Light Field Reconstruction via Deep Adaptive Fusion of Hybrid Lenses

Jing Jin, Mantang Guo, Junhui Hou, Senior Member, IEEE, Hui Liu, 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 confidence 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, fusion, depth.

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], densely-sampled LF reconstruction [6], and motion deblurring [7]. Recent research also demonstrates that LF is a promising media for virtual/augment reality [8], [9].

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 [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21]. 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 [19], [21], 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 [18], [19], [20], [21], 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 Fig. 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 confidence-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 concerning 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 confidence map for the output of each module. And the final HR LF image is obtained by adaptively fusing the two intermediate predictions based on their confidence maps, in which only their advantages are collected.

This paper follows the overall framework proposed in our previous conference paper [22], 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 exploits 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 Section IV.

Being aware of these challenges, we carefully redesign both the network architecture and the training strategy to promote the effectiveness of the framework on real hybrid data. Particularly, we make the following efforts:

1) we remove the central view from the LR views to avoid the influence of different central view characteristics between training and testing, and accordingly we modify the SR-Net and Warp-Net to adapt them to the stack of LR side views instead of the LF image for learning the LF representations and explicit geometry;

2) in the Warp-Net, we use a multi-scale structure to explore the long-distance correlations among views;

3) we further enhance the utilization of the high-frequency information of the HR central view from two perspectives, i.e., in the SR-Net, the HR features are fused with the features of each side view more sufficiently, and in the Warp-Net, the HR details of the central view are utilized to enhance the estimated disparity maps estimated from the LR views; and

4) we propose a training strategy tailored to real data, i.e., LF images with large disparities and augmented with color perturbation are used to construct the training dataset.

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 [22], 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. Section II comprehensively reviews existing methods for image super-resolution. Section III presents the proposed method. Section IV demonstrates the advantages of the proposed method through
extensive experiments on both real and simulated hybrid data. Finally, Section V 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 [23], [24], [25], [26], [27] have been proposed. Witnessing the great representation ability of deep learning [28], Dong et al. [29], [30] 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 [31], [32], [33], [34], [35]. Various loss functions were also proposed to encourage more visually pleasing results, e.g., the perceptual loss [36] and the adversarial loss [37]. More recently, the attention mechanism incorporating non-local operations were introduced to enhance the feature representation and further improve the SR performance [38], [39]. We refer the readers to [40], [41] 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. [42], [43] 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. [44] 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. [45] improved this framework by enhancing the feature extractor. Yang et al. [46] 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. [47] 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. Shao et al. [48] predicted the homography matrix between the cross-resolution image pair in a progressively multi-scale manner, with each scale-level learning a sub-homography from a local attention map by Transformer. Zhou et al. [49] first constructed an SR multi-plane image (MPI) at the LR view and then generated the final SR image by fusing the coarse SR result synthesized from the MPI and the bicubic-upsampled LR image via a CNN.

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 [10], [11], [50], [51]. 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. [52] 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. [12] 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. [53] 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. [54] 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 challenge, Yeung et al. [55] proposed to use 4-D convolution and more efficient spatial-angular separable convolution (SAS-conv) on LF images. More recently, Wang et al. [56] proposed the spatial-angular interaction module to repetitively incorporate spatial and angular information. Jin et al. [57] 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. [58], in which an HR RGB camera is co-located with a Stack-Hartmann sensor. Boominathan et al. [18] proposed a patch-based method named PaSR to improve the resolution with the hybrid input. Based on PaSR, Wang et al. [19] 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 causes blurring. Zhao et al. [21] proposed a method named HCSR to separate the high-frequency details from the HR image and warp them to all SAIs to reconstruct an HR LF image. Besides spatial

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
super-resolution, the hybrid LF imaging system was also used to generate LF videos [59].

III. PROPOSED FRAMEWORK

Notation: Let $L = \{I_u \in \mathbb{R}^{H \times W} | u \subseteq \mathcal{U}\}$ 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., $\mathcal{U} = \{u | \mathbf{u} = (u, v), 1 \leq u \leq M, 1 \leq v \leq N\}$, and $I_u$ denotes the SAI at $\mathbf{u}$.

