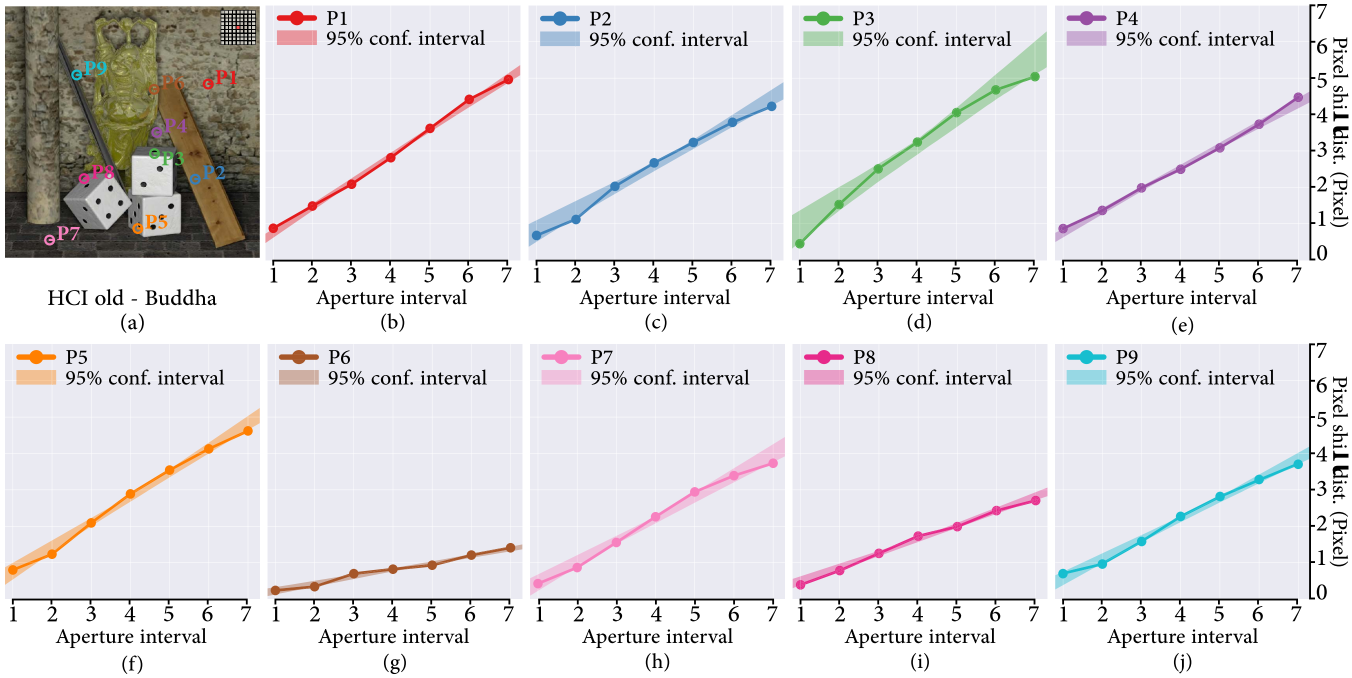


Fig. 7. Relationships between the pixel shift distance and the aperture interval.

TABLE I

MEASUREMENT OF THE LINEAR DEPENDENCE BETWEEN THE FEATURE POINT SHIFT DISTANCE AND THE APERTURE INTERVAL

Feature Points	Statistical measurements		
	R^2	Adjusted R^2	PCC
P1	0.996	0.996	0.998
P2	0.991	0.989	0.996
P3	0.989	0.986	0.994
P4	0.998	0.997	0.999
P5	0.994	0.993	0.997
P6	0.986	0.984	0.993
P7	0.984	0.981	0.992
P8	0.994	0.993	0.997
P9	0.991	0.990	0.996
Average	0.991	0.990	0.996

the corresponding propagated views and the ground truth, i.e.,

$$\begin{aligned}\ell_\omega = &\left\|\mathcal{L}_0(\boldsymbol{x}) - P\left(\mathcal{L}_1(\boldsymbol{x}), A_{0\leftarrow 1}(\boldsymbol{x})\right)\right\|_1 \\ &+ \left\|\mathcal{L}_1(\boldsymbol{x}) - P\left(\mathcal{L}_0(\boldsymbol{x}), A_{1\leftarrow 0}(\boldsymbol{x})\right)\right\|_1 \\ &+ \frac{1}{N}\sum_{i=1}^{N}\left\|\mathcal{L}_{k_i}(\boldsymbol{x}) - P\left(\mathcal{L}_1(\boldsymbol{x}), A_{k_i\leftarrow 1}(\boldsymbol{x})\right)\right\|_1 \\ &+ \frac{1}{N}\sum_{i=1}^{N}\left\|\mathcal{L}_{k_i}(\boldsymbol{x}) - P\left(\mathcal{L}_0(\boldsymbol{x}), A_{k_i\leftarrow 0}(\boldsymbol{x})\right)\right\|_1. \end{aligned} \tag{21}$$

Finally, we smooth the estimated aperture disparity with [61]

$$\begin{aligned}\ell_s = &\|\nabla_x A_{0\leftarrow 1}(\boldsymbol{x})\|_1 + \left\|\nabla_y A_{0\leftarrow 1}(\boldsymbol{x})\right\|_1 \\ &+ \|\nabla_x A_{1\leftarrow 0}(\boldsymbol{x})\|_1 + \left\|\nabla_y A_{1\leftarrow 0}(\boldsymbol{x})\right\|_1, \end{aligned} \tag{22}$$

where the notation ∇ denotes the differential operation.

