Learning Dynamic Interpolation for Extremely Sparse Light Fields with Wide Baselines

Mantang Guo [*]   Jing Jin [*]   Hui Liu   Junhui Hou

Department of Computer Science, City University of Hong Kong, Hong Kong SAR

{mantanguo2-c,jingjin25-c,hliu99-c}@my.cityu.edu.hk, jh.hou@cityu.edu.hk

Abstract

In this paper, we tackle the problem of dense light field (LF) reconstruction from sparsely-sampled ones with wide baselines and propose a learnable model, namely dynamic interpolation, to replace the commonly-used geometry warping operation. Specifically, with the estimated geometric relation between input views, we first construct a lightweight neural network to dynamically learn weights for interpolating neighbouring pixels from input views to synthesize each pixel of novel views independently. In contrast to the fixed and content-independent weights employed in the geometry warping operation, the learned interpolation weights implicitly incorporate the correspondences between the source and novel views and adapt to different image content information. Then, we recover the spatial correlation between the independently synthesized pixels of each novel view by referring to that of input views using a geometry-based spatial refinement module. We also constrain the angular correlation between the novel views through a disparity-oriented LF structure loss. Experimental results on LF datasets with wide baselines show that the reconstructed LFs achieve much higher PSNR/SSIM and preserve the LF parallax structure better than state-of-the-art methods. The source code is publicly available at `https://github.com/MantangGuo/DI4SLF`.

1. Introduction

Densely-sampled light field (LF) images record not only appearance but also geometry information of 3D scenes, which enable wide applications, such as 3D reconstruction [30, 24, 3], image post-refocusing [20], and virtual reality [7, 38]. However, densely-sampled LFs raise great challenges for the acquisition. For example, camera array [33] or computer-controlled gantry [29] are either bulky and ex-
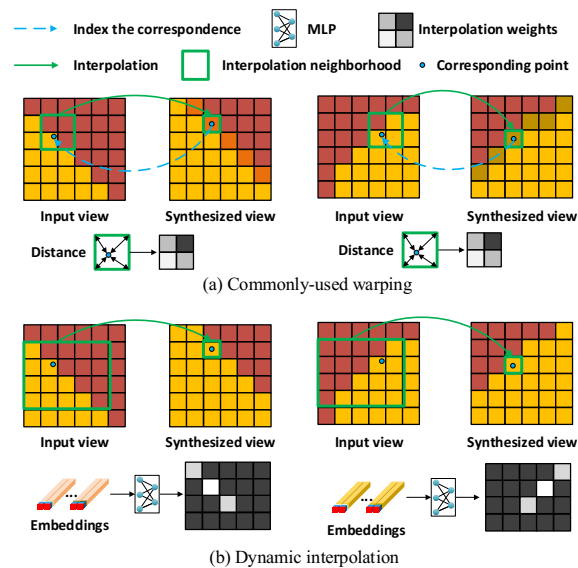
Figure 1. Comparison of the commonly-used warping operation and the proposed dynamic interpolation. In contrast to the fixed and content-independent weights employed in the warping operation (taking bilinear interpolation weights as an example), we propose to dynamically learn geometry-aware and content-adaptive interpolation weights from carefully constructed embeddings.

pensive or limited in capturing static scenes, while cost-effective commercial LF cameras [17, 21] suffer from a trade-off between the spatial and angular resolution due to the limited sensor resolution [8, 9].

Although many computational methods have been proposed to reconstruct densely-sampled LFs from sparsely-sampled ones, the wide baseline between input views remains a great challenge. To be more specific, non-depth-based methods [23, 37, 35, 31, 36, 4, 5] investigate the implicit signal distribution of LF data to learn the mapping from sparse to dense LFs. These methods inevitably suffer from the aliasing problem and lead to artifacts when the LF is extremely under-sampled. In comparison, depth-based methods [32, 12, 34, 11, 10] perform much better by employing the explicit geometry information. These methods follow the general pipeline of warping-based view synthe-

sis, and mainly focus on improving the disparity estimation and post-processing refinement. However, the reconstruction quality is still limited.

In this paper, we tackle the challenging problem of LF reconstruction from extremely sparse and wide-baseline inputs, based on an insight that commonly-used warping operation confronts with natural limitations. Specifically, the warping operation synthesizes pixels of the novel view by performing interpolation using their neighboring pixels from input views. The employed interpolation weights are determined by fitting a simple and smooth curve using a small set of neighbors, which inevitably impacts the reconstruction quality as the content information is not considered. To this end, we propose a learnable module, namely **dynamic interpolation**, to replace the commonly-used warping operation. As shown in Fig. 1, dynamic interpolation uses a lightweight neural network to dynamically predict **geometry-aware** and **content-adaptive** interpolation weights for novel view synthesis. As the pixels of the novel views are independently synthesized, we subsequently recover the spatial correlation between them by referring to that of input views using a geometry-based refinement module. We also constrain the angular correlation between the novel views through a disparity-oriented LF structure loss. Extensive experimental results demonstrate the significant superiority of the proposed model on LF datasets with wide baselines over warping-based methods as well as other state-of-the-art ones.

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

- we deeply analyze the geometry warping operation for handling the challenge of LF reconstruction from wide-baseline inputs, and figure out the essential limitation lies in the weakness of the interpolation weights; and
- we reformulate the LF reconstruction from a new perspective and propose dynamic interpolation, which is capable of overcoming the limitation of the geometry warping operation.

2. Related Work

The existing LF reconstruction methods could be roughly divided into two categories: non-learning-based methods and learning-based methods.

Non-learning-based methods usually adopt various prior assumptions to regularize the LF data, i.e., Gaussian-based priors [16, 15, 19], sparse priors [18, 23, 28], and low-rank [13]. These methods either require many sparse samplings, or have high computational complexity. Explicitly estimating the scene depth information, and then using it to warp input sub-aperture images (SAIs) to novel ones is another kind of methods for LF reconstruction. Wanner and Goldluecke [32] estimated disparity maps at input view by calcu-

lating the structure tensor of epipolar plane images (EPIs), and then used the estimated disparity maps to warp input SAIs to the novel viewpoints. This method makes the reconstruction quality rely heavily on the accuracy of the depth estimation. Zhang et al. [39] proposed a disparity-assisted phase-based method that can iteratively refine the disparity map to minimize the phase difference between the warped novel SAI and the input SAI. However, the angular positions of synthesized SAIs are restricted to the neighbor of input views, which cannot reconstruct LFs with large baselines.

