Light Field Reconstruction via Attention-Guided Deep Fusion of Hybrid Lenses

Jing Jin, Hui Liu, Junhui Hou, Senior Member, IEEE, and Hongkai Xiong, Senior Member, IEEE

Abstract—This paper explores the problem of reconstructing high-resolution light field (LF) images from hybrid lenses, including a high-resolution camera surrounded by multiple low-resolution cameras. The performance of existing methods is still limited, as they produce either blurry results on plain textured areas or distortions around depth discontinuous boundaries. To tackle this challenge, we propose a novel end-to-end learning-based approach, which can comprehensively utilize the specific characteristics of the input from two complementary and parallel perspectives. Specifically, one module regresses a spatially consistent intermediate estimation by learning a deep multidimensional and cross-domain feature representation, while the other module warps another intermediate estimation, which maintains the high-frequency textures, by propagating the information of the high-resolution view. We finally leverage the advantages of the two intermediate estimations adaptively via the learned attention maps, leading to the final high-resolution LF image with satisfactory results on both plain textured areas and depth discontinuous boundaries. Besides, to promote the effectiveness of our method trained with simulated hybrid data on real hybrid data captured by a hybrid LF imaging system, we carefully design the network architecture and the training strategy. Extensive experiments on both real and simulated hybrid data demonstrate the significant superiority of our approach over state-of-the-art ones. To the best of our knowledge, this is the first end-to-end deep learning method for LF reconstruction from a real hybrid input. We believe our framework could potentially decrease the cost of high-resolution LF data acquisition and benefit LF data storage and transmission. The code will be publicly available at https://github.com/jingjin25/LFhybridSR-Fusion.

Index Terms—Light field, super-resolution, hybrid imaging system, deep learning, attention.

I. INTRODUCTION

The light field (LF) describes all light rays through every point along every direction in a free space [1]. An LF image can be interpreted as multiple views observed from viewpoints regularly distributed over a 2-D grid. Therefore, LF images contain not only color information but also geometric structure of the scene in an implicit manner. The rich information enables many applications such as 3-D reconstruction [2], image post-refocusing [3], material recognition [4], saliency detection [5], and motion deblurring [6]. Recent research also demonstrates that LF is a promising media for virtual/augment reality [7], [8].

A high-quality LF image can be captured by a densely positioned array of high-resolution (HR) cameras. However, it is neither practical nor necessary to do so with so many separate HR units. Recent commercialized LF cameras provide a convenient way to capture LF images. However, the captured LF images always suffer from low spatial resolution due to the limitation of sensor resolution. To overcome this limitation, many methods for reconstructing HR LF images have been proposed [9]–[20]. Among them, LF reconstruction with a hybrid input is a promising way. A hybrid LF imaging system can be built by a sparse grid of low-resolution (LR) image sensors that surround a central HR camera [18], [20], as shown in Fig. 1. These heterogeneous sensors simultaneously sample along the angular and spatial dimensions of the LF at different sampling rates, and provide sufficient information for subsequent algorithms to calculate an HR LF. The LR views are useful for recording the geometry information of the scene, while the HR central view captures delicate textures and high-frequency information of the scene. To produce an HR LF image, a post-process algorithm is necessary to combine the information of the hybrid input.

Although multiple algorithms have been proposed to reconstruct an HR LF from the hybrid input [17]–[20], they still have limited performance. Generally, these methods comprise several steps that are independently designed, and the final results would be compromised by any inaccuracy of each step. Furthermore, these methods fail to fully describe the complicated relation between the HR central view and the LR side views as well as the one within the high-dimensional LF image.

We propose a learning-based framework to reconstruct an HR LF image with a hybrid input in an end-to-end manner. The proposed framework produces impressive performance. As illustrated in Figure 1, our framework achieves the goal with two complementary and parallel research lines, namely SR-Net and Warp-Net, and the advantages of them are combined via attention-guided fusion. The SR-Net up-samples the LR views to the desired resolution by learning a deep representation from both components of the hybrid input. The results of this module are spatially consistent with respect to the scene content, but always blurred, especially when the up-sampling scale is relatively large. In Warp-Net, the HR view is warped to synthesize an HR LF using the disparity maps estimated from the LR views. The predictions by this module inherit the delicate textures and high-frequency information from the HR view, but always have artifacts caused by occlusion or disparity inaccuracy. Observing the complementary behavior between these two modules, we learn a pixel-wise attention...
Fig. 1. Illustration of the proposed framework. The hybrid imaging system [18] captures an HR central view and multiple LR side views. Two sub-networks that are complementary to each other are involved to reconstruct the HR LF image, and the predictions of them are adaptively fused based on learned attention maps. Specifically, the SR-Net spatially super-resolves the input LR side views under the guidance of the HR central view, and the Warp-Net warps the HR central view with the disparity maps estimated from the LR side views. Finally, the predictions produced by these two modules are adaptively fused based on the learned attention maps to generate an HR LF image. The blue frames indicate that the central view of the reconstructed HR LF comes from the input.

This paper follows the overall framework proposed in our previous conference paper [21], namely HybridLF-Net. Yet, HybridLF-Net was merely designed for simulated hybrid data, i.e., the LR side views are generated by down-sampling an HR LF image, and its effectiveness on real hybrid data captured from a typical hybrid imaging system is not explicitly considered. To be specific, the SR-Net of HybridLF-Net explores the LF features using spatial-angular separable (SAS) convolutions and the Warp-Net of HybridLF-Net estimates disparity maps using a plain and shallow convolutional network on the LR LF image. However, there is a significant gap between real and simulated hybrid data, such as the color inconsistency across views, the relatively large disparity, and the inaccurate LF structure among views (i.e., the LR side views and the down-sampled HR central view no longer form an accurately calibrated LF image). Consequently, HybridLF-Net cannot work well on real hybrid data. That is, the SAS-based feature extraction manner in SR-Net and the LF-based disparity estimation are inappropriate, and the accuracy of the disparity maps estimated by Warp-Net is insufficient. See the quantitative and qualitative results in Sec. IV.

Benefiting from the carefully designed training strategy and network architecture, our framework trained with simulated hybrid data can work well on real hybrid data. Extensive experiments on the hybrid data captured by a real imaging system, as well as that simulated from LFs, demonstrate the significant superiority of our method over HybridLF-Net [21], as well as other state-of-the-art ones. That is, our method can reconstruct HR LF images with higher quality and better parallax structure effectively and efficiently.