A. Overview

As shown in Fig. 1, a typical hybrid LF imaging system captures an HR central view, denoted by $I_{u_0}^h \in \mathbb{R}^{\alpha H \times \alpha W}$, surrounded by a set of LR side views, denoted by $S^l = L \setminus I_{u_0}^h = \{I_u^l \in \mathbb{R}^{H \times W} | u \subseteq \mathcal{U}\}$, 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 $\hat{L}^h = \{I_u^h \in \mathbb{R}^{\alpha H \times \alpha W} | u \subseteq \mathcal{U}\}$, and the corresponding ground-truth one is denoted as $\hat{L}^l = \{I_u^l | u \subseteq \mathcal{U}\}$. The problem of reconstructing $\hat{L}^h$ from the hybrid input can be implicitly formulated as

$$\hat{L}^h = f(I_{u_0}^h, S^l).$$

(1)

To reconstruct $\hat{L}^h$, the specific properties of the hybrid input $I_{u_0}^h$ and $S^l$ have to be fully explored. Specifically, $I_{u_0}^h$ 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 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 $I_{u_0}^h$, the SR-Net aims to super-resolve $S^l$ via fusing the high-frequency information from $I_{u_0}^h$, 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 confidence map, while the Warp-Net inversely warps $I_{u_0}^h$ to side views with the disparity estimated from $S^l$, generating another intermediate HR LF image as well as its confidence map. Finally, the two intermediate predictions are adaptively fused based on the learned confidence 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.

B. SR-Net

As depicted in Fig. 2, the SR-Net comprises three modules, i.e., LF feature extraction, HR feature extraction, and hybrid feature fusion. The three modules are connected to promote sufficient exploration of the information contained in the hybrid input.

1) LF Feature Extraction: Multiple observations from different perspectives contained in $S^l$ provide supplementary information, available online, of the scene (i.e., details absent at a certain view may be present in another one), which will be beneficial to the reconstruction quality. To capture such information, we stack the LR side views $\{I_u^l\}$ along the feature channel and utilize sequential convolutional layers with dense connections [34], [60] to extract LR features. Note that we suppress the feature channels of each group of densely-connected features using a bottleneck layer to reduce the number of parameters. The sub-pixel convolutional layer [32] is applied to up-sample the

---

**Fig. 2.** Illustration of the network architectures of the SR-Net and the Warp-Net. The SR-Net super-resolves the LR side views by fusing the high-frequency information from the HR central view, while the Warp-Net synthesizes an HR LF image by warping the HR central view based on the disparity map estimated from the LR side views. We refer readers to the supplementary file, available online, for the detailed architecture.
extracted features to the desired spatial resolution, i.e.,

\[ \mathcal{F}^l = \text{UP} \left( f_{sr-l} \left( S^l \right) \right), \tag{2} \]

where \( \mathcal{F}^l = \{ F_u^l | u \subset \mathcal{U} \} \) is the set of LF features for LR side views, \( \text{UP}(\cdot) \) is the up-sampling layer, and \( f_{sr-l}(\cdot) \) is the feature extraction layers.

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

\[ I_u^l(x) = I_u^l(x + d (u' - u)), \tag{7} \]

where \( d \) is the disparity of point \( I_u^l(x) \). We use a network with the U-Net structure [62], denoted as \( f_{\text{warp-di}}(\cdot) \), to exploit the view correlations in \( S^l \), and the output is upsampled to generate the initial disparity map, denoted as \( D^{\text{init}} \).

\[ D^{\text{init}} = \text{UP} \left( f_{\text{warp-di}} \left( S^l \right) \right). \tag{8} \]