Recently, many deep learning-based methods have been proposed to reconstruct dense LFs from sparse samplings. Yoon et al. [37] reconstructed novel SAIs from spatially up-sampled horizontal, vertical and surrounding SAI-pairs by using three separate networks. This method can only regress novel SAIs from adjacent ones, and could not process sparse LFs with large disparities. Wu et al. [35] used a 2-D image super-resolution network to recover high-frequency details along the angular dimension of the interpolated EPI. Analogously, Wang et al. [31] restored the high-frequency details of EPI stacks by using 3-D convolutional neural networks (CNNs). These methods process 2-D or 3-D slices of a 4-D sparse LF, which cannot fully explore the spatial-angular correlations implied in the LF. Yeung et al. [36] proposed the computational efficient spatial-angular separable convolution for reconstructing a dense LF from a sparse one in a single forward pass.

The pipeline of warping-based non-learning LF reconstruction methods is also employed by several deep learning-based ones. Kalantari et al. [12] used two sequential networks to separately estimate the disparity map at the novel view, and predicted the color of novel SAI from warped images, respectively. Wu et al. [34] extracted depth information from the sheared EPI volume, and then used it to reconstruct high angular-resolution EPIs. These methods either ignore the angular relations between synthesized SAIs, or underuse the spatial information of the input SAIs during the reconstruction. Srinivasan et al. [25] reconstructed an LF from a single 2-D image with predicting 4-D ray depths. This method only works on dataset with small disparities, and is restricted by its generalization ability. Jin et al. [10] explicitly learned the disparity map at the novel viewpoint from input SAIs. They synthesized the coarse novel SAIs individually by fusing the warped input SAIs with confidence maps. Then they used a refinement network to recover the parallax structure by exploring the complementary information from the coarse LF. Zhou et al. [40] predicted the multiplane image at a reference view by using a CNN to represent the scene's content. Then the novel view can be synthesized from the multiplane image representation with homography and alpha compositing.
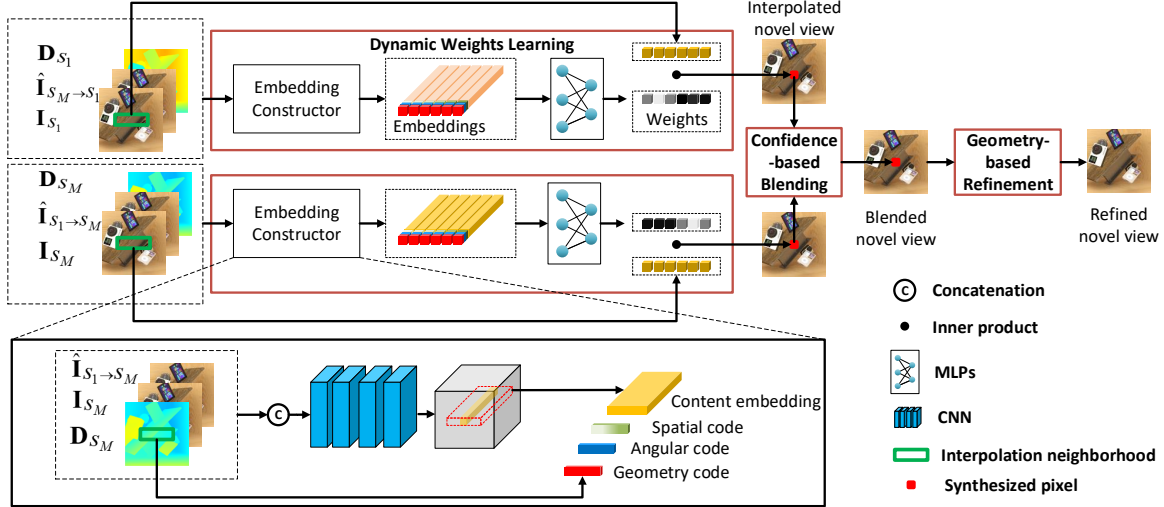
Figure 2. The flowchart of the proposed *dynamic interpolation* model for LF reconstruction from extremely sparse (taking $M = 2$ as an example) and wide-baseline inputs. Our proposed model consists of three components: dynamic weight learning, confidence-based blending, and geometry-based spatial refinement.

3. Problem Analysis

Denote by $\mathcal{L}(u, x, y) \in \mathbb{R}^{U \times H \times W}$ a 3-D LF containing U sub-aperture images (SAIs) of each spatial resolution $H \times W$, which are sampled along a 1-D straight line. The SAI at the angular position of u is denoted as $\mathbf{I}_u$. Given an extremely sparse LF with M SAIs, denoted by $\mathcal{S} = \{\mathbf{I}_{s_1}, \cdots, \mathbf{I}_{s_M}\}$, where $M \ll U$, our goal is to synthesize the unsampled SAIs, denoted by $\mathcal{T} = \{\mathbf{I}_{t_1}, \cdots, \mathbf{I}_{t_N}\}$, where $N = U - M$, so that the densely-sampled LF denoted by $\widetilde{\mathcal{L}}$ can be reconstructed as close to $\mathcal{L}$ as possible. This problem can be implicitly formulated as:

$$\widetilde{\mathcal{L}} = \mathcal{S} \bigcup \widetilde{\mathcal{T}} = f(\mathcal{S}), \tag{1}$$

where $\widetilde{\mathcal{T}}$ is the set of synthesized novel SAIs, and f denotes the mapping function to be learned. Not that we also denote the input and target SAI as $\mathbf{I}_s$ and $\mathbf{I}_t$, respectively, in the rest of the paper for simplification.

The SAIs of a 3-D LF image are the observations of the same scene from different viewpoints. Under the assumption of Lambertian and non-occlusion, projections of the same scene point will have the same intensity at different SAIs. This relation can be described as:

$$\mathbf{I}_t(\mathbf{x}_t) = \mathbf{I}_s(\mathbf{x}_t'), \tag{2}$$

where $\mathbf{x}_t = (x_t, y_t)$ is the spatial coordinate, and $\mathbf{x}_t'$ is the location of the corresponding pixel of $\mathbf{I}_t(\mathbf{x}_t)$ at $\mathbf{I}_s$. Given the disparity value of $\mathbf{I}_t(\mathbf{x}_t)$, denoted by d, $\mathbf{x}_t'$ can be easily computed as $\mathbf{x}_t' = (x_t', y_t') = (x_t + d(s - t), y_t)$. Based on this relation, the pixels of $\mathbf{I}_t$ can be estimated by collecting their corresponding pixels on $\mathbf{I}_s$. However, as the values of $\mathbf{x}_t'$ are always fractional, interpolation is required to computed the intensity of the corresponding pixel by the weighted sum of the neighboring pixels. The interpolation process can be formulated as:

$$\mathbf{I}_t(\mathbf{x}_t) = \sum_{\mathbf{x}_s \in \mathcal{P}_{\mathbf{x}_t'}} w(\mathbf{x}_s - \mathbf{x}_t'; \phi_w) \mathbf{I}_s(\mathbf{x}_s), \tag{3}$$

where $\mathcal{P}_{\mathbf{x}_t'}$ is the set of neighbors of $\mathbf{x}_t'$, and w is the function with the parameter ϕ_w which defines the interpolation weights for the pixels of $\mathbf{I}_s$.