The rest of this paper is organized as follows. Sec. 2 comprehensively reviews existing methods for image super-resolution. Sec. 3 presents the proposed method. Sec. 4 demonstrates the advantages of the proposed method through extensive experiments on both real and simulated hybrid data. Finally, Sec. 5 concludes this paper.

II. RELATED WORK

A. Single Image Super-resolution

Single image super-resolution (SISR) is a classical problem in the field of image processing. To solve this ill-posed inverse problem, a considerable number of regularization-based and example-based methods [22]–[26] have been proposed. Witnessing the great representation ability of deep learning [27], Dong et al. [28], [29] pioneered deep learning-based methods for SISR, which learn the mapping from LR to HR images in a data-driven manner. Later, deeper network architectures equipped with enhanced feature extraction techniques such
as residual and dense connections were widely exploited to improve the SR performance [30]–[34]. Various loss functions were also proposed to encourage more visually pleasing results, e.g., the perceptual loss [35] and the adversarial loss [36]. More recently, the attention mechanism incorporating non-local operations were introduced to enhance the feature representation and further improve the SR performance [37], [38]. We refer the readers to [39], [40] for a comprehensive survey on SISR.

B. Reference-based Image Super-resolution

Reference-based super-resolution (RefSR) utilizes rich and accurate details from a reference image to assist the SR process. Benefiting from the extra information provided by the reference image, RefSR can achieve significantly superior performance to SISR. Zheng et al. [41], [42] proposed to align the feature maps from the reference image to the target LR image via estimating an optical flow. This method requires a high similarity between the reference and LR images, e.g., different views of the same scene in an LF image. Different from such a global alignment, the idea of local texture matching and transfer was proposed to handle more generic scenarios, where the reference image shares less similar content with the LR image or the correspondences between them have a long distance. Zhang et al. [43] proposed to search for the matching patches from the reference image in the feature space and then swap the matched features to represent the LR image. Xie et al. [44] improved this framework by enhancing the feature extractor. Yang et al. [45] applied the attention mechanism to transfer and fuse HR features from the reference image into LR features based on their relevance embedding. Shim et al. [46] utilized stacked deformable convolutional layers equipped with a multi-scale structure and non-local blocks to match similar content between the LR and reference features.

These RefSR methods can be directly applied to reconstruct an HR LF image from a hybrid input by super-resolving each LR view individually. However, it is difficult to preserve the LF structure as the consistency between the reconstructed views is not considered.

C. LF Image Super-resolution

Different from SISR, LF image super-resolution aims at simultaneously increasing the spatial resolution of all sub-aperture images (SAIs) in an LF image. On top of the target to recover high-frequency details for each SAI, LF super-resolution should also maintain the LF parallax structure. To characterize the relation between SAIs, many methods define a physical model to reconstruct the observed LR SAIs using the desired HR ones. Afterwards, the inverse problem is solved by different priors [9], [10], [47], [48]. These methods always require accurate disparity estimation, which is challenging.

Recent years have witnessed progress on learning-based methods for LF super-resolution. Farrugia et al. [49] constructed a training set by 3D-stacks of 2-D-patches cropped from different SAIs of paired LF images, and then learned a linear mapping between the subspace of the LR and HR patch stacks. Yoon et al. [11] is the first to apply convolutional neural network (CNN) on LF images. However each SAI of an LF image is processed independently in their network, which neglects the angular relationship. Therefore, Yuan et al. [50] proposed to refine the result after separately applying an SISR approach on each SAI. For the same purpose of keeping the geometric consistency in the reconstructed LF image, Wang et al. [51] adopted a recurrent neural network to learn the relations between adjacent SAIs along horizontal and vertical directions. To take advantage of the complementary information between SAIs introduced by the LF structure and address the high-dimensionality challenging, Yeung et al. [52] proposed to use
4-D convolution and more efficient spatial-angular separable convolution (SAS-conv) on LF images. More recently, Wang et al. [53] proposed the spatial-angular interaction module to repetitively incorporate spatial and angular information. Jin et al. [54] proposed an All-to-One module to fuse the combinatorial geometry embedding between the target and auxiliary views in the LF image.

D. LF Image Super-resolution with a Hybrid Input

LF hybrid imaging system was first proposed by Lu et al. [55], in which an HR RGB camera is co-located with a Shack-Hartmann sensor. Boominathan et al. [17] proposed a patch-based method named PaSR to improve the resolution with the hybrid input. Based on PaSR, Wang et al. [18] improved the performance by iterating between patch-based super-resolution and depth-based synthesis, where the synthesized images were used to update the patch dictionary. The patch-based approaches avoid the need to calibrate and register the DSLR camera and the LF camera. However, the average aggregation approaches avoid the need to calibrate and register the DSLR and depth-based synthesis, where the synthesized images of deep CNNs, we investigate a deep neural network that can well capture the characteristics of the input to learn such a mapping function \( f \).

As shown in Fig. 1 our framework consists of two sub-networks, namely SR-Net and Warp-Net. To be specific, by learning deep representations from both \( S^l \) and \( I^h_{u_0} \), the SR-Net aims to super-resolve \( S^l \) via fusing the high-frequency information from \( I^h_{u_0} \) (i.e., to equally increase the spatial resolution of all views contained in \( S^l \)), leading to an intermediate HR LF image and its corresponding attention map, while the Warp-Net inversely warps \( I^h_{u_0} \) to side views with the disparity estimated from \( S^l \), generating another intermediate HR LF image as well as its attention map. Finally, the two intermediate predictions are adaptively fused based on the learned attention maps such that only their respective advantages can be leveraged into a better output. Note that our framework is trained end-to-end. In what follows, we will introduce the details of the proposed framework as well as comprehensive analyses.

III. PROPOSED FRAMEWORK

Notation. Let \( \mathcal{L} = \{I_u \in \mathbb{R}^{H \times W} | u \subset \mathcal{U}\} \) denote an LF image with \( M \times N \) views of resolution \( H \times W \), \( \mathcal{U} \) be the set of 2-D angular coordinates of the views, i.e., \( \mathcal{U} = \{(u, v), 1 \leq u \leq M, 1 \leq v \leq N\} \), and \( I_u \) denotes the SAI at \( u \).