\( D^{\text{init}} \) roughly describes the scene geometry but lacks high-frequency details to warp the HR central view. Therefore, we further refine \( D^{\text{init}} \) by combining the HR information from \( I_u^h \). We utilize sequential convolutional layers to extract features from \( D^{\text{init}} \) and \( I_u^h \), 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 \( D^{\text{init}} \) at individual views, i.e.,

\[ R^d_u = f_{\text{warp-f}} \left( \text{CAT} \left( F^d_u, F^h_u \right) \right). \tag{9} \]

Finally, we estimate the HR disparity map denoted as \( D^h = \{ D_u^h | u \subset \mathcal{U} \} \) as

\[ D_u^h = R^d_u + D^{\text{init}}. \tag{10} \]

2) **Inverse Warping:** Based on \( D^h \), another intermediate HR LF image, denoted as \( \hat{L}^u \), can be synthesized by inversely warping \( I_u^h \) to each viewpoint. To make this module end-to-end trainable, we employ the differentiable bicubic interpolation [63] to realize the process of inverse warping

\[ \hat{I}_u^\text{warp} = \text{WARP} \left( I_u^h, I_u^h, u - u_0 \right). \tag{11} \]

To train the Warp-Net, we minimize the absolute error between the synthesized HR LF image \( \hat{L}^u \) and its ground-truth, i.e.,

\[ \ell_{\text{warp}} = \sum_u \sum_x \left| I_u^h(x) - \hat{I}_u^\text{warp}(x) \right|. \tag{12} \]

Moreover, we use an edge-aware smoothness loss [64, 65] to regularize the estimated disparity map, i.e.,

\[ \ell_{\text{smooth}} = \frac{1}{2} \sum_u \sum_x \text{Exp} \left( -\lambda \left| \frac{\partial D_u^h}{\partial x}(x) \right| \right) \left| \frac{\partial D_u^h}{\partial x}(x) \right| \]

\[ + \text{Exp} \left( -\lambda \left| \frac{\partial D_u^h}{\partial y}(x) \right| \right) \left| \frac{\partial D_u^h}{\partial y}(x) \right|, \tag{13} \]

where the edge weight \( \lambda \) is set to 150 according to [65].

**Remark:** By reusing pixels from \( I_u^h \), we expect the high-frequency details of the scene that are challenging to predict can be directly transferred from \( I_u^h \) to each view of \( \hat{L}^u \). For example, for regions with continuous depths and complicated textures, Warp-Net performs quite well. See the visual results.
in Fig. 7. However, \( \tilde{I}_{\text{warp}} \) 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 \( I_u \) to wrong positions, resulting in distortion. Second, pixels observed in views of \( I_u \) but occluded in \( I_v \) will be occupied by the occluder after warping, causing errors. 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 [66].

D. Confidence-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 \( \tilde{L}^{sr} \) and \( \tilde{L}^{warp} \), in which their advantages are leveraged. And such an adaptive fusion process is achieved under the guidance of their own pixel-wise confidence maps.

Both confidence maps are 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 confidence maps denoted as \( C^{sr} = \{ C^{sr} \in \mathbb{R}^{H \times W} | u \subset \mathbb{U} \} \) and \( C^{warp} = \{ C^{warp} \in \mathbb{R}^{H \times W} | u \subset \mathbb{U} \} \) for the SR-Net and Warp-Net, respectively, and then apply the Softmax normalization across \( C^{sr} \) and \( C^{warp} \), generating \( \tilde{C}^{sr} = \{ \tilde{C}^{sr} \in \mathbb{R}^{H \times W} | u \subset \mathbb{U} \} \) and \( \tilde{C}^{warp} = \{ \tilde{C}^{warp} \in \mathbb{R}^{H \times W} | u \subset \mathbb{U} \} \). The final reconstruction \( \hat{I}_u^h \) is produced by the weighted sum of \( \tilde{L}^{sr} \) and \( \tilde{L}^{warp} \)

\[
\hat{I}_u^h = \tilde{I}^{sr}_u \odot \tilde{C}^{sr}_u + \tilde{I}^{warp}_u \odot \tilde{C}^{warp}_u, \tag{14}
\]

where \( \odot \) is the element-wise multiplication operator. Such an adaptive fusion process is trained under the supervision of minimizing the \( \ell_2 \) distance between the final reconstructed HR LF image and the ground truth one

\[
\ell_{\text{fusion}} = \sum_u \sum_x \left| I_u^h(x) - \hat{I}_u^h(x) \right|. \tag{15}
\]