The above mentioned procedure is adopted by the commonly-used warping operation. However, we provide an insight that this procedure has natural limitations from two aspects:

(1) It requires to estimate d to locate $\mathbf{x}_t'$. However, estimating the disparities of unsampled SAIs from the input SAIs is challenging. Moreover, as $\mathcal{P}_{\mathbf{x}_t'}$ is always a small set of pixels surrounding $\mathbf{x}_t'$, e.g., 2 (or 2×2) neighbors for linear (or bilinear) interpolation [25] and 4 (or 4×4) neighbors for cubic (or bicubic) interpolation [12], the reconstruction results greatly rely on the accuracy of the disparity estimation.

(2) The weight function $w(\cdot; \phi_w)$ is defined by fitting a simple and smooth curve using the small set $\mathcal{P}_{\mathbf{x}_t'}$, which neglects the content information. Thus, even with an accurate estimation of d, it is difficult to produce high-quality results, especially on areas with texture edges, occlusion boundaries, and non-Lambertian objects.

Base on these observations, we propose a novel model, namely *dynamic interpolation*, to synthesize $\widetilde{\mathcal{T}}$ from $\mathcal{S}$, which overcomes the limitations of the commonly-used warping operation by learning geometry-aware and content-adaptive interpolation weights.

4. Proposed Method

Overview. As shown in Fig. 2, the proposed model, namely *dynamic interpolation*, mainly consists of three components, i.e., dynamic weight learning, confidence-

based blending, and geometry-based spatial refinement. Specifically, we first independently synthesize each pixel of $\widetilde{\mathbf{I}}_t$ by applying interpolation over its neighboring pixels from $\mathbf{I}_s$. In contrast to commonly-used warping operation, the interpolation weights used in our model are dynamically learned. We also estimate confidence maps to blend the pixels interpolated from different input SAIs, which further handles the occlusion problems. Then, we recover the spatial correlation between the independently synthesized pixels by referring to that of $\mathbf{I}_s$.

In this paper, we set $M = 2$. It is worth noting that our framework could be straightforwardly extended to 4-D LFs with larger M. In what follows, we will introduce the technical details of the proposed method.

4.1. Dynamic Weight Learning

This module aims at learning the interpolation weights to independently synthesize each pixel of a novel SAI, denoted by $\mathbf{I}_t(\mathbf{x}_t)$, from $\mathbf{I}_s$. The interpolation weight for each pixel $\mathbf{I}_s(\mathbf{x}_s)$ is predicted by a multilayer perceptron (MLP), and the following information is embedded in the MLP:

(1) **The correspondence relation** between $\mathbf{I}_t$ and $\mathbf{I}_s$, which helps to implicitly locate the corresponding pixel of $\mathbf{I}_t(\mathbf{x_t})$ in $\mathbf{I}_s$. The correspondence embedding consists of three components, i.e., a geometric code $E_{\mathbf{x}_t,\mathbf{x}_s}^{geo}$, a spatial code $\mathbf{E}_{\mathbf{x}_t,\mathbf{x}_s}^{spa}$, and an angular code $E_{\mathbf{x}_t,\mathbf{x}_s}^{ang}$. Specifically, $E_{\mathbf{x}_t,\mathbf{x}_s}^{geo}$ is the disparity value of $\mathbf{I}_s(\mathbf{x}_s)$, i.e.,

$$E_{\mathbf{x}_t,\mathbf{x}_s}^{geo} = \mathbf{D}_s(\mathbf{x}_s), \tag{4}$$

where $\mathbf{D}_s$ is the disparity map of $\mathbf{I}_s$, which is estimated from $\mathcal{S}$ using a pre-trained optical-flow model. $E_{\mathbf{x}_t,\mathbf{x}_s}^{spa}$ and $E_{\mathbf{x}_t,\mathbf{x}_s}^{ang}$ describe the the spatial and angular distance between $\mathbf{I}_t(\mathbf{x}_t)$ and $\mathbf{I}_s(\mathbf{x}_s)$, i.e.,

$$\mathbf{E}_{\mathbf{x}_t,\mathbf{x}_s}^{spa} = \mathbf{x}_s - \mathbf{x}_t, \ E_{\mathbf{x}_t,\mathbf{x}_s}^{ang} = s - t. \tag{5}$$

These information can directly determine whether $\mathbf{I}_s(\mathbf{x}_s)$ corresponds to $\mathbf{I}_t(\mathbf{x}_t)$ under the estimated geometric relation, and thus, greatly helps the MLP to locate informative pixels in $\mathbf{I}_s$ and allocate larger weights to them.

(2) **The content information** around $\mathbf{I}_s(\mathbf{x}_s)$, which helps to understand complicated scenarios, such as texture edges, occlusion boundaries, and non-Lambertian objects. To construct the content embedding, denoted by $\mathbf{E}_{\mathbf{x}_t,\mathbf{x}_s}^{ctt}$, we first backward warp the other SAI in $\mathcal{S}$ to $\mathbf{I}_s$ based on $\mathbf{D}_s$, and the resulting image is denoted by $\widehat{\mathbf{I}}_{s'\to s}$. Then, we employ a sub-CNN $f_c(\cdot)$ to learn the content information, i.e.,

$$\mathbf{E}_{\mathbf{x}_t,\mathbf{x}_s}^{ctt} = f_c\left(\mathbf{x}_s, \mathbf{x}_t, \mathbf{I}_s, \widehat{\mathbf{I}}_{s'\to s}, \mathbf{D}_s\right). \tag{6}$$

It is expected that f_c is able to detect the texture edges of $\mathbf{I}_s$ and understand the occlusion and non-Lambertian relations by comparing $\mathbf{I}_s$ and $\widehat{\mathbf{I}}_{s'\to s}$ with the assistance of $\mathbf{D}_s$.