A. Overview

As shown in Fig. 1 a typical hybrid LF imaging system captures an HR central view, denoted by \( I^h_{u_0} \in \mathbb{R}^{H \times W \times \alpha} \), surrounded by a set of LR side views, denoted by \( S^l = \mathcal{L} \setminus I_{u_0} = \{I^l_u \in \mathbb{R}^{H \times W} | u \subset \mathcal{U}\} \), where \( u_0 \) denotes the angular coordinate of the central view, \( \mathcal{U} = \mathcal{U} \setminus u_0 \), \( \alpha \) is the up-sampling scale factor, and \( \setminus \) means the subtraction of sets. An HR 4-D LF image to be reconstructed is denoted as \( \tilde{L}^h = \{I^h_u \in \mathbb{R}^{H \times \alpha \times W} | u \subset \mathcal{U}\} \), and the corresponding ground-truth one is denoted as \( L^h = \{I^h_u | u \subset \mathcal{U}\} \). The problem of reconstructing \( \tilde{L}^h \) from the hybrid input can be implicitly formulated as

\[
\tilde{L}^h = f(I^h_{u_0}, S^l).
\]

To reconstruct \( \tilde{L}^h \), the specific properties of the hybrid input \( I^h_{u_0} \) and \( S^l \) have to be fully explored. Specifically, \( I^h_{u_0} \) with high spatial resolution captures high-frequency details of the scene, while \( S^l \) with multiple observations from different perspectives records geometric information. Moreover, the image characteristics of the real hybrid data, e.g., the relatively large disparity and color inconsistency across views, have to be considered. Considering the powerful representation ability of deep CNNs, we investigate a deep neural network that can...
dense connections $f_{sr-f} (\cdot)$ to learn a residual map, denoted as $R_u^{sr}$:
\[
R_u^{sr} = f_{sr-f} \left( \text{CAT} \left( \{ F_u^l, F_u^h \} \right) \right). \tag{4}
\]
Finally, we add the residual map to upsampled LR view by the bicubic interpolation $\text{BIC}(\cdot)$ to reconstruct the HR view, i.e.,
\[
\hat{I}_u^{sr} = R_u^{sr} + \text{BIC} \left( I_u^l \right), \tag{5}
\]
which constructs the intermediate super-resolved LF by SR-Net, i.e., $\hat{L}_u^{sr} = \{ \hat{I}_u^{sr} | u \subset \mathcal{U} \}$.

The SR-Net is trained by minimizing the absolute error between $\hat{L}_u^{sr}$ and the ground-truth HR LF images:
\[
\ell_{\text{sr}} = \sum_u \sum_x \left| I_u^h(x) - \hat{I}_u^{sr}(x) \right|. \tag{6}
\]

Remark. This module relies on the powerful modelling capacity of the deep CNN to super-resolve $S^l$ for an intermediate HR LF image. By combining features extracted from $S^l$ and $I_{u_0}^h$ for the learning of HR residuals, it is expected that the SR-Net can reconstruct the HR LF image as well as possible. However, its output still suffers from blurry effects caused by the $\ell_1$ loss \cite{58,35}, although $I_{u_0}^h$ contains the high-frequency information of the scene. Additionally, convolutional layers may have difficulties to transfer the high-frequency information from $I_{u_0}^h$ to $S^l$, because the local operation may be insufficient to cover the large disparity between them. See the analysis in Sec. IV-B4 and Fig. 6. In other words, the high-frequency information embedded in $I_{u_0}^h$ cannot be very effectively propagated to the output of the SR-Net. To this end, we further develop the following Warp-Net.

C. Warp-Net

As illustrated in Fig. 2 there are two phases involved in this sub-network, i.e., disparity estimation and inverse warping. The Warp-Net first learns an HR disparity map for each view by exploring the unique LF structure of $S^l$ and combining the HR information of $I_{u_0}^h$, and the resulting HR disparity map is further used to inversely warp $I_{u_0}^h$, leading to another intermediate HR LF image as well as its attention map.

1) Disparity estimation: In this phase, we estimate the disparity maps of the LF image by exploring the view relation, i.e., the LF structure embedded in the LR side views. Specifically, under the Lambertian assumption and in the absence of occlusions, such a relation can be expressed as
\[
I_u^h(x) = I_u^l \left( x + d(u' - u) \right), \tag{7}
\]
where $d$ is the disparity of point $I_u^l(x)$. We use a network with the U-Net structure \cite{59}, denoted as $f_{\text{warp-dis}}(\cdot)$, to exploit the view correlations in $S^l$, and the output is upsampled to generate the initial disparity map, denoted as $\mathcal{D}^{\text{init}} = \{ D_u^{\text{init}} | u \subset \mathcal{U} \}$:
\[
\mathcal{D}^{\text{init}} = \text{UP} \left( f_{\text{warp-dis}} (S^l) \right). \tag{8}
\]

$\mathcal{D}^{\text{init}}$ roughly describes the scene geometry, but lacks high-frequency details to warp the HR central view. Therefore, we further refine $\mathcal{D}^{\text{init}}$ by combining the HR information from $I_{u_0}^h$. We utilize sequential convolutional layers to extract features from $\mathcal{D}^{\text{init}}$ and $I_{u_0}^h$, producing $F^d$ and $F^h$, respectively. The extracted geometry and image features are combined via concatenation and then fused using densely-connected convolutional layers denoted as $f_{\text{warp-f}}(\cdot)$ to reconstruct residual maps for $\mathcal{D}^{\text{init}}$ at individual views, i.e.,
\[
R_u^d = f_{\text{warp-f}} \left( \text{CAT} \left( \{ F_u^d, F_u^h \} \right) \right). \tag{9}
\]
Finally, we estimate the HR disparity map denoted as $\mathcal{D}^h = \{ D_u^h | u \subset \mathcal{U} \}$ as
\[
D_u^h = R_u^d + D_u^{\text{init}}. \tag{10}
\]

2) Inverse warping: Based on $\mathcal{D}^h$, another intermediate HR LF image, denoted as $\hat{L}_u^{\text{warp}} = \{ \hat{I}_u^{\text{warp}} | u \subset \mathcal{U} \}$, can be synthesized by inversely warping $I_{u_0}^h$ to each viewpoint. To make this module be end-to-end trainable, we employ the differentiable bicubic interpolation \cite{60} to realize the process of inverse warping:
\[
\hat{I}_u^{\text{warp}} = \text{WARP} \left( I_{u_0}^h, \mathcal{D}^h_u, u - u_0 \right). \tag{11}
\]