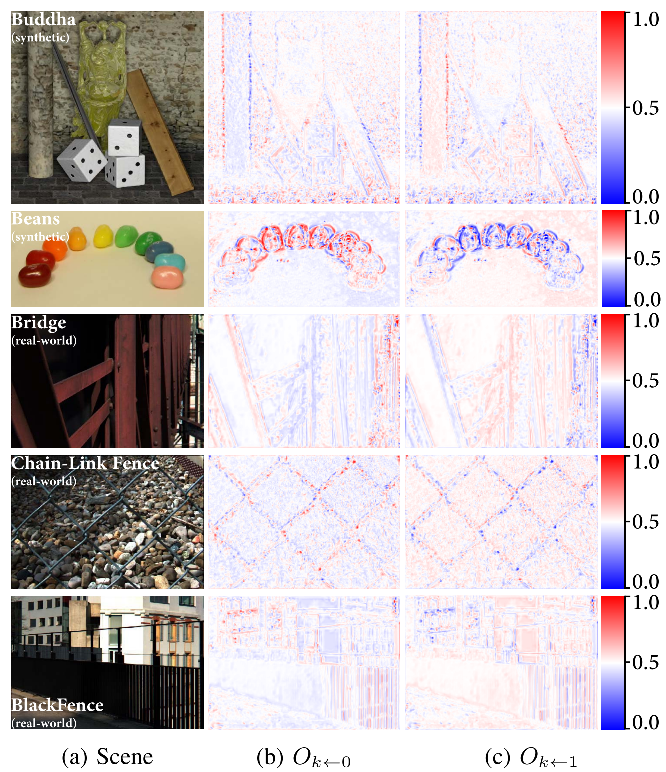


Fig. 8. Illustration of the WCMs of different scenes. All WCMs have $k = 0.5$, i.e., they correspond to the image midway between two input images, taken from various datasets.

Our overall objective is a summation of these loss functions, i.e.,

$$\ell = \lambda_1 \ell_r + \lambda_2 \ell_c + \lambda_3 \ell_\omega + \lambda_4 \ell_s, \tag{23}$$

where λ_1, λ_2, λ_3 and λ_4 are the respective weights. Empirically, they are set to 200, 1000, 100 and 1, respectively.

Fig. 9. We compare our approach against multiple state-of-the-art methods on several difficult Lytro light field scenes from Kalantari *et al.* [16]. The scene "Flower 1" contains a significant number of occluded regions and the scene "Cars" contains many thin structures, such as the fibre of the paper (in the magenta box) and the reflection of tree branches (in the red box). Our approach produces more realistic results compared with the selected algorithms. We also show the EPI at the dashed line in each zoomed region.

IV. Experiments

A. Datasets and Implementation Details

The effectiveness of data-driven methods often depends significantly on the quality of the training data. Compared with many view synthesis approaches that directly learn to produce the target image, our model attempts to approximate the image relations or correspondences, which increase the model robustness to different types of light field images. To demonstrate this, we train our networks using 100 real-world light field images captured with a Lytro Illum camera provided by the Stanford Lytro Archive [62]. Due to hardware limitation of plenoptic cameras, many corner angular samples are outside the field of view. Therefore, for each scene, we select the center 9×9 views in the experiments. To validate the effectiveness of our proposed framework, we conduct experiments on both synthetic and real scenes. The former are selected from HCI datasets [63], [64] generated using the Blender software, while the latter are from multiple public datasets captured with the Lytro Illum cameras, including Stanford Lytro Archive [62] (not included in the training set), Kalantari *et al.*'s [16] and Flower [54].

TABLE II

Quantitative Comparisons (PSNR/SSIM) of the Proposed Approach With Continuous View Synthesis Algorithms for Angular Super-Resolution $2\times2 \rightarrow 8\times8$. The Input Light Fields for Each Algorithm Are Sampled at Four Corners

Scenes	Methods			
	Kalantari *et al.* [16]	Soft3D [43]*	FPFR [31]	**Ours**
Flowers1	33.32 / 0.960	32.84 / 0.960	34.51 / **0.972**	**34.64 / 0.972**
Flowers2	31.94 / 0.960	32.74 / 0.961	**34.20 / 0.972**	34.03 / 0.965
Cars	31.65 / 0.966	30.93 / 0.962	32.28 / 0.967	**33.02 / 0.973**
Seahorse	31.87 / 0.970	31.22 / 0.961	**34.36** / 0.970	33.61 / **0.972**
StoneLion	40.57 / 0.979	38.53 / 0.973	40.02 / 0.975	**40.84 / 0.981**
Leaves1	35.84 / 0.973	31.74 / 0.947	34.76 / 0.966	**36.95 / 0.978**
Leaves2	34.17 / 0.963	32.47 / 0.965	32.86 / 0.957	**34.28 / 0.966**
Average	34.19 / 0.967	32.92 / 0.961	34.71 / 0.969	**35.33 / 0.972**

*Note that Soft3D [43] and FPFR [31] did not release their codes. For FPFR, we used the code provided by the authors. For Soft3D, we used the code reimplemented by the authors of reference [66].

Our project is implemented using PyTorch, and the optimization model is trained on a Ubuntu 16.04.4 computer with