Finally, the geometry and content embedding, denoted by $\mathbf{E}_{\mathbf{x}_t,\mathbf{x}_s}$, is constructed as:

$$\mathbf{E}_{\mathbf{x}_t,\mathbf{x}_s} = \mathsf{CAT}\left(E_{\mathbf{x}_t,\mathbf{x}_s}^{geo}, \mathbf{E}_{\mathbf{x}_t,\mathbf{x}_s}^{spa}, E_{\mathbf{x}_t,\mathbf{x}_s}^{ang}, \mathbf{E}_{\mathbf{x}_t,\mathbf{x}_s}^{ctt}\right), \tag{7}$$

where $\mathsf{CAT}(\cdot)$ is the concatenation operation, and the interpolation weights for $\mathbf{I}_s(\mathbf{x}_s)$ to synthesize $\mathbf{I}_t(\mathbf{x}_t)$, denoted by $W_{\mathbf{x}_t,\mathbf{x}_s}$, is predicted as:

$$W_{\mathbf{x}_t,\mathbf{x}_s} = f_w\left(\mathbf{E}_{\mathbf{x}_t,\mathbf{x}_s}\right), \tag{8}$$

where $f_w(\cdot)$ is the learnable MLP.

To reduce the computational cost, the interpolation is performed over the neighborhood of $\mathbf{I}_t(\mathbf{x}_t)$ in $\mathbf{I}_s$, instead of the whole range of $\mathbf{I}_s$. Suppose the disparity range of $\mathbf{I}_t$ is $[-d_{max}, d_{max}]$, the neighborhood of $\mathbf{I}_t(\mathbf{x}_t)$ in $\mathbf{I}_s$ is defined as: $\mathcal{P}_{\mathbf{x}_t} = \{\mathbf{x} = (x,y)|x_t - d_{max}(s-t) \le x \le x_t + d_{max}(s-t), y = y_t\}$. Then, we predict the intensity of $\mathbf{I}_t(\mathbf{x}_t)$ by applying interpolation on $\mathcal{P}_{\mathbf{x}_t}$ based on the learned weights, and the predicted result is denoted by $\widetilde{\mathbf{I}}_{s\to t}(\mathbf{x}_t)$, i.e.,

$$\widetilde{\mathbf{I}}_{s\to t}(\mathbf{x}_t) = \sum_{\mathbf{x}_s \in \mathcal{P}_{\mathbf{x}_t}} W_{\mathbf{x}_t,\mathbf{x}_s} \mathbf{I}_s(\mathbf{x}_s). \tag{9}$$

Remark: Compared with the commonly-used warping operation, our dynamic interpolation has the following advantages:

(1) Instead of relying on the disparity estimation accuracy of the novel SAI, we utilize the disparity map directly estimated between input SAIs, which is much more reliable. Moreover, the geometry information is implicitly incorporated by learning the weights for each pixel in the possibly maximum neighborhood, which might improve the tolerance of the disparity estimation error.

(2) Instead of fitting a simple curve using a small set of pixels for interpolation, we learn the weights using an MLP and provide content information over a relatively large field to make the weights adaptive to various and complicated neighboring correlations.

4.2. Confidence-based Blending

Although the weight learning module has the ability of handling the problem of occlusion boundaries by embedding the content information, it is still difficult to synthesize the pixels whose correspondences are completely occluded in $\mathbf{I}_s$ by interpolation on only one of the input SAIs. Fortunately, the object occluded from one viewpoint might be visible from another one. Therefore, we blend the images synthesized from different input SAIs under the guidance of their confidence maps, which indicate the non-occlusion pixels with higher values.

To predict the confidence value for each pixel position $\mathbf{x}_t$ in the synthesized SAI, we first aggregate the geometry and content embeddings for each neighbors of $\mathbf{x}_t$ in $\mathbf{I}_s$ by concatenation, and then apply another MLP, denoted by $f_b(\cdot)$, on the aggregated feature, i.e.,

$$\widetilde{\mathbf{C}}_{s\to t}(\mathbf{x}_t) = f_b\left(\mathsf{CAT}\{\mathbf{E}_{\mathbf{x}_t,\mathbf{x}_s}|\mathbf{x}_s \in \mathcal{P}_{\mathbf{x}_t}\}\right), \tag{10}$$

where $\widetilde{\mathbf{C}}_{s\to t}$ is the confidence map for $\widetilde{\mathbf{I}}_{s\to t}$. Based on the learned confidence map, the SAIs synthesized from different input SAIs are combined to produce the intermediate

result of the novel SAI, denoted by $\widetilde{\mathbf{I}}_t^b$, i.e.,

$$\widetilde{\mathbf{I}}_t^b = \sum_{s \in \{s_1, \cdots, s_M\}} \widetilde{\mathbf{C}}_{s \to t} \odot \widetilde{\mathbf{I}}_{s \to t}, \qquad (11)$$

where $\odot$ is the element-wise multiplication operator.

4.3. Geometry-based Spatial Refinement

As pixels in $\widetilde{\mathbf{I}}_t^b$ are independently synthesized, the spatial correlations among them are not considered. To further enhance the quality of $\widetilde{\mathbf{I}}_t^b$, we propose a refinement module to recover its spatial correlation by inferring that from the input SAIs using a sub-CNN. Considering the wide baseline between $\widetilde{\mathbf{I}}_t^b$ and $\mathbf{I}_s$, directly applying a network will have difficulties to perceive the corresponding information from $\mathbf{I}_s$. Therefore, we adopt a geometry-based spatial refinement, which first explicitly locates the correspondences in $\mathbf{I}_s$ at the patch level, and then learn the spatial correlations from $\mathbf{I}_s$ to refine $\widetilde{\mathbf{I}}_t^b$.

Let $\widetilde{\mathbf{H}}_t^{\mathbf{x}_t^o}$ denote a patch of $\widetilde{\mathbf{I}}_t^b$ centered at $\mathbf{x}_t^o = (x_t^o, y_t^o)$. To locate its similar patch in $\mathbf{I}_s$, we first estimate the disparity map of $\widetilde{\mathbf{I}}_t^b$ by forward warping $\mathbf{D}_s$, resulting in $\widetilde{\mathbf{D}}_t$, and then calculate the patch-level disparity of $\widetilde{\mathbf{H}}_t^{\mathbf{x}_t^o}$ by averaging the disparity over all of its contained pixels, leading to $\widetilde{d}_h$. Then, the central position of the corresponding patch of $\widetilde{\mathbf{H}}_t^{\mathbf{x}_t^o}$ at $\mathbf{I}_s$, denoted by $\mathbf{x}_s^o = (x_s^o, y_s^o)$, can be estimated as:

$$x_s^o = x_t^o + \widetilde{d}_h(s - t). \qquad (12)$$

Based on $\mathbf{x}_s^o$, we can collect the corresponding patch of $\widetilde{\mathbf{H}}_t^{\mathbf{x}_t^o}$ at $\mathbf{I}_s$, which is denoted by $\mathbf{H}_s^{\mathbf{x}_s^o}$.