To train the Warp-Net, we minimize the absolute error between the synthesized HR LF image $\hat{L}_u^{\text{warp}}$ and its ground-truth, i.e.,
\[
f_{\text{warp}} = \sum_u \sum_x \left| I_u^h(x) - \hat{I}_u^{\text{warp}}(x) \right|. \tag{12}
\]

Moreover, we use an edge-aware smoothness loss \cite{61,62} to regularize the estimated disparity map, i.e.,
\[
f_{\text{smooth}} = \frac{1}{2} \sum_u \sum_x \text{Exp} \left( -\lambda \left| \frac{\partial D_u^h}{\partial x}(x) \right| \right) \left| \frac{\partial D_u^h}{\partial x}(x) \right| + \text{Exp} \left( -\lambda \left| \frac{\partial D_u^h}{\partial y}(x) \right| \right) \left| \frac{\partial D_u^h}{\partial y}(x) \right|, \tag{13}
\]
where the edge weight $\lambda$ is set to 150 according to \cite{62}.

Remark. By reusing pixels from $I_{u_0}^h$, we expect the high-frequency details of the scene that are challenging to predict can be directly transferred from $I_{u_0}^h$ to each view of $\hat{L}_u^{\text{warp}}$. For example, for regions with continuous depths and complicated textures, Warp-Net performs quite well. See the visual results in Figure 6. However, $\hat{L}_u^{\text{warp}}$ inevitably has distortion caused by inaccurate disparity estimations or occlusions. Specifically, it is difficult to obtain accurate disparities without the ground-truth disparities for supervision, especially in challenging regions, such as textureless regions. Such inaccurate disparities will warp pixels of $I_{u_0}^h$ to wrong positions, resulting in distortion. Second, pixels observed in views of $I_{u_0}^h$ but occluded in $I_{u_0}^h$ will be occupied by the occluder after warping, causing error. Interestingly, the SR-Net suffers less from the distortion induced by these two factors. For example, the textureless regions, where the disparities cannot be accurately estimated, correspond to low-frequency contents, which can be relatively easily predicted by the SR-Net. Besides, the powerful regression ability of the SR-Net can predict the occluded pixels to some extent \cite{63}. 
D. Attention-Guided Fusion

As aforementioned, the SR-Net is capable of predicting the overall content of an HR LF image but fails to recover its delicate textures and sharp edges, while the Warp-Net is able to propagate the high-frequency information to all views but suffers from the distortion caused by occlusions and inaccurate disparity estimation. Fortunately, their advantages are complementary to each other. Therefore, we finally reconstruct an HR LF image by adaptively fusing \( \hat{L}_{sr} \) and \( \hat{L}_{warp} \), in which their advantages are leveraged. And such an adaptive fusion process is achieved under the guidance of their own pixel-wise attention maps.

Both attention maps are respectively learned from the features extracted by the SR-Net and Warp-Net. Specifically, we first use an additional layer parallel to the output layer at the last level to generate the attention maps denoted as \( C_{sr} = \{ C_{sr}^u \in \mathbb{R}^{\alpha H \times \alpha W} | u \in \mathcal{U} \} \) and \( C_{warp} = \{ C_{warp}^u \in \mathbb{R}^{\alpha H \times \alpha W} | u \in \mathcal{U} \} \), for the SR-Net and Warp-Net, respectively, and then apply the Softmax normalization across \( C_{sr} \) and \( C_{warp} \), generating \( \tilde{C}_{sr} = \{ \tilde{C}_{sr}^u \in \mathbb{R}^{\alpha H \times \alpha W} | u \in \mathcal{U} \} \) and \( \tilde{C}_{warp} = \{ \tilde{C}_{warp}^u \in \mathbb{R}^{\alpha H \times \alpha W} | u \in \mathcal{U} \} \). The final reconstruction \( \hat{L}_u \) is produced by the weighted sum of \( \tilde{C}_{sr} \) and \( \tilde{C}_{warp} \):

\[
\hat{L}_u = \tilde{I}_{sr}^u \odot \tilde{C}_{sr}^u + \tilde{I}_{warp}^u \odot \tilde{C}_{warp}^u,
\]

where \( \odot \) is the element-wise multiplication operator. Such an adaptive fusion process is trained under the supervision of minimizing the \( \ell_1 \) distance between the final reconstructed HR LF image and the ground truth one:

\[
\ell_{fusion} = \sum_u \sum_x \left| I_u^h(x) - \hat{L}_u^h(x) \right|.
\]

Combining all modules, we train the whole network end-to-end with the following loss function:

\[
\ell = \ell_{fusion} + \ell_{sr} + \ell_{warp} + \gamma \ell_{smooth},
\]

where the weight factor for smoothness loss \( \gamma \) is empirically set to 0.1.

IV. EXPERIMENTS

A. Implementation Details and Data Augmentation

In our network, we set the kernel size of all convolutional layers to \( 3 \times 3 \) except that of the bottleneck layers, whose kernel size is \( 1 \times 1 \), and applied zero-padding to keep the spatial resolution unchanged. During training, we randomly cropped images to patches of spatial resolution \( 128 \times 128 \). The set the batch size to 1 and used Adam optimizer \([4]\) with \( \beta_1 = 0.9 \) and \( \beta_2 = 0.999 \). The learning rate was initialized as \( 1e^{-4} \) and decreased by a factor of 0.5 every 250 epochs.

Due to the limited number of images in current high-quality LF datasets, it is necessary to apply data augmentation to increase the diversity of the training samples. However, commonly used methods for data augmentation, including image rotation and flip, do not work for LF data. Specifically, if we apply these transformations on each SAI separately, the LF structure in Eq. (7) would be destroyed. For example, applying the flip operation along the \( y \) dimension, we have:

\[
\begin{align*}
I_{u,v}(x, W - y) &= I_{u + \Delta u, v + \Delta v}(x + d \Delta u, W - (y + d \Delta v)) \quad (17) \\
I_{u,v}(x, W - y) &= I_{u + \Delta u, v + \Delta v}(x + d \Delta u, (W - y) - d \Delta v),
\end{align*}
\]

where it can be seen that the relation between the flipped view \( I_u \) and \( I_{u + \Delta u} \) disobey Eq. (7). Therefore, we propose a new data augmentation method tailored for LF data, i.e., applying the image geometric augmentation methods on angular and spatial dimensions simultaneously. With our new strategy, taking the flip augmentation along the \( y \) dimension as an example again, we have:

\[
\begin{align*}
I_{u,v,N-v}(x, W - y) &= I_{u + \Delta u, N-v + \Delta v}(x + d \Delta u, W - (y + d \Delta v)) \quad (18) \\
I_{u,v,N-v}(x, W - y) &= I_{u + \Delta u, (N-v) - \Delta v}(x + d \Delta u, (W - y) - d \Delta v),
\end{align*}
\]

where the LF structure described in Eq. (7) still holds in the flipped LF image.