TABLE III
QUANTITATIVE COMPARISONS (PSNR/SSIM) OF THE PROPOSED APPROACH WITH LEARNING-BASED METHODS

Scenes	Data Type	Methods				
		Kalantari *et al.* [16]	Wu *et al.* [22]	Yeung *et al.* [24]	HDDRNet [12]	**Ours**
Bedroom	Synthetic	34.87 / 0.914	32.52 / 0.890	34.26 / 0.895	34.29 / 0.892	**34.48 / 0.906**
Cotton	Synthetic	41.98 / 0.964	38.49 / 0.944	41.47 / 0.956	42.32 / 0.961	**43.63 / 0.973**
Dino	Synthetic	36.88 / 0.951	34.21 / 0.904	40.38 / 0.951	40.34 / 0.952	**40.78 / 0.957**
Origami	Synthetic	30.91 / 0.903	28.26 / 0.898	31.54 / 0.912	32.18 / 0.912	**33.59 / 0.919**
Average	—	36.16 / 0.933	33.37 / 0.909	36.91 / 0.929	37.28 / 0.929	**38.12 / 0.939**
Occlusions	Real-world	32.18 / 0.897	30.48 / 0.867	**33.19** / 0.908	32.78 / 0.909	33.10 / **0.912**
Reflective	Real-world	35.28 / 0.923	33.21 / 0.893	36.82 / 0.933	36.77 / 0.931	**37.01 / 0.950**
Average	—	33.73 / 0.910	31.85 / 0.880	35.01 / 0.920	34.78 / 0.920	**35.06 / 0.931**

an Intel Xeon(R)@2.20HGz CPU and a Titan X GPU. The input images are randomly cropped to 224×224. Moreover, we use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is set to 10^{-4} and reduced by a factor of 0.1 for every 5 epochs. The training takes approximately 7 days.

B. Aperture Disparity Map

One crucial component of our synthesis algorithm is the ADM. Each is estimated in an unsupervised manner, where the signals from the labels are directly used to reconstruct the images. Fig. 6 gives an example of the estimated disparity maps between the boundary images ($\mathcal{L}_0(\boldsymbol{x})$ and $\mathcal{L}_1(\boldsymbol{x})$) and different intermediate images. We use the center 5×5 views to visualize the estimated intermediate ADMs between the images with different intervals. The input images are indicated by red squares while the ADMs are denoted by blue squares. The figure shows the ADMs between different horizontal and vertical views from both real-world and synthetic scenes. Both estimated ADMs are learned in an unsupervised manner. According to the color-coding panel, the pixels of foreground and background are shifted in opposite directions, and the colors tend to be darker when the interval between the two views is larger.

Another issue we need to tackle is a quantitative assessment of the quality of the computed ADMs. The linear relations (listed in Eqs. (11) to (16)) offer an approach to address this. One can evaluate the linearity of the estimated ADMs in terms of the viewpoint distance between the input SAIs. To do this, we feed a sequence of image pairs with different aperture intervals into the disparity estimation network and analyze the output ADMs $A_{1\leftarrow 0}(x)$ and $A_{0\leftarrow 1}(x)$.

The aperture interval controls how different the two input images are. For example, if the two SAIs are with angular coordinates $(0, 0)$ and $(0, 2)$ respectively, their aperture interval is 2. In the experiment, we track several feature points of an input image and record the shift distance of each according to the output ADMs. These feature points are selected using corner detection [65]. Fig. 7 shows the results of 9 randomly picked feature points throughout the image on both the foreground objects and the background wall. The distribution of these points in a synthetic scene (*Buddha*) are shown in Fig. 7a, while the other plots (Figs. 7b to 7j) show the relationship between the point shift distance and the aperture interval.

To further demonstrate the linearity between these two variables, we also compute their linear regression, and plot the regression line and 95% confidence intervals in the Figs. 7b to 7j. Table I presents some statistics to measure the linear correlation between the shift distance and aperture interval, including the coefficient of determination (R^2), adjusted R^2, and Pearson correlation coefficient (PCC). The average values of all three measurements are over 0.99, which suggests that the two variables are highly linearly related.

C. Warping Confidence Map

The ADM aims at estimating the displacement of each pixel in an image, but it can fail in regions with occlusions. To handle this, we introduce the WCM in order to consolidate the information from multiple views to predict a pixel value. Fig. 8 presents the WCMs of images midway between two input images, using both synthetic and real-world scenes. For each WCM, pixels in red denote those at these positions, the source image contributes more; the darker color in red, the higher contribution it makes. The case is opposite for blue. For both real-world and synthetic scenes, the object boundary regions tend to be darker in red or blue. As discussed in Section III-D, the occlusions always appear at the boundary regions, which magnify the distinction of the contributions (on the occluded pixels) from two input images. Take the "buddha" scene (the first row in Fig. 8) as an example. In WCMs $O_{k\leftarrow 0}$ and $O_{k\leftarrow 1}$, there are different colors near the left and right boundaries of the pillar. This demonstrates that the left input image ($\mathcal{L}_0$) contributes more on the left boundary region while the right input image ($\mathcal{L}_1$) contributes more on the right boundary region, as interpreted in Fig. 3 (Line $P'MO''$ and Line $R'NQ''$).

V. RESULTS AND DISCUSSIONS

To validate the effectiveness of our proposed framework, we conduct experiments on both synthetic and real-world light fields. We use common classical quantitative metrics, namely peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) to assess the performance of the algorithms.