To recover the spatial correlation among pixels in $\widetilde{\mathbf{H}}_t^{\mathbf{x}_t^o}$, we feed the concatenation of $\widetilde{\mathbf{H}}_t^{\mathbf{x}_t^o}$ and its corresponding patches in all the input SAIs, i.e., $\{\mathbf{H}_s^{\mathbf{x}_s^o} | s \in \{s_1, \cdots, s_M\}\}$, into a sub-CNN to predict a residual map for refinement. We then merge the refined patches to produce the final prediction of the novel SAI, i.e.,

$$\widetilde{\mathbf{I}}_t = f_r(\widetilde{\mathbf{I}}_t^b), \qquad (13)$$

where $f_r(\cdot)$ denotes the geometry-based spatial refinement module.

4.4. Disparity-oriented Loss

The final and intermediate predictions of the novel SAIs are supervised by the ground-truth one, i.e., the loss function for the reconstruction of $\mathbf{I}_t$ is defined as:

$$\ell_t^r = \left\| \widetilde{\mathbf{I}}_t - \mathbf{I}_t \right\|_1 + \left\| \widetilde{\mathbf{I}}_t^b - \mathbf{I}_t \right\|_1 + \sum_{s \in \{s_1, \cdots, s_M\}} \left\| \widetilde{\mathbf{I}}_{s \to t} - \mathbf{I}_t \right\|_1 . \qquad (14)$$

Moreover, as each novel SAI is reconstructed individually, we propose a disparity-oriented LF structure loss to constrain the angular correlation between them. The relation described in Eq. (2) can be constrained by minimizing the gradients along the directions of the straight lines in EPIs of the LF. Considering the existence of occlusions and non-Lambertian, we instead minimizing the distance between the gradients of predicted EPIs and the ground-truth ones. Note that the directions of EPI lines are located under the guidance of the ground-truth disparity of the LFs, which are easily available in the training datasets. Such a disparity-oriented LF structure loss can be formulated as

$$\ell^d = \left\| \nabla_d \widetilde{\mathcal{E}} - \nabla_d \mathcal{E} \right\|_1, \qquad (15)$$

where ∇_d is the gradient operator along the direction defined by the ground-truth disparity d of each pixel, and $\widetilde{\mathcal{E}}$ and $\mathcal{E}$ are the EPIs of $\widetilde{\mathcal{L}}$ and $\mathcal{L}$, respectively.

Our framework is end-to-end trained using the final objective function defined as: $\ell = \sum_{t \in \{t_1, \cdots, t_N\}} \ell_t^r + \lambda \ell^d$, where $\lambda \geq 0$ is the weight factor for the disparity-oriented loss.

5. Experimental Results

5.1. Training Details and Datasets

Both the content embedding network $f_c(\cdot)$ and the spatial refinement network $f_r(\cdot)$ are 2-D CNNs that consist of 4 residual blocks [6] with the kernel of size 3×3. Zeros-padding was applied to keep the spatial size unchanged. We refer readers to the *Supplementary Material* for the detailed network architecture. At each iteration of the training phase, a 32×32 patch randomly cropped from the LF image was synthesized. We used the *unfold* function in *PyTorch* to efficiently locate the neighborhood of each synthesized pixel.

The batch size was empirically set to 1. The learning rate was initially set to $1e^{-4}$ and reduced to $1e^{-5}$ after 8000 epochs. We used Adam [14] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ as the optimizer.

We trained our framework using 29 LF images from the Inria Sparse LF dataset [22]. Each LF image contains 9×9 SAIs with a disparity range of $[-20, 20]$ between adjacent SAIs. We took the 3rd and 7th SAIs at the same row as inputs, which have a wide baseline up to 80 pixels, to train the framework. The test dataset consists of 7 LF images from the Inria Sparse LF dataset [22] and 14 LF images from the MPI LF archive [1]. Note that MPI [1] is a high angular-resolution LF dataset where each LF image contains 101 SAIs distributed on a scanline. Thus, we can construct testing LFs with different baselines by sampling SAIs with different intervals (see details in Section 5.2).

5.2. Comparison with State-of-the-art Methods

We compared the proposed method with three state-of-the-art deep learning-based methods for LF reconstruction, including Kalantari *et al.* [12], Wu *et al.* [34], and Jin *et al.* [10]. All the methods were retrained on the same dataset

Table 1. Quantitative comparisons (PSNR/SSIM) of different methods over the Inria Sparse LF dataset [22].

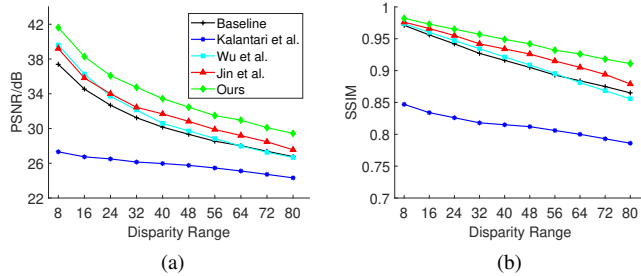| Light Field	Disparity range	Baseline	Kalantari *et al.* [12]	Wu *et al.* [34]	Jin *et al.* [10]	Ours (PWCNet)	Ours (RAFT)
Electro_devices	[-19.6, 32.8]	28.49/0.871	24.66/0.691	28.51/0.866	32.77/0.936	33.04/0.941	**35.43/0.960**
Flying_furniture	[-34.0, 62.4]	28.39/0.838	28.83/0.784	27.38/0.783	31.69/0.896	30.06/0.881	**33.93/0.935**
Coffee_beans_vases	[10.8, 58.4]	27.17/0.886	21.54/0.579	23.04/0.836	28.08/0.927	**29.63**/0.936	29.55/**0.943**
Dinosaur	[-57.6, 72.8]	23.00/0.773	22.21/0.731	23.07/0.788	26.61/0.897	24.94/0.861	**27.50/0.904**
Flowers	[-40.4, 66.0]	23.05/0.757	21.96/0.667	23.52/0.767	24.36/0.842	**25.24/0.860**	24.86/0.849
Rooster_clock	[-34.4, 21.2]	31.43/0.904	22.71/0.710	29.05/0.887	27.69/0.929	35.90/0.946	**38.16/0.966**
Smiling_crowd	[-40.4, 64.8]	18.87/0.722	17.01/0.596	19.52/0.729	21.01/0.822	20.90/0.824	**22.87/0.877**
Average		25.77/0.821	22.70/0.680	25.05/0.802	27.46/0.893	28.53/0.893	**30.33/0.919**

(a)   (b)

Figure 3. Quantitative comparisons (PSNR/SSIM) of different methods under different disparity ranges (pixels) between input SAIs on the MPI LF dataset [1]. All subfigures share the same legend shown in the first one.

with the officially released codes and suggested configurations. Note that the 2-D angular convolutional layers used by Jin *et al.* [10] were degenerated to 1-D convolutional layers to adapt to the 3-D LF.