Moreover, in real hybrid data, different views usually have obvious brightness and color inconsistency due to the change of illumination, camera lens, and viewpoints. To increase the robustness of the model to color inconsistency across views, we augmented the training samples by randomly and independently changing the brightness, contrast, saturation, and hue of each side view of the input, while keeping the color of the supervision data unchanged. We will validate the effectiveness of the color augmentation in Sec. IV-B4.

B. Evaluation on Real Hybrid Data

1) Dataset and Training Strategy: To evaluate the proposed framework, we adopted the real hybrid data captured by the hybrid imaging prototype built in [18] (in Fig. 1), which attaches eight low-cost LR side cameras around a central high-quality HR DSLR camera. Each scene image captured by this prototype consists of eight low-quality side views of spatial resolution around \( 900 \times 1482 \), and a high-quality central view of spatial resolution around \( 1729 \times 2846 \). These \( 3 \times 3 \) views are nearly regularly placed on a 2-D plane with the same rotation and translation parameters. To learn a model suitable for such real data, we particularly designed the training strategy. As the ground-truth HR LF images are not available for supervision in the real hybrid dataset, we simulated hybrid data from publicly available LF datasets for training, i.e., we spatially down-sampled off-center views of LF images from the Inria Dense dataset [65] and the HCI benchmark [66], which contain synthetic LF images of spatial resolution \( 512 \times 512 \), angular resolution of \( 9 \times 9 \),
and disparity in the range of $[-4, 4]$. Considering the angular resolution of the real data for testing and the observation that real hybrid data usually have relatively large disparities, we also uniformly sampled $3 \times 3$ SAIIs from $9 \times 9$ SAIIs of LF images, leading 44 simulated hybrid inputs with a disparity range of $[-16, 16]$ for training.

We converted the RGB images to YUV color space, and only used the Y component for training. During testing, to address the problem of color inconsistency across views, we first converted the input hybrid data to YUV color space, and then applied the trained model to reconstruct the luminance and two chrominance channels separately.

2) Comparison with state-of-the-art methods: To demonstrate the advantages of the proposed method, we compared it with state-of-the-art methods, including two traditional stepwise methods for LF SR from hybrid inputs, i.e., PaSR [17] and iPADS [18], a deep learning-based method for LF SR from hybrid inputs, i.e., HybridLF-Net [21], and a deep learning-based RefSR method, i.e., CrossNet [41]. Additionally, based on RDN [33], a state-of-the-art SISR method, we developed a strong baseline network, namely M-RDN-H, to handle a hybrid LF input. Specifically, as shown in Fig. 3, M-RDN-H first extracts features from the stacked LR side views and HR central view separately, and then concatenates these feature maps together. The combined features are fed into the network with a similar structure to RDN to learn the mapping from the LR to HR space. The LR side views are up-sampled using bicubic interpolation to share the same spatial resolution of the real data for testing and the observation that conventional methods process the images in RGB space directly.

Comparison of visual results. Fig. 4 provides visual comparisons of the reconstructed LFs by different methods, where it can be observed that:

- all of PaSR [17], CrossNet [41], M-RDN-H, and HybridLF-Net [21] suffer from serious blurry effects, such as the letters on the notebook cover, the barcode, and the wood texture on the scene/object. To be specific, PaSR [17] searches for 9 nearest neighbors in the feature space for each LR patch, and then reconstructs this patch by weighted averaging the corresponding HR patches. Such an average operation causes the loss of the high-frequency details in the HR patches. CrossNet [41] estimates a flow between the LR and HR input views, which is further used to align the two views in feature space for reconstruction. However, as the flow is predicted between the cross-domain images, i.e., the LR and HR views, without a proper guidance, i.e., the prediction process is only driven by the final reconstruction loss, it is hard to accurately align the HR features to the target view when the disparity increases, resulting in insufficient propagation of the high-frequency details. M-RDN-H is a pure regression-based method, whose limited performance could be caused by the relatively large disparity between input views as the local convolutions have difficulties to explore the long-distance correlations. Besides, as HybridLF-Net [21] was built on simulated hybrid data, it fails to handle the challenges posed by the color inconsistency and large disparity of real hybrid data. Thus, the high-frequency details in the HR view are not effectively propagated to side views.
- iPADS [18] suffers from distortions around depth discontinuous boundaries. As iPADS renders HR side views by warping the HR central view based on an estimated depth map, high-frequency details can be preserved relatively well on plain areas. However, this method inevitably causes distortions due to depth inaccuracy and occlusions, as we analysed in Sec. IIIC and
- our approach produces satisfactory results on both textured areas and occlusion boundaries. Owing to the attention-guided fusion framework, the results by our approach keep the high-frequency details explicitly propagated from the HR central view and the geometric structure around occlusion boundaries simultaneously.

We refer the readers to the associated video demo for more results.

Comparison of the LF parallax structure. The most valuable information of LF data is the LF parallax structure as described in Eq. 7, which implicitly represents the geometry of the scene/object. To evaluate the ability of different methods in preserving the LF parallax structure, we visually compared the depth/disparity maps estimated from the reconstructed HR LF images by different methods using an identical LF depth estimation algorithm [67]. Fig. 5 shows the results, where it can be observed that our approach can produce much better
Fig. 4. Visual comparisons of different methods on real hybrid data. For each algorithm, we provide the zoom-in images of the red and blue blocks. The colored grid on the top-left corner of each image indicates its angular position.
disparity maps. Specifically, the disparity maps from PaSR [17] and iPADS [18] present obviously blurry around object edges. The reason is that patch matching and depth-based warping generally cause blurry or distortion around depth discontinuous boundaries, leading to view inconsistency in these areas. The disparity maps from CrossNet [41], M-RDN-H, and HybridLF-Net [21] keep sharp edges of the objects, but show obvious errors on areas with weak textures and large disparities, especially the background. In contrast, the disparity maps from our approach keep sharper edges and describe more accurate geometry for both foreground objects and backgrounds, demonstrating the stronger ability of our method in preserving the LF parallax structure than other methods.