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

\[
\ell = \ell_{\text{fusion}} + \ell^{sr} + \ell^{warp} + \gamma \ell_{\text{smooth}}, \tag{16}
\]

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. We set the batch size to 1 and used Adam optimizer [67] 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 (7) would be destroyed. For example, applying the flip operation along the \( y \) dimension, we have

\[
I_{u,v}(x, W - y) = I_{u+\Delta u,v+\Delta v}(x + d\Delta u, W - (y + d\Delta v)) = I_{u+\Delta u,v+\Delta v}(x + d\Delta u, (W - y) - d\Delta v), \tag{17}
\]

where it can be seen that the relation between the flipped view \( I_{u,v} \) and \( I_{u+\Delta u,v} \) disobeys (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

\[
I_{u,N-v}(x, W - y) = I_{u+\Delta u,N-(v+\Delta v)}(x + d\Delta u, W - (y + d\Delta v)) = I_{u+\Delta u,N-(v+\Delta v)}(x + d\Delta u, (W - y) - d\Delta v), \tag{18}
\]

where the LF structure described in (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 examples 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 Section 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 LF imaging prototype built in [19], as shown in Fig. 3, 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 $1729 \times 2846$. The intrinsic parameters of the cameras and the extrinsic parameters with reference to the central camera were calibrated using a standard camera calibration toolbox and bundle adjustment software, respectively. The side-view images were rectified to share the same rotation with the central image by projection and homography warping. Finally, the $3 \times 3$ views are nearly regularly placed on a 2-D plane with the same rotation. We also refer readers to [19] for more details about the settings of the prototype.

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 [68] and the HCI benchmark [69], 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 $3 \times 3$ SAI s from $9 \times 9$ SAI s 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 [18] and iPADS [19], a deep learning-based method for LF SR from hybrid inputs, i.e., HybridLF-Net [22], and a deep learning-based RefSR method, i.e., CrossNet [42]. Additionally, based on RDN [34], 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. 4, 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 HR view before being fed into the network. Note that we re-trained all the learning-based methods under comparison using the same training dataset and strategy as Ours for fair comparisons. The YUV channels are also processed in the same way for learning-based methods, while conventional methods process the images in RGB space directly.

Comparison of Visual Results: Fig. 5 provides visual comparisons of the reconstructed LF s by different methods, where it can be observed that

- all of PaSR [18], CrossNet [42], M-RDN-H, and HybridLF-Net [22] suffer from serious blurry effects, such as the letters on the notebook cover, the barcode, and the wood texture on the wall. To be specific, PaSR [18] 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 [42] 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 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 exploring the long-distance correlations. Besides, as HybridLF-Net [22] 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 [19] 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 analyzed in Section III-C; and

- our approach produces satisfactory results on both textured areas and occlusion boundaries. Owing to the confidence-guided fusion framework, the results of 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$^1$ for more results.

$^1$https://github.com/jingjin25/LFHybridSR-Fusion
Comparison of the LF Parallax Structure: The most valuable information of LF data is the LF parallax structure as described in (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 [70]. Fig. 6 shows the results, where it can be observed that our approach can produce much better disparity maps. Specifically, the disparity maps from PaSR [18], and iPADS [19] 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 [42], M-RDN-H, and HybridLF-Net [22] 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 to preserve the LF parallax structure than other methods.