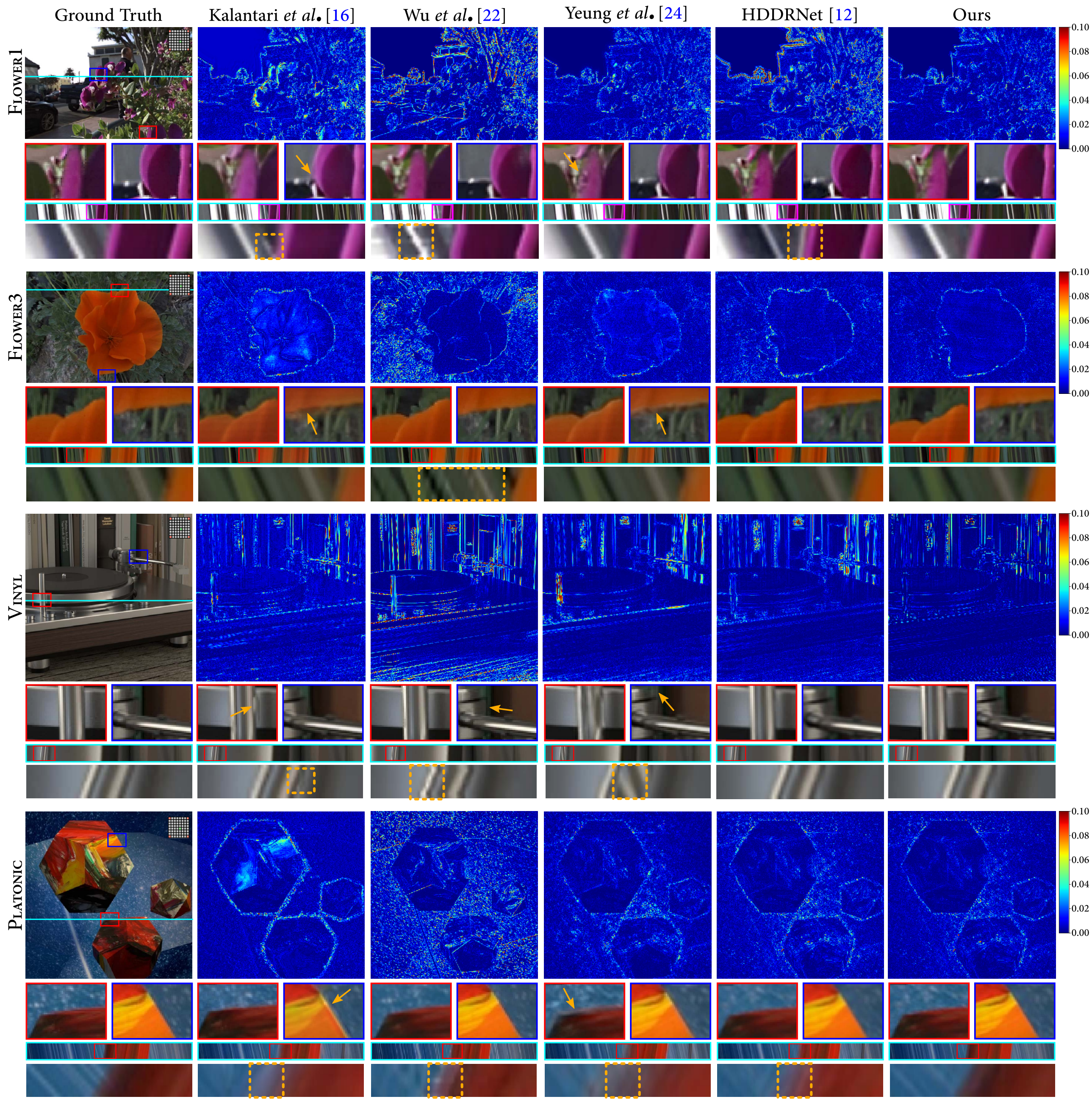


Fig. 10. Visual comparisons of different algorithms on the (5, 5) synthesized view for the task $2 \times 2 \rightarrow 8 \times 8$ on both real-world and synthetic scenes. We select two regions (red and blue boxes) to compare the spatial results of different algorithms. For each reconstructed light field, the EPIs at the position highlighted by the cyan line are also visualized. We zoom in one selected region (red box) of each EPI for better comparison, and in the zoomed EPI we highlight the region that is obviously different from the ground truth.

A. Comparison With Continuous Synthesis Methods

We first evaluate the performance of our method against the recent techniques for continuous view synthesis. Algorithms of this kind attempt to learn a continuous representation of the plenoptic function from a sparsely-sampled input light field. The comparisons are conducted against three recent learning-based methods, namely Kalantari *et al.* [16], Soft3D [43] and Shi *et al.* [31]. Table II presents the quantitative results of different algorithms on the Lytro images for angular $2 \times 2 \rightarrow 8 \times 8$ resolution enhancement. As shown, our method achieves the best quantitative results compared with all these methods.

Fig. 9 compares the visual performance of different algorithms. As shown, the approach of Kalantari *et al.* tends to generate artifacts near the object boundary, such as the leaves and the petal boundary in "Flower 1". Soft3D [43] gives relatively smooth results near the petal boundary. As for "Cars", all of the other three methods lose the information of the thin structures. In this picture, three regions (the colored boxes) with thin structures are selected and zoomed in to highlight the differences. For each zoomed region, we also show the EPI near the dashed line. In comparison, our model produces more realistic spatial results for both thin objects and boundary regions. The disparity is hard to estimated near

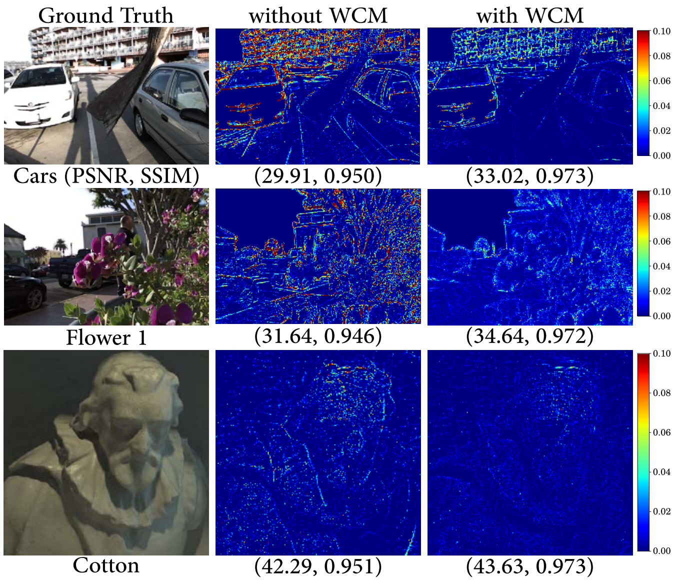


Fig. 11. Performance comparison of the proposed model with and without WCM module.