3) Efficiency: We also compared the computational complexities of different methods by measuring the running time (in second) of the testing phase and the number of parameters of deep learning-based methods. All methods were tested on a desktop with Intel Xeon Silver 4215R CPU @ 3.20GHz, 128 GB RAM and NVIDIA Quadro RTX 8000. As listed in Table I, it can be observed that learning-based methods, i.e., CrossNet, M-RDN-H, HybridLF-Net, and Ours, are much faster than conventional methods, i.e., PaSR and iPADS. Although our approach takes a slightly longer time than CrossNet and M-RDN-H, its model size is much smaller than theirs. Taking the trade-off between the computational complexity and reconstruction quality, we believe our method is the best one.

4) Ablation study: Here, we provided ablation studies to validate the effectiveness of the framework and the training strategy.

Effectiveness of the fusion manner. To investigate the difference between the SR-Net and Warp-Net and their contributions to the final output, and consequently validate the effectiveness of the fusion component, we visually compared the intermediate predictions by SR-Net and Warp-Net, the corresponding attention maps, and the final output. As shown in Fig. 6, it can be seen that for plain areas (highlighted...
Fig. 6. Visual comparisons of intermediate predictions by the SR-Net and Warp-Net. Note that as the sum of the attention maps of the SR-Net and Warp-Net is equal to 1, we only visualized the attention map of the SR-Net. The zoomed-in red frames highlight the advantages of Warp-Net, while the zoomed-in blue frames highlight the advantages of SR-Net.

in red frames), the SR-Net produces seriously blurry results and fails to recover the textured details, while the Warp-Net can accurately propagate the high-frequency textures from the HR input view. The attention maps also show that the Warp-Net has higher weights for the final reconstruction in these areas. For areas with discontinuous depth (highlighted in blue frames), the predictions of the Warp-Net have distortions while those of the SR-Net maintain the content and provide more contributions to the final outputs. Therefore, we can conclude that the SR-Net and Warp-Net present advantages separately in different areas, and the fusion component is indeed able to leverage the advantages of these two modules to produce better final results.

**Effectiveness of the color augmentation.** We compared the reconstruction results of our method trained with (w/) and without (w/o) the color augmentation strategy. As shown in Fig. 7, it can be seen that our method trained w/o color augmentation produces blurry results with color inconsistency in both spatial and angular domains. More specifically, the colors of SR-Net’s results are mainly influenced by the variant color of the individual side view, while the colors of Warp-Net’s results are the same as that of the HR central view. Consequently, the attention-based fusion results show inconsistent colors inside each view and cross different views. In contrast, the results by our method trained w/ color augmentation preserve intra-view high-frequency details and inter-view color consistency, demonstrating the effectiveness of the color augmentation strategy.

C. Evaluation on Simulated Hybrid Data

To have a quantitative understanding of the advantages of our method, here we also conducted evaluation on simulated hybrid data, which can provide ground-truth HR LF images, although there is a significant gap between real and simulated hybrid data, such as the large disparity and color inconsistency cross views.

1) **Datasets and training details:** We generated simulated hybrid data by down-sampling off-center views of an LF image and only retaining the resolution of the central view. In order to evaluate the performance of different methods on LFs with a higher angular resolution, We used the same training dataset as the experiment in Sec. IV-B but with $5 \times 5$ uniformly sampled SAIs to train another two models for $4\times$ and $8\times$ reconstruction, respectively. The color augmentation was not applied during training as the color inconsistency issue does not appear on simulated data. The rest 19 LF images in the
datasets were used for testing. We converted the LF images to 
YUV color space, and only used the Y components for training 
and quantitative evaluation. When generating visual results, 
the U and V components were up-sampled using bicubic 
interpolation.

2) Comparison with state-of-the-art methods: We com-
pared the proposed approach with state-of-the-art methods for 
LF reconstruction from the hybrid input, including PaSR [17], 
CrossNet [41], M-RDN-H, and HybridLF-Net [21]. We also 
provided comparisons with LF SR methods, i.e., SAS-conv 
[52] and M-RDN. Similar to M-RDN-H, we constructed the 
baseline model M-RDN by modifying RDN [33] to adapt 
to LF data, in which all SAI’s of an LF image are stacked 
along the feature channel and then fed into the residual dense 
network for spatial SR. Fig. 3 shows the network architecture 
of M-RDN. Note that all the learning-based methods were 
re-trained with our training dataset for fair comparisons.

Comparison of quantitative results. We used PSNR and 
SSIM to quantitatively measure the quality of the reconstructed 
HR LF images from simulated hybrid data by different meth-
ods, and the corresponding results are listed in Table II, where 
We can observe that:

- the methods with a hybrid input, including PaSR, Cross-
Net, M-RDN-H, HybridLF-Net, and Ours, significantly 
outperform those with only an LR LF input, including 
SAS-conv and M-RDN, which indicates that the extra 
HR view indeed makes contributions by providing more 
high-frequency information about the scene, and the 
five methods for hybrid inputs have the ability to take 
advantage of such valuable information to some extent. 
Also, this observation validates the potential of the hybrid 
LF imaging;
- among methods with a hybrid input, the traditional 
method PaSR is inferior to others, indicating that a 
simple model with a small capacity is not enough to 
model the intricate relations contained in the hybrid 
input, while learning-based methods, including CrossNet, 
M-RDN-H, HybridLF-Net, and Ours, have much larger 
capacities; and
- our approach achieves the highest PSNR/SSIM in average 
both at scales and exceeds the second best methods (i.e., 
HybridLF-Net [21]) by around 0.5 dB at 4× and 1 db at 
8× reconstruction, demonstrating the great advantage of 
our method.