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. 8, 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 confidence-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 evaluations 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 Section 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 compared the proposed approach with state-of-the-art methods for LF reconstruction from the hybrid input, including PaSR [18], CrossNet [42], M-RDN-H, and HybridLF-Net [22]. We also provided comparisons with LF SR methods, i.e., SAS-conv [55] and M-RDN. Similar to M-RDN-H, we constructed the baseline model M-RDN by modifying RDN [34] to adapt to LF data, in which all SAsIs of an LF image are stacked along the feature channel and then fed into the residual dense network for spatial SR. Fig. 4 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 methods, and the corresponding results are listed in Table II, where we can observe that:

- the methods with a hybrid input, including PaSR, CrossNet, 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 at both scales and exceeds the second best methods (i.e.,
Fig. 7. Visual comparisons of intermediate predictions by the SR-Net and Warp-Net. Note that as the sum of the confidence maps of the SR-Net and Warp-Net is equal to 1, we only visualized the confidence 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.

Fig. 8. Visual comparisons of the reconstruction results by our method trained w/ and w/o color augmentation. We also provided the HR central view and LR views for reference. See the associated video demo for more results.

HybridLF-Net [22]) 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 different methods for 4× and 8× reconstruction from simulated hybrid data in Figs. 9 and 10. 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 edges and clearer scenes, which are closer to the ground-truth ones. Particularly, for 8× reconstruction, it is very difficult to recover
**Table II**

| Scale | LF | Bicubic | SAS-conv [55] | M-RDN | PSr [14] | CroNet [42] | M-RDN-H | Hybrid-F.Net [21] | Ours |
|-------|-----|---------|--------------|--------|---------|-------------|---------|-----------------|------|
| Room  | 30.5/9.8/0.493 | 33.87/9.047/0.238 | 36.1/9.4/0.229 | 34.51/8.958/0.229 | 37.83/9.77/0.050 | 39.67/9.96/0.039 | 39.97/9.97/0.039 | 39.98/9.98/0.039 |
| Bed   | 30.5/9.8/0.493 | 33.87/9.047/0.238 | 36.1/9.4/0.229 | 34.51/8.958/0.229 | 37.83/9.77/0.050 | 39.67/9.96/0.039 | 39.97/9.97/0.039 | 39.98/9.98/0.039 |
| Sideboard | 24.3/7.4/0.454 | 27.95/8.76/0.189 | 27.52/8.61/0.177 | 26.74/8.79/0.173 | 29.50/9.30/0.074 | 34.18/9.73/0.035 | 33.78/9.79/0.036 | 33.92/9.77/0.035 |
| Town  | 28.1/6.8/0.417 | 32.12/9.34/0.200 | 31.73/9.28/0.179 | 31.45/8.70/0.070 | 36.58/9.79/0.035 | 39.3/9.99/0.016 | 40.49/9.99/0.017 | 40.52/9.99/0.017 |
| Antiques | 34.6/8.4/0.312 | 36.10/8.93/0.155 | 37.68/9.42/0.102 | 39.70/9.42/0.122 | 44.10/9.93/0.095 | 47.01/9.29/0.077 | 47.50/9.97/0.018 | 48.07/9.97/0.015 |
| Cameras_brush | 28.7/8.0/0.374 | 33.44/8.61/0.132 | 32.77/9.55/0.156 | 34.51/9.54/0.154 | 37.13/9.73/0.048 | 38.36/9.94/0.024 | 39.07/9.98/0.025 | 39.19/9.98/0.025 |
| Chair  | 28.1/6.8/0.417 | 32.12/9.34/0.200 | 31.73/9.28/0.179 | 31.45/8.70/0.070 | 36.58/9.79/0.035 | 39.3/9.99/0.016 | 40.49/9.99/0.017 | 40.52/9.99/0.017 |
| Coffee_time | 21.2/6.0/0.372 | 27.82/6.89/0.257 | 25.53/8.31/0.201 | 26.74/8.82/0.254 | 31.73/9.27/0.057 | 37.46/9.92/0.035 | 38.46/9.92/0.035 | 38.56/9.91/0.035 |
| Flowers_clock | 30.9/8.0/0.360 | 33.66/8.90/0.055 | 35.22/9.74/0.080 | 34.95/9.55/0.080 | 39.27/9.90/0.032 | 41.59/9.99/0.021 | 41.97/9.99/0.021 | 42.37/9.99/0.020 |
| Lamps | 33.0/9.2/0.248 | 29.23/9.46/0.111 | 32.95/9.68/0.114 | 33.88/9.64/0.108 | 36.28/9.92/0.024 | 37.76/9.74/0.024 | 37.99/9.72/0.023 | 38.63/9.69/0.021 |
| Microphone_flower | 23.8/9.3/0.375 | 27.25/9.16/0.103 | 26.47/9.05/0.143 | 29.36/9.52/0.112 | 31.20/9.72/0.040 | 38.50/9.94/0.014 | 38.56/9.95/0.017 | 38.62/9.98/0.013 |
| Pencil_case | 29.9/8.5/0.365 | 32.91/9.13/0.244 | 31.23/9.02/0.224 | 31.43/8.75/0.097 | 36.82/9.63/0.056 | 39.13/9.90/0.028 | 39.08/9.75/0.045 | 39.13/9.90/0.028 |
| Raster_clock | 28.7/8.5/0.365 | 32.91/9.13/0.244 | 31.23/9.02/0.224 | 31.43/8.75/0.097 | 36.82/9.63/0.056 | 39.13/9.90/0.028 | 39.08/9.75/0.045 | 39.13/9.90/0.028 |
| Roses_table | 30.9/8.0/0.360 | 33.66/8.90/0.055 | 35.22/9.74/0.080 | 34.95/9.55/0.080 | 39.27/9.90/0.032 | 41.59/9.99/0.021 | 41.97/9.99/0.021 | 42.37/9.99/0.020 |
| Toy_friends | 30.5/9.8/0.493 | 33.87/9.047/0.238 | 36.1/9.4/0.229 | 34.51/8.958/0.229 | 37.83/9.77/0.050 | 39.67/9.96/0.039 | 39.97/9.97/0.039 | 39.98/9.98/0.039 |
| Toys | 29.8/8.0/0.365 | 32.91/9.13/0.244 | 31.23/9.02/0.224 | 31.43/8.75/0.097 | 36.82/9.63/0.056 | 39.13/9.90/0.028 | 39.08/9.75/0.045 | 39.13/9.90/0.028 |
| Whales | 30.3/9.1/0.307 | 33.45/8.91/0.091 | 34.54/9.07/0.082 | 36.58/9.54/0.089 | 40.24/9.89/0.007 | 42.30/9.92/0.017 | 42.60/9.93/0.019 | 43.20/9.90/0.019 |