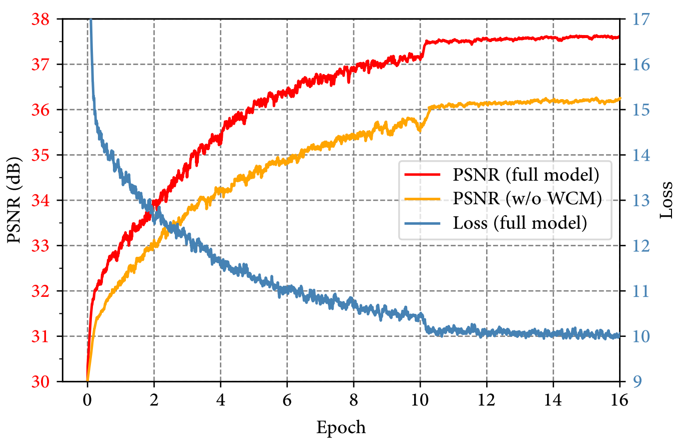


Fig. 12. Learning curves of different variants of the proposed model.

the thin structures, leading to the missing or aliasing of these structures in the reconstructed EPIs. From the EPI results in the scene "Cars", one can also see that the reconstructed light field using our method also retain better disparity information. In both examples, our results are closer to the ground truth. In addition, our approach takes about 0.23s per synthesized view on average to reconstruct a 8×8 light field from its 2×2 views at a spatial resolution of 376×541, which is nearly 40 times faster than Kalantari *et al.* (about 9s per view) and 3.7 times faster than FPFR [31] (about 0.85s per view).

B. Comparison With Other Learning-Based Methods

Next, we compare the performance of our method against several recent learning-based methods, including Wu *et al.* [22], Yeung *et al.* [24] and HDDRNet [12]. Since the method of Kalantari *et al.* [16] has been compared with these methods in the respective paper, we also include it as a reference. Table III and Fig. 10 present the quantitative and visual results, respectively. Wu *et al.* [22] design a "blur-restoration-deblur" pipeline to overcome angular aliasing,

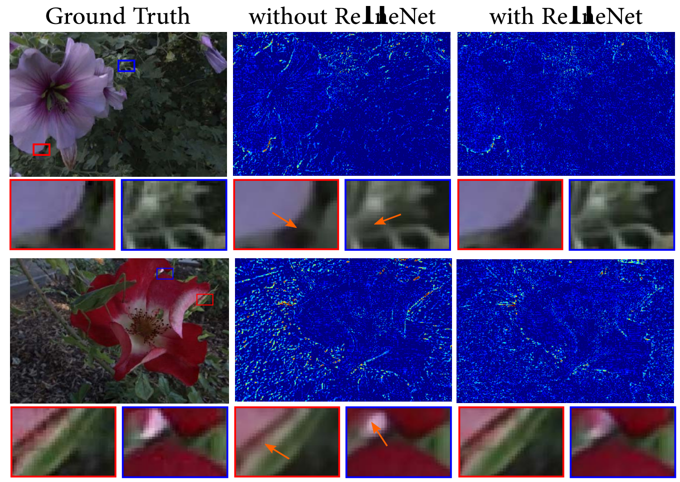


Fig. 13. Illustration of the visual improvements by adding the RefineNet module. The second and third columns present the reconstruction error maps without and with RefineNet.

TABLE IV

VERIFYING THE EFFECTIVENESS OF REFINENET. WE COMPARE THE RECONSTRUCTION QUALITY OF THE LIGHT FIELDS GENERATED WITH AND WITHOUT THIS MODULE UNDER THE TASK $2 \times 2 \rightarrow 8 \times 8$ OVER 20 FLOWER SCENES [54] AND 10 HCI SCENES [63]

Metrics	without RefineNet		with RefineNet	
	Flower (20)	HCI (10)	Flower (20)	HCI (10)
PSNR	35.33	33.56	37.05	34.92
SSIM	0.945	0.912	0.956	0.930

TABLE V

ABLATION STUDY ON DIFFERENT LOSS TERMS FOR TRAINING THE PROPOSED FRAMEWORK ON THE FLOWER DATASET [54]

Settings	PSNR	SSIM
without perceptual loss	35.32	0.943
without warping loss	35.29	0.920
without smoothness loss	35.33	0.944
Full loss	**35.33**	**0.945**

but the algorithm requires at least three views to generate acceptable results. For $2 \times 2 \rightarrow 8 \times 8$ task, because only two views are available for inputs, the insufficient angular information leads to aliasing effects in their reconstructed EPIs, especially in the EPIs of the two Flower scenes in Fig. 10. Yeung *et al.* [24] and HDDRNet [12] achieve state-of-the-art performance. They both adopt the 4D convolution to fully exploit the spatio-angular information, which result in their superior results.

However, for the synthetic scenes, their results tend to have more errors near the object boundaries. Another drawback for these end-to-end methods is that they can only generate the light fields with fixed angular dimensions. This is due to the reshape operation (between the channel and angular dimensions) when upsampling the angular resolution. As a result, when the number of input (or output) views varies, both have to train a new model to fit for such changes. In comparison, we increase the number of views by utilizing the continuous representation among the views and adopt the