To verify the advantage of the proposed dynamic interpolation in comparison to the commonly-used warping operation, we developed a baseline model by replacing the dynamic weight learning module with a disparity-based warping operation while leaving the confidence-based blending and the geometry-based refinement unchanged. Note that to ensure fair comparisons, the disparity maps of the novel views used for warping were estimated from the same inputs as Ours, i.e., the disparity maps of the input views from a pre-trained optical-flow model.

Moreover, to demonstrate the effects of the input disparity estimation accuracy on the performance of our model, we adopted two different optical-flow methods, namely RAFT [27] and PWCNet [26], and used them to separately train two models, denoted as Ours (RAFT) and Ours (PWC-Net), respectively.

Quantitative comparisons on the Inria Sparse LF dataset. We reconstructed 5×5 LF images in the Inria dataset [22] in a row-by-row manner, and calculated the PSNR and SSIM values between the reconstructed LFs and ground-truth ones in Y channel to quantitatively evaluate different methods. Table 1 lists the results, where it can be observed that:

- Ours (RAFT) achieves significantly higher PSNR and SSIM than Baseline. Although the disparity maps between input SAIs are also provided to Baseline, its performance is still limited by the fixed and content-independent weights employed in the commonly-used warping operation, which demonstrates the advantage of learned dynamic weights in our method;

- Ours (PWCNet) is worse than Ours (RAFT) but still much better than all the compared methods. Although the estimated disparity maps with different levels of quality are involved, our method may adaptively aggregate content features to provide gains for learning dynamic weights, which can demonstrate the advantages of learning dynamic weights over commonly-used ones. Besides, it is highly expected that our framework will be further improved with more powerful and advanced optical flow estimation proposed in future;

- our method achieves higher performance than Wu *et al.* [34]. The reason maybe that Wu *et al.* [34] reconstructed LFs in the perspective of reconstructing 2-D EPIs, and neglected modeling the correlations between two spatial dimensions, which limits the performance. On the contrary, our method employs a geometry-based refinement module to refine the correlations among pixels of novel SAIs, and further improves the reconstruction quality; and

- both Kalantari *et al.* [12] and Jin *et al.* [10] achieve worse performances than our method. In addition to that they cannot estimate disparities well from extremely sparse LFs with a limited receptive field of CNNs, they also cannot handle severe artifacts brought by the warping operation. By contrast, our method can effectively mitigate the warping errors in reconstructing extremely sparse LFs by dynamically learning content-adaptive weights for each pixel of the novel SAI. Besides, the geometry-based refinement module can further refine novel SAIs to improve the quality.

Quantitative comparisons on the MPI LF dataset. Furthermore, we also evaluated the robustness of different methods under different disparity ranges on the MPI dataset
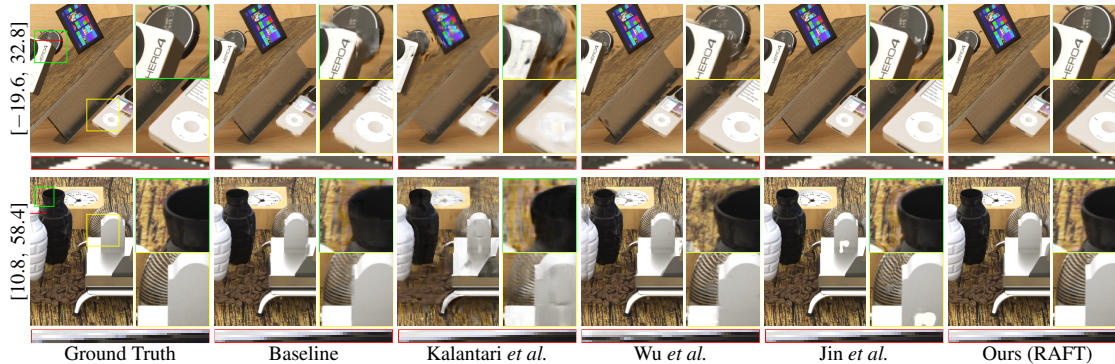
Figure 4. Visual comparisons of reconstructed LFs from different methods over the Inria Sparse LF dataset [22]. The disparity range between input SAIs of each LF is shown on the left.
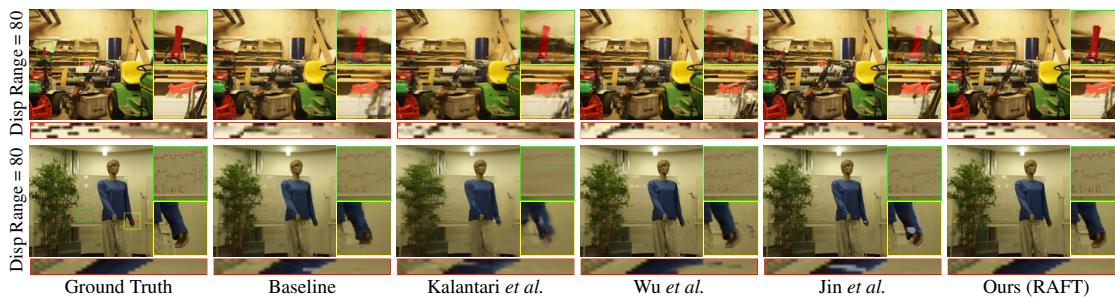

Figure 5. Visual comparisons of reconstructed LFs from different methods over the MPI LF dataset [1]. The disparity range between input SAIs reaches 80 pixels for each reconstructed LF.

Table 2. Comparisons of running time (in seconds per view) and model parameter size (M) of different methods on the Inria Sparse LF dataset [22].

	Baseline	Kalantari *et al.* [12]	Wu *et al.* [34]	Jin *et al.* [10]	Ours
Time	**0.65**	6.32	27.77	0.69	2.78
# Params	0.95	2.55	**0.24**	2.22	0.69

[1]. Each scene contains a high-angular densely-sampled LF image composed of 101 SAIs distributed on a scanline with spatial resolution 720×960. The disparity between adjacent SAI is around 1 pixel. We can sample SAIs with different intervals along the angular dimension to construct LFs with different disparity ranges. Specifically, we separately set 10 disparity ranges from 8 to 80 pixels between two input SAIs. For each disparity range, we evenly sampled 3 SAIs between two input SAIs as ground truth. The PSNR/SSIM shown in Fig. 3 indicates that although the reconstruction quality of all methods decreases along with the increase of the disparity range, Ours consistently achieves better reconstruction quality than the other methods under all disparity ranges, demonstrating the robustness of our method towards different disparity ranges.