Comparison of visual results. We visually compared dif-
ferent methods for 4× and 8× reconstruction from simulated 
hybrid data in Figs. 8 and 9. These results further demonstrate 
the significant advantages of the proposed approaches over the 
state-of-the-art ones, i.e., our approach can reconstruct sharper 
details and edges, which are closer to the ground-truth 
ones. Particularly, for 8× reconstruction, it is very difficult 
to recover the details without the guidance of an HR view. 
From Fig. 9, it can be seen that the patterns in the results 
of SAS-conv and M-RDN are seriously distorted. In contrast, 
CrossNet, M-RDN-H, HybridLF-Net, and Ours accept less 
influence of the scale increasing and can still produce accept-
able results. Moreover, our algorithm successfully preserves 
the high-frequency details and reconstructs sharper images.

Comparison of the LF parallax structure. Comparing 
the 2-D epipolar plane image (EPI) is a straightforward way 
to evaluate the LF structure qualitatively. In the EPI of an 
LF image, the projections of a single scene point observed in 
different views construct a straight line. Therefore, we present 
EPIs constructed from the predictions of different algorithms 
for comparison. As shown in Figs. 8 and 9, we can observe that 
the EPIs of our algorithm have clearer line texture and more 
accurate slopes, which demonstrates that our network preserves 
the LF structure better than others. Besides, as the ground-
truth HR LF images are available in this scenario, we also 
evaluated the LF parallax structure of the reconstructed HR 
LF images by different methods both qualitatively. Specifically, 
considering that SSIM is a well-known metric to measure 
the structural similarity between images, we computed the 
SSIM values over EPIs. As listed in Table III, the superiority 
of our method is demonstrated again, based on the fact that 
our method produces the highest SSIM values, especially on
the 8× reconstruction, which poses great challenges to other methods in preserving the LF parallax structure.

V. CONCLUSION

We have presented a novel learning-based framework for reconstructing an HR LF image from a hybrid input in an end-to-end fashion. Owing to The elegant and innovative network architecture and the proposed framework, a lightweight CNN, to comprehensively exploit the underlying properties of the input from two complementary and parallel perspectives. Owing to the careful design and the training and data augmentation strategies, our framework trained with simulated hybrid data is able to adapt to real hybrid data by a typical hybrid network with state-the-art techniques, our framework trained with simulated hybrid data.

ACKNOWLEDGMENT

We would like to thank the authors of [18] for sharing the datasets captured by their hybrid LF imaging system.

REFERENCES

[1] M. Levoy and P. Hanrahan, “Light field rendering,” in Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, 1996, pp. 31–42.
Fig. 8. Visual comparisons of different methods on $4 \times$ reconstruction from simulated hybrid data. For each algorithm, we provide the zoom-in images of the red block and EPIs constructed at the green line.

[2] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross, “Scene reconstruction from high spatio-angular resolution light fields,” ACM Transactions on Graphics, vol. 32, no. 4, pp. 73–1, 2013.
[3] J. Fiss, B. Curless, and R. Szeliski, “Refocusing plenoptic images using depth-adaptive splatting,” in IEEE International Conference on Computational Photography (ICCP), 2014, pp. 1–9.
[4] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, “A 4d light-field dataset and cnn architectures for material recognition,” in European Conference on Computer Vision (ECCV), 2016, pp. 121–138.
[5] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, “Saliency detection on light field,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2816–2823.
[6] P. P. Srinivasan, R. Ng, and R. Ramamoorthi, “Light field blind motion deblurring,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2354–2362.
[7] F.-C. Huang, K. Chen, and G. Wetzstein, “The light field stereoscope: immersive computer graphics via factored near-eye light field displays with focus cues,” ACM Transactions on Graphics, vol. 34, no. 4, p. 60, 2015.
[8] J. Yu, “A light-field journey to virtual reality,” IEEE MultiMedia, vol. 24, no. 2, pp. 104–112, 2017.
[9] S. Wanner and B. Goldluecke, “Variational light field analysis for disparity estimation and super-resolution,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 3, pp. 606–619, 2014.
[10] M. Rossi and P. Frossard, “Geometry-consistent light field super-resolution via graph-based regularization,” IEEE Transactions on Image Processing, vol. 27, no. 9, pp. 4207–4218, 2018.
[11] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon, “Learning a deep convolutional network for light-field image super-resolution,” in IEEE International Conference on Computer Vision Workshops (ICCVW), 2015, pp. 24–32.
[12] J. Jin, J. Hou, H. Yuan, and S. Kwong, “Learning light field angular super-resolution via a geometry-aware network,” in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020, pp. 11141–11148.
[13] M. Guo, J. Hou, J. Jin, J. Chen, and L.-P. Chau, “Deep spatial-angular regularization for compressive light field reconstruction over coded apertures,” in European Conference on Computer Vision (ECCV), 2020.
[14] W. F. H. Yeung, J. Hou, J. Chen, Y. Y. Chung, and X. Chen, “Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular cues,” in European Conference on Computer Vision (ECCV), 2018, pp. 137–152.
[15] J. Jin, J. Hou, J. Chen, H. Zeng, S. Kwong, and J. Yu, “Deep coarse-to-fine dense light field reconstruction with flexible sampling and geometry-aware fusion,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2020.
[16] J. Jin, J. Hou, J. Chen, W. F. H. Yeung, and S. Kwong, “Light field spatial super-resolution via cnn guided by a single high-resolution rgb image,” in IEEE International Conference on Digital Signal Processing (DSP), 2018, pp. 1–5.
[17] V. Boominathan, K. Mitra, and A. Veeraraghavan, “Improving resolution and depth-of-field of light field cameras using a hybrid imaging system,” in IEEE International Conference on Computational Photography (ICCP), 2014, pp. 1–10.
[18] Y. Wang, Y. Liu, W. Heidrich, and Q. Dai, “The light field attachment: Turning a dslr into a light field camera using a low budget camera ring,” IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 10, pp. 2357–2364, 2017.
[19] X. Wang, L. Li, and G. Hou, “High-resolution light field reconstruction using a hybrid imaging system,” Applied optics, vol. 55, no. 10, pp. 2580–2593, 2016.
[20] M. Zhao, G. Wu, Y. Li, X. Hao, F. Lu, and Y. Liu, “Cross-scale reference-based light field super-resolution,” IEEE Transactions on Computational Imaging, vol. 4, no. 3, pp. 406–418, 2018.
[21] J. Jin, J. Hou, J. Chen, S. Kwong, and J. Yu, “Light field super-resolution via attention-guided fusion of hybrid lenses,” in ACM International Conference on Multimedia (ACM MM), 2020, p. 193–201.
[22] J.-B. Huang, A. Singh, and N. Ahuja, “Single image super-resolution from transformed self-exemplars,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5197–5206.
[23] H. Chang, D.-Y. Yeung, and Y. Xiong, “Super-resolution through neighbor embedding,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2004, pp. 1–1.
Fig. 9. Visual comparisons of different methods on $8 \times$ reconstruction from simulated hybrid data. For each algorithm, we provide the zoom-in images of the red block and EPIs constructed at the green line.