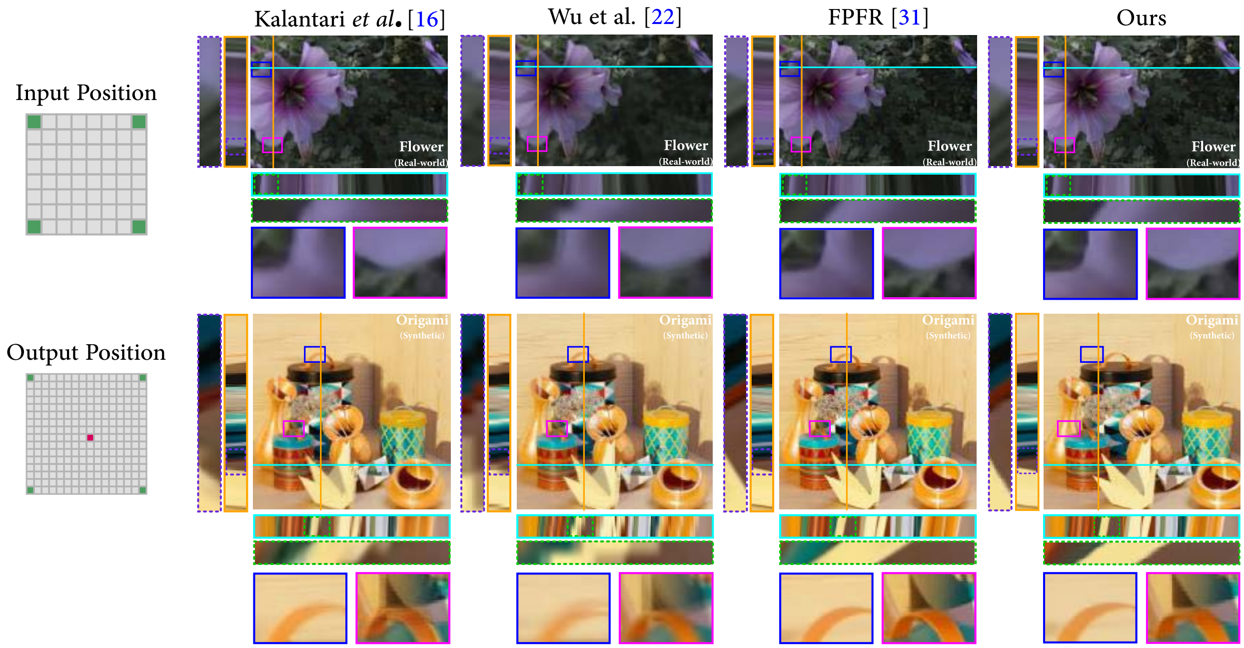


Fig. 14. Visual comparisons on light field dense reconstruction on both real-world (top) and synthetic scenes (bottom). We present the results of 16×16 reconstruction from 2×2 corner views sampled from the 8×8 input image grid. The $(9, 9)^{\text{th}}$ SAI (the red square) of the light field recovered from different methods are presented. Three selected regions have been zoomed in for clarity. Horizontal and vertical EPIs corresponding to the highlighted lines with different colors are shown, and a selected region in each EPI is zoomed in for better viewing.

fully-convolutional framework as the refinement module. Such design enables our approach to handle the input light field with various angular dimension, and the resulting approach achieves the best quantitative results and can produce images close to the ground truth.

C. Ablation Study

The ablation studies are conducted on several crucial components to evaluate our framework. Regarding WCM, we compare the performance of the variants with and without this module, and present the results in Fig. 11. Comparing the error maps in the second and third columns, we can observe an obvious drop on the accuracy of reconstruction near the object boundary regions if WCM is removed. For the scenes contain complex occlusions, such as "Cars" and "Flower 1", the quantitative measurement PSNR reduces even more, by about 3dB.

In addition to the results, we also explore the convergence behavior of different variants, as shown in Fig. 12. By properly handling the occlusions, our model can converge to a better point.

Second, regarding RefineNet, we compare the light fields generated by our approach with and without this module. Table IV shows the quantitative results on both real-world and synthetic scenes. We use 20 real-world scenes randomly selected from the "Flower" dataset [54]. Each light field contains a flower in the foreground, which has a relatively larger disparity than the background. The synthetic scenes are randomly selected from HCI [63] and we use 10 light fields. As shown, the RefineNet module improves the reconstruction results by over 1.3dB on PSNR, and it also shows improvement in terms of SSIM. Fig. 13 illustrates the visual improvements by adding this module. The reconstructed errors are further reduced and some spatial details can be recovered better.

Third, our ablation studies also evaluate the functions of different loss terms. The reconstruction loss term ℓ_r is essential for the network training, and therefore we conduct the ablation studies to evaluate the effectiveness of the other three terms. Table V shows the quantitative results of the proposed network (without the RefineNet module) trained with different loss terms on the "Flower" dataset. As shown, the full model achieves the best performance.

D. Application on Image-Based Rendering

IBR aims at producing the images at new camera positions from a set of captured samples. For light field, one major benefit of rendering-based methods is that they do not necessarily require explicit geometric models and can generate the new views by straightforward interpolation [57]. However, in order to produce the plausible views, such techniques often require the light field to be densely sampled [7]. In comparison, our method can reconstruct the densely-sampled light field with sufficient angular resolution to enable the rendering applications.

To evaluate the effectiveness of our method on image-based rendering applications, we compare the performance of dense reconstruction on both synthetic and real-world scenes. Specifically, we compare the performance of different algorithms when reconstructing 16×16 densely-sampled light fields from 2×2 corner views sampled from the 8×8 input image grid. Fig. 14 visualizes the $(9, 9)^{\text{th}}$ SAI of the reconstructed light field images. As shown, compared with Kalantari *et al.* [16] and Wu *et al.* [22], our method produces more realistic results with sharp textures, and constructs the EPIs with clear slopes. We also compare with FPFR [31], and both methods have competitive performance, according to the visual results shown in the last two columns of Fig. 14.

VI. Conclusion

In this paper, we propose a depth-free algorithm for the reconstruction of arbitrary intermediate views of the light fields. To efficiently describe the parallax between any two given views, we define the ADM based on the epipolar property of the light field. By incorporating also the WCM, our method can efficiently address the occlusions near the object boundaries. Both ADM and WCM are approximated using a dense network. In addition, we further adopt a 4D CNN with alternating filters in the refinement stage to improve the quality of synthesized images. Experimental results have demonstrated that the proposed model achieves state-of-the-art performance for both synthetic and real-world light fields.