Qualitative comparisons. We visually compared reconstructed wide-baseline LFs by different methods. One of the challenges for the wide-baseline LF reconstruction is fairly reconstructing a large number of occlusion regions while are not so many in narrow-baseline ones. From reconstructed SAIs and zoomed-in regions shown in Fig. 4

and Fig. 5, it can be observed that our method can produce sharp edges at the occlusion boundaries, while other methods produce either severe distortions or heavy blurry effects at these regions. Moreover, our method can produce better high-frequency details at texture regions than other methods, which demonstrate that our method can still achieve high-quality LF reconstructions even under such large input disparity range. Please refer to the *Supplementary Material* for more visual results.

Comparisons of the LF parallax structure. Moreover, as the parallax structure is one of the most important values of LF data, we thus managed to compare the parallax structures of LFs reconstructed by different methods. On the one hand, comparisons of EPIs in Figs. 4 and 5 indicate that our method can preserve clearer linear structures than other methods, even for lines corresponding to large-disparity regions, validating the strong ability of our method in preserving parallax structure on extremely sparse LFs. On the other hand, it is expected that depth/disparity maps estimated from high-quality LFs shall be close to those are estimated from ground-truth ones. Thus, we compared the depth maps estimated from reconstructed LFs of different methods by a identical LF-based depth estimation method [2]. As shown in Fig. 6, our method can produce sharper edges at occlusion boundaries and preserve smoothness at regions with uniform depth, which are closest to the ground truth. Such observations also demonstrate the advantage of our method on preserving the LF parallax structure.
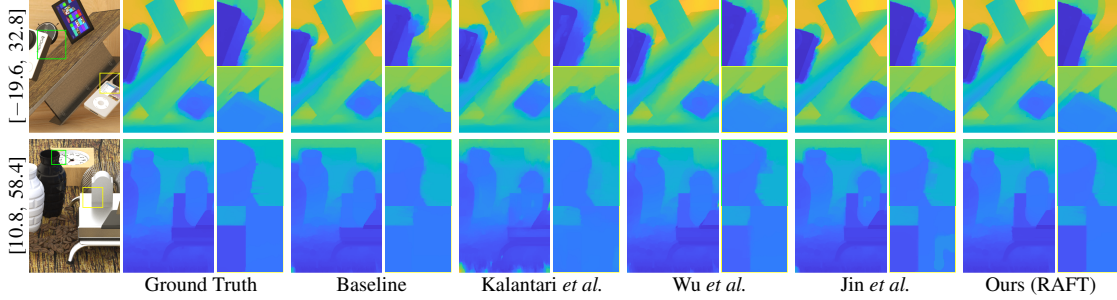
Figure 6. Visual comparisons of estimated depth maps from different methods over the Inria Sparse LF dataset [22]. The disparity range between input SAIs of each LF is shown on the left.

Efficiency comparisons. We compared the efficiency and model size of different methods. All the methods were implemented on a Linux server with Intel CPU E5-2699 @ 2.20GHz, 128GB RAM and Tesla V100. As listed in Table 2, we can see that Ours is much faster than Kalantari *et al.* [12] and Wu *et al.* [34] but slower than Baseline and Jin *et al.* [10]. Besides, our model size is smaller than Baseline, Kalantari *et al.* [12], and Jin *et al.* [10], but larger than Wu *et al.* [34]. Taking the reconstruction accuracy, efficiency, and model size together, we believe our method is the best.

5.3. Ablation Study

We carried out comprehensive ablation studies to validate the effectiveness of three key components involved in our framework, i.e., content embedding, the geometry-based refinement module, and the disparity-oriented loss term. Specifically, each component was sequentially added to the base model until all the three components were included to form the complete model. As shown in Table 3 and Fig. 7, it can be seen that there is a significant increase of performance when adding the content embedding to the base model, which verifies the advantage brought by detecting the texture edges of input views, and understanding the occlusion and non-Lambertian relations between input views. The visual comparisons shown in Fig. 7 (examples 1-3) also verify the advantage. Moreover, the geometry-based refinement module can also bring around 0.3 dB increase of PSNR based on model with the content embedding. The examples 4-5 in Fig. 7 also shown that some fine structures such as delicate objects and textures are obviously broken without this module. It verifies the effectiveness of our refining pixel correlations in novel SAI through being guided by the correct ones from input SAIs. By comparing results in the last two rows in Table 3 and the EPIs in examples 6-7 in Fig. 7, we can see that supervising the parallax structure by the ground-truth disparity during training is helpful to reconstruct high-quality LFs. We also provide the intermediate visual results before and after confidence-based blending in Fig. 8, where it can be observed that the confidence-based blending handles the occlusion regions by leveraging the advantages of the left and right results under

Table 3. Ablation study. "×" denotes that the corresponding component is not included, while "√" denotes being included.

Content embedding	Geometry-based refinement	Disparity-oriented loss	PSNR	SSIM
×	×	×	29.05	0.901
√	×	×	29.83	0.915
√	√	×	30.14	**0.920**
√	√	√	**30.33**	0.919

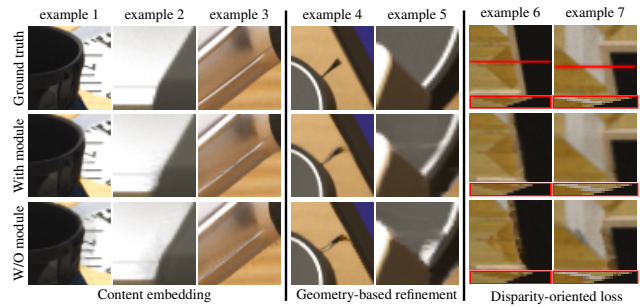the guidance of their confidence maps.

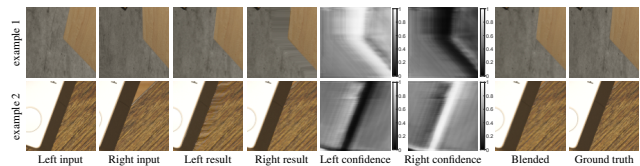Figure 7. Effectiveness of the three key modules.

Figure 8. Effectiveness of confidence-based blending.