[24] R. Timofte, V. De Smet, and L. Van Gool, “A+: Adjusted anchored neighborhood regression for fast super-resolution,” in IEEE Asian Conference on Computer Vision (ACCV), 2014, pp. 111–126.

[25] J. Sun, Z. Xu, and H.-Y. Shum, “Image super-resolution using gradient profile prior,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[26] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution via sparse representation,” IEEE Transactions on Image Processing, vol. 19, no. 11, pp. 2861–2873, 2010.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems (NeurIPS), 2012, pp. 1097–1105.

[28] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in European Conference on Computer Vision (ECCV), 2014, pp. 184–199.

[29] ———, “Image super-resolution using deep convolutional networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 295–307, 2016.

[30] J. Kim, J. Kwon Lee, and K. Mu Lee, “Accurate image super-resolution using very deep convolutional networks,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1646–1654.

[31] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1874–1883.

[32] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Deep laplacian pyramid networks for fast and accurate super-resolution,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 624–632.

[33] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2472–2481.

[34] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, “Enhanced deep residual networks for single image super-resolution,” in IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 136–144.

[35] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in European Conference on Computer Vision (ECCV), 2016, pp. 694–711.

[36] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 105–114.

[37] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in European Conference on Computer Vision (ECCV), 2018, pp. 286–301.

[38] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, “Second-order attention network for single image super-resolution,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11 065–11 074.

[39] J. Tian and K.-K. Ma, “A survey on super-resolution imaging,” Signal, Image and Video Processing, vol. 5, no. 3, pp. 329–342, 2011.

[40] Z. Wang, J. Chen, and S. C. Hoi, “Deep learning for image super-resolution: A survey,” arXiv preprint arXiv:1902.06068, 2019.

[41] H. Zheng, M. Ji, H. Wang, Y. Liu, and L. Fang, “Crossnet: An end-to-end reference-based super resolution network using cross-scale warping,” in European Conference on Computer Vision (ECCV), 2018, pp. 87–104.

[42] Y. Tan, H. Zheng, Y. Zhu, X. Yuan, X. Lin, B. David, and F. Lu, “Crossnet++: Cross-scale large-parallax warping for reference-based super-resolution,” IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 01, pp. 1–1, 2020.

[43] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, “Image super-resolution by neural texture transfer,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7982–7991.

[44] Y. Xie, J. Xiao, M. Sun, C. Yao, and K. Huang, “Feature representation matters: End-to-end learning for reference-based image super-resolution,” in European Conference on Computer Vision (ECCV), 2020, pp. 230–245.

[45] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, “Learning texture transformer network for image super-resolution,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5791–5800.
[46] G. Shim, J. Park, and I. S. Kweon, “Robust reference-based super-resolution with similarity-aware deformable convolution,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8425–8434.

[47] K. Mitra and A. Veeraraghavan, “Light field denoising, light field super-resolution and stereo camera based refocussing using a gmm light field patch prior,” in IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012, pp. 22–28.

[48] S. Heber and T. Pock, “Shape from light field meets robust pca,” in European Conference on Computer Vision (ECCV), 2014, pp. 751–767.

[49] R. A. Farrugia, C. Galea, and C. Guillemot, “Super resolution of light field images using linear subspace projection of patch-volumes,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 7, pp. 1058–1071, 2017.

[50] Y. Yuan, Z. Cao, and L. Su, “Light-field image superresolution using a combined deep cnb based on epi,” IEEE Signal Processing Letters, vol. 25, no. 9, pp. 1359–1363, 2018.

[51] 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 Transactions on Image Processing, vol. 27, no. 9, pp. 4274–4286, 2018.

[52] W. F. H. Yeung, J. Hou, X. Chen, J. Chen, Z. Chen, and Y. Y. Chung, “Light field spatial super-resolution using deep efficient spatial-angular separable convolution,” IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2319–2330, 2018.

[53] Y. Wang, L. Wang, J. Yang, W. An, J. Yu, and Y. Gao, “Spatial-angular interaction for light field image super-resolution,” in European Conference on Computer Vision (ECCV), 2020, pp. 290–308.

[54] J. Jin, J. Hou, J. Chen, and S. 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), 2020, pp. 2260–2269.

[55] C.-H. Lu, S. Muenzel, and J. W. Fleischer, “High-resolution light-field microscopy,” in Imaging and Applied Optics, 2013.

[56] 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 Transactions on Graphics, vol. 36, no. 4, p. 133, 2017.

[57] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.

[58] M. Mathieu, C. Couprie, and Y. LeCun, “Deep multi-scale video prediction beyond mean square error,” International Conference on Learning Representations (ICLR), pp. 1–11, 2016.

[59] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241.

[60] M. Jaderberg, K. Simonyan, A. Zisserman et al., “Spatial transformer networks,” in Advances in Neural Information Processing Systems (NeurIPS), 2015, pp. 2017–2025.

[61] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” in IEEE International Conference on Computer Vision (ICCV), 1998, pp. 839–846.

[62] R. Jonschkowski, A. Stone, J. T. Barron, A. Gordon, K. Konolige, and A. Angelova, “What matters in unsupervised optical flow,” ECCV, pp. 1–1, 2020.

[63] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, “Learning-based view synthesis for light field cameras,” ACM Transactions on Graphics, vol. 35, no. 6, p. 193, 2016.

[64] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[65] J. Shi, X. Jiang, and C. Guillemot, “A framework for learning depth from a flexible subset of dense and sparse light field views,” IEEE Transactions on Image Processing, vol. 28, no. 12, pp. 5867–5880, 2019.

[66] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, “A dataset and evaluation methodology for depth estimation on 4d light fields,” in Asian Conference on Computer Vision (ACCV), 2016, pp. 19–34.

[67] J. Chen, J. Hou, Y. Ni, and L.-P. Chau, “Accurate light field depth estimation with superpixel regularization over partially occluded regions,” IEEE Transactions on Image Processing, vol. 27, no. 10, pp. 4889–4900, 2018.