Acknowledgment

The authors would like to thank Dr. Xiaoran Jiang, the author of [66], for helping them with the experiments of Soft3D. They would also like to thank the strong support from Dr. Jinglei Shi, the author of [31], and his group. They would also like to thank the Digital Health Laboratory, Department of Orthopaedics and Traumatology, Faculty of Medicine, The University of Hong Kong, for the supporting of model training and testing. Part of the work was done during an internship at Huawei.

References

[1] E. Y. Lam, "Computational photography with plenoptic camera and light field capture: Tutorial," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 32, no. 11, pp. 2021–2032, Nov. 2015.

[2] R. Ng, "Digital light field photography," Ph.D. dissertation, Dept. Comput. Sci., Stanford Univ. California, Stanford, CA, USA, 2006.

[3] N. Chen, C. Zuo, E. Lam, and B. Lee, "3D imaging based on depth measurement technologies," *Sensors*, vol. 18, no. 11, p. 3711, Oct. 2018.

[4] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Comput. Sci. Tech. Rep.*, vol. 2, no. 11, pp. 1–11, 2005.

[5] X. Sun, Z. Xu, N. Meng, E. Y. Lam, and H. K.-H. So, "Data-driven light field depth estimation using deep convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 367–374.

[6] H. Schilling, M. Diebold, C. Rother, and B. Jahne, "Trust your model: Light field depth estimation with inline occlusion handling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4530–4538.

[7] J.-X. Chai, S.-C. Chan, H.-Y. Shum, and X. Tong, "Plenoptic sampling," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn. - SIGGRAPH*, 2000, pp. 307–318.

[8] N. Meng, X. Sun, H. K.-H. So, and E. Y. Lam, "Computational light field generation using deep nonparametric Bayesian learning," *IEEE Access*, vol. 7, pp. 24990–25000, 2019.

[9] N. Viganò, P. M. Gil, C. Herzog, O. de la Rochefoucauld, R. van Liere, and K. J. Batenburg, "Advanced light-field refocusing through tomographic modeling of the photographed scene," *Opt. Exp.*, vol. 27, no. 6, pp. 7834–7856, 2019.

[10] N. Meng, T. Zeng, and E. Y. Lam, "Spatial and angular reconstruction of light field based on deep generative networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4659–4663.

[11] Y. Wang, L. Wang, J. Yang, W. An, J. Yu, and Y. Guo, "Spatial-angular interaction for light field image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2020, pp. 290–308.

[12] N. Meng, H. K.-H. So, X. Sun, and E. Y. Lam, "High-dimensional dense residual convolutional neural network for light field reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 873–886, Mar. 2021.

[13] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis, "Depth synthesis and local warps for plausible image-based navigation," *ACM Trans. Graph.*, vol. 32, no. 3, p. 30, Jun. 2013.

[14] R. S. Overbeck, D. Erickson, D. Evangelakos, M. Pharr, and P. Debevec, "A system for acquiring, processing, and rendering panoramic light field stills for virtual reality," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–15, Jan. 2019.

[15] G. Wu *et al.*, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.

[16] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, p. 193, 2016.

[17] H. Shum and S. B. Kang, "Review of image-based rendering techniques," *Proc. SPIE*, vol. 4067, pp. 2–13, May 2000.

[18] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 286–301.

[19] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "Deep stereo: Learning to predict new views from the world's imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5515–5524.

[20] N. Meng, E. Y. Lam, K. K. Tsia, and H. K.-H. So, "Large-scale multi-class image-based cell classification with deep learning," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 5, pp. 2091–2098, Sep. 2019.

[21] X. Sun, N. Meng, Z. Xu, E. Y. Lam, and H. K.-H. So, "Sparse hierarchical nonparametric Bayesian learning for light field representation and denoising," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 3272–3279.

[22] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai, "Light field reconstruction using convolutional network on EPI and extended applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1681–1694, Jul. 2019.

[23] N. Meng, X. Wu, J. Liu, and E. Lam, "High-order residual network for light field super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, Apr. 2020, pp. 11757–11764.

[24] H. W. F. Yeung, J. Hou, J. Chen, Y. Y. Chung, and X. Chen, "Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 137–152.

[25] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," *ACM Trans. Graph.*, vol. 37, no. 4, p. 65, Aug. 2018.

[26] M. DuVall, J. Flynn, M. Broxton, and P. Debevec, "Compositing light field video using multiplane images," in *Proc. ACM SIGGRAPH Posters*, Jul. 2019, pp. 1–2.

[27] J. Flynn *et al.*, "DeepView: View synthesis with learned gradient descent," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2367–2376.

[28] B. Mildenhall *et al.*, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Trans. Graph.*, vol. 38, no. 4, p. 29:1–29:14, 2019.

[29] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-aware video frame interpolation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3703–3712.

[30] Y.-L. Liu, Y.-T. Liao, Y.-Y. Lin, and Y.-Y. Chuang, "Deep video frame interpolation using cyclic frame generation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 8794–8802.

[31] J. Shi, X. Jiang, and C. Guillemot, "Learning fused pixel and feature-based view reconstructions for light fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2555–2564.

[32] M. Levoy, "Light fields and computational imaging," *Computer*, vol. 39, no. 8, pp. 46–55, Aug. 2006.

[33] A. Kubota, K. Aizawa, and T. Chen, "Reconstructing dense light field from array of multifocus images for novel view synthesis," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 269–279, Jan. 2007.

[34] Z. Lin and H.-Y. Shum, "A geometric analysis of light field rendering," *Int. J. Comput. Vis.*, vol. 58, no. 2, pp. 121–138, Jul. 2004.