6. Conclusion

We have presented a learning-based method for densely-sampled LF reconstruction from extremely sparse ones. More precisely, we focused on addressing the challenging problem of wide-baseline inputs and proposed a novel dynamic interpolation model. By learning geometry-aware and content-adaptive interpolation weights via a lightweight neural network, our method overcomes the limitations of commonly-used warping operation, and efficiently reconstructs LFs with much higher quality, compared with state-of-the-art methods.

References

[1] Vamsi Kiran Adhikarla, Marek Vinkler, Denis Sumin, Rafał Mantiuk, Karol Myszkowski, Hans-Peter Seidel, and Piotr Didyk. Towards a quality metric for dense light fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 58–67, 2017.

[2] Jie Chen, Junhui Hou, Yun Ni, and Lap-Pui Chau. Accurate light field depth estimation with superpixel regularization over partially occluded regions. *IEEE Transactions on Image Processing*, 27(10):4889–4900, 2018.

[3] Chunle Guo, Jing Jin, Junhui Hou, and Jie Chen. Accurate light field depth estimation via an occlusion-aware network. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020.

[4] Mantang Guo, Junhui Hou, Jing Jin, Jie Chen, and Lap-Pui Chau. Deep spatial-angular regularization for compressive light field reconstruction over coded apertures. In *European Conference on Computer Vision (ECCV)*, pages 278–294. Springer, 2020.

[5] Mantang Guo, Junhui Hou, Jing Jin, Jie Chen, and Lap-Pui Chau. Deep spatial-angular regularization for light field imaging, denoising, and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] Fu-Chung Huang, Kevin Chen, and Gordon Wetzstein. The light field stereoscope: immersive computer graphics via factored near-eye light field displays with focus cues. *ACM Transactions on Graphics*, 34(4):60, 2015.

[8] Jing Jin, Junhui Hou, Jie Chen, and Sam Kwong. Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2260–2269, 2020.

[9] Jing Jin, Junhui Hou, Jie Chen, Sam Kwong, and Jingyi Yu. Light field super-resolution via attention-guided fusion of hybrid lenses. In *ACM International Conference on Multimedia (ACM MM)*, pages 193–201, 2020.

[10] Jing Jin, Junhui Hou, Jie Chen, Huanqiang Zeng, Sam Kwong, and Jingyi Yu. Deep coarse-to-fine dense light field reconstruction with flexible sampling and geometry-aware fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[11] Jing Jin, Junhui Hou, Hui Yuan, and Sam Kwong. Learning light field angular super-resolution via a geometry-aware network. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 11141–11148, 2020.

[12] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics*, 35(6):1–10, 2016.

[13] Mahdad Hosseini Kamal, Barmak Heshmat, Ramesh Raskar, Pierre Vandergheynst, and Gordon Wetzstein. Tensor low-rank and sparse light field photography. *Computer Vision and Image Understanding*, 145:172–181, 2016.

[14] Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic gradient descent. In *International Conference on Learning Representations (ICLR)*, pages 1–15, 2015.

[15] Anat Levin and Fredo Durand. Linear view synthesis using a dimensionality gap light field prior. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1831–1838. IEEE, 2010.

[16] Anat Levin, William T Freeman, and Frédo Durand. Understanding camera trade-offs through a bayesian analysis of light field projections. In *European Conference on Computer Vision (ECCV)*, pages 88–101. Springer, 2008.

[17] Lytro. http://lytro.com, 2016.

[18] Kshitij Marwah, Gordon Wetzstein, Yosuke Bando, and Ramesh Raskar. Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Transactions on Graphics*, 32(4):1–12, 2013.

[19] Kaushik Mitra and Ashok Veeraraghavan. Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 22–28. IEEE, 2012.

[20] Ren Ng et al. *Digital light field photography*. Stanford University, 2006.

[21] RayTrix. 3d light field camera technology. https://raytrix.de/.

[22] Jinglei Shi, Xiaoran Jiang, and Christine Guillemot. A framework for learning depth from a flexible subset of dense and sparse light field views. *IEEE Transactions on Image Processing*, 28(12):5867–5880, 2019.

[23] Lixin Shi, Haitham Hassanieh, Abe Davis, Dina Katabi, and Fredo Durand. Light field reconstruction using sparsity in the continuous fourier domain. *ACM Transactions on Graphics*, 34(1):1–13, 2014.

[24] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim. Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4748–4757, 2018.

[25] Pratul P Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. Learning to synthesize a 4d rgbd light field from a single image. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2243–2251, 2017.

[26] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8934–8943, 2018.

[27] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision (ECCV)*, pages 402–419. Springer, 2020.

[28] Suren Vagharshakyan, Robert Bregovic, and Atanas Gotchev. Light field reconstruction using shearlet transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):133–147, 2017.

[29] Vaish Vaibhav and Adams Andrew. The (new) stanford light field archive. http://lightfield.stanford.edu/acq.html.

[30] Tingchun Wang, Alexei A. Efros, and Ravi Ramamoorthi. Depth estimation with occlusion modeling using light-field cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2170–2181, 2016.

[31] Yunlong Wang, Fei Liu, Zilei Wang, Guangqi Hou, Zhenan Sun, and Tieniu Tan. End-to-end view synthesis for light field imaging with pseudo 4dcnn. In *European Conference on Computer Vision (ECCV)*, pages 333–348, 2018.

[32] Sven Wanner and Bastian Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):606–619, 2013.

[33] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. High performance imaging using large camera arrays. In *ACM Transactions on Graphics*, pages 765–776. 2005.

[34] Gaochang Wu, Yebin Liu, Qionghai Dai, and Tianyou Chai. Learning sheared epi structure for light field reconstruction. *IEEE Transactions on Image Processing*, 28(7):3261–3273, 2019.

[35] Gaochang Wu, Mandan Zhao, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field reconstruction using deep convolutional network on epi. In *IEEE Confer-ence on Computer Vision and Pattern Recognition (CVPR)*, pages 6319–6327. IEEE, 2017.

[36] Henry Wing Fung Yeung, Junhui Hou, Jie Chen, Yuk Ying Chung, and Xiaoming Chen. Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues. In *European Conference on Computer Vision (ECCV)*, pages 137–152, 2018.

[37] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Learning a deep convolutional network for light-field image super-resolution. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 24–32. IEEE, 2015.

[38] Jingyi Yu. A light-field journey to virtual reality. *IEEE MultiMedia*, 24(2):104–112, 2017.

[39] Zhoutong Zhang, Yebin Liu, and Qionghai Dai. Light field from micro-baseline image pair. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3800–3809. IEEE, 2015.

[40] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4):1–12, 2018.