[35] A. M. K. Siu and R. W. H. Lau, "Image registration for image-based rendering," *IEEE Trans. Image Process.*, vol. 14, no. 2, pp. 241–252, Feb. 2005.

[36] A. Fitzgibbon, Y. Wexler, and A. Zisserman, "Image-based rendering using Image-based priors," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 141–151, Feb. 2005.

[37] S. M. Seitz and C. R. Dyer, "View morphing," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn. - SIGGRAPH*, 1996, pp. 21–30.

[38] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light field reconstruction using sparsity in the continuous Fourier domain," *ACM Trans. Graph.*, vol. 34, no. 1, p. 12, 2014.

[39] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Light field reconstruction using shearlet transform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 133–147, Jan. 2018.

[40] K. Mitra and A. Veeraraghavan, "Light field denoising, light field superresolution and stereo camera based refocussing using a GMM light field patch prior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 22–28.

[41] A. Levin and F. Durand, "Linear view synthesis using a dimensionality gap light field prior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1831–1838.

[42] P. Hedman, S. Alsisan, R. Szeliski, and J. Kopf, "Casual 3D photography," *ACM Trans. Graph.*, vol. 36, no. 6, p. 234:1–234:15, Nov. 2017.

[43] E. Penner and L. Zhang, "Soft 3D reconstruction for view synthesis," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 235-1–235-11, 2017.

[44] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 606–619, Mar. 2014.

[45] S. E. Chen and L. Williams, "View interpolation for image synthesis," in *Proc. 20th Annu. Conf. Comput. Graph. Interact. Techn. - SIGGRAPH*, 1993, pp. 279–288.

[46] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 261–270.

[47] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2012, pp. 611–625.

[48] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[49] T.-C. Wang, J.-Y. Zhu, N. K. Kalantari, A. Efros, and R. Ramamoorthi, "Light field video capture using a learning-based hybrid imaging system," *ACM Trans. Graph.*, vol. 36, no. 4, p. 133, 2017.

[50] N. Meng, Z. Ge, T. Zeng, and E. Y. Lam, "LightGAN: A deep generative model for light field reconstruction," *IEEE Access*, vol. 8, pp. 116052–116063, 2020.

[51] M. S. K. Gul and B. K. Gunturk, "Spatial and angular resolution enhancement of light fields using convolutional neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2146–2159, May 2018.

[52] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Light-field image super-resolution using convolutional neural network," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 848–852, Jun. 2017.

[53] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, and T. Tan, "LFNet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4274–4286, Sep. 2018.

[54] P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng, "Learning to synthesize a 4D RGBD light field from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2243–2251.

[55] Y. Chen, M. Alain, and A. Smolic, "Self-supervised light field view synthesis using cycle consistency," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Sep. 2020, pp. 1–6.

[56] Y. Gao, R. Bregovic, and A. Gotchev, "Self-supervised light field reconstruction using shearlet transform and cycle consistency," *IEEE Signal Process. Lett.*, vol. 27, pp. 1425–1429, 2020.

[57] M. Levoy and P. Hanrahan, "Light field rendering," in *ACM Conf. Comput. Graph. Interact. Techn.*, 1996, pp. 31–42.

[58] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[59] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.

[60] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[61] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4463–4471.

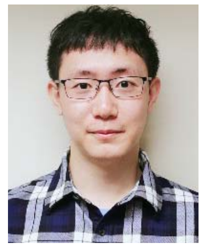
[62] *Stanford Lytro Light Field Archive*. Accessed: Oct. 2018. [Online]. Available: http://lightfields.stanford.edu/

[63] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. Asian Conf. Comput. Vis.* New York, NY, USA: Springer, Mar. 2016, pp. 19–34.

[64] S. Wanner, S. Meister, and B. Goldlüecke, "Datasets and benchmarks for densely sampled 4D light fields," *Vis., Model. Vis.*, vol. 13, pp. 225–226, Sep. 2013.

[65] J. Shi and Tomasi, "Good features to track," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR*, Aug. 1994, pp. 593–600.

[66] X. Jiang, J. Shi, and C. Guillemot, "A learning based depth estimation framework for 4D densely and sparsely sampled light fields," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2257–2261.



Nan Meng received the bachelor's degree from the University of Electronic Science and Technology of China in 2015 and the Ph.D. degree in electrical engineering from The University of Hong Kong in 2020. He is currently a Researcher and the Director with the Department of Orthopaedics and Traumatology, The University of Hong Kong. His research interests include machine learning, light field reconstruction, light field rendering, medical imaging, optical-based diagnosis, and scoliosis analysis.

Kai Li received the bachelor's degree from the University of Electronic Science and Technology of China in 2015. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. His research interests include reinforcement learning, game theory, multi-agent systems, online learning, and computer vision.

Jianzhuang Liu (Senior Member, IEEE) received the Ph.D. degree in computer vision from The Chinese University of Hong Kong, Hong Kong, in 1997. From 1998 to 2000, he was a Research Fellow with Nanyang Technological University, Singapore. From 2000 to 2012, he was a Postdoctoral Fellow, an Assistant Professor, and an Adjunct Associate Professor with The Chinese University of Hong Kong. In 2011, he joined the Shenzhen Institute of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China, as a Professor. He is currently a Principal Researcher with Huawei Technologies Company Ltd., Shenzhen. He has authored more than 150 articles. His research interests include computer vision, image processing, deep learning, and graphics.

Edmund Y. Lam (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Stanford University. From 2010 to 2011, he was a Visiting Associate Professor with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology. He is currently a Professor of Electrical and Electronic Engineering with The University of Hong Kong, and also serves as the Computer Engineering Program Director. His research interest includes computational imaging. He is also a Fellow of OSA, SPIE, IS&T, and HKIE. He was a recipient of the IBM Faculty Award.