**Fig. 9.** Visual comparisons of different methods on 4x 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.
the details without the guidance of an HR view. From Fig. 10, 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 acceptable 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. 9 and 10, 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 $8 \times$ reconstruction, which poses great challenges to other methods in preserving the LF parallax structure.

We further compared the depth maps estimated from the reconstructed HR LF images by different methods on the simulated data quantitatively. Note that the ground-truth depth map of Bedroom is not available, and thus we excluded it for comparison. We still adopted the LF depth estimation algorithm in [70] and used the mean squares error (MSE) between the estimated depth map and its ground-truth to measure the accuracy. Table IV lists the quantitative results of each LF image, where it can be seen that the depth maps from the reconstructed LFs by our method achieve the highest and second-highest accuracy on the majority of scenes under the $4 \times$ and $8 \times$ tasks. Besides, the MSE values of our method and other compared methods are even lower than those of GT in some scenes. The reason is that LF depth estimation itself is an open problem, and no method can guarantee perfect estimation, potentially resulting in errors.

---

**Table III**

| Scale | Bicubic | SAS-conv [53] | M-RDN | PaSR [18] | CrossNet [42] | M-RDN-H | HybridLF-Net [22] | Ours |
|-------|---------|---------------|-------|-----------|---------------|---------|-------------------|------|
| $4 \times$ | 0.819 | 0.900 | 0.891 | 0.916 | 0.957 | 0.973 | 0.974 | 0.976 |
| $8 \times$ | 0.724 | 0.805 | 0.791 | 0.882 | 0.929 | 0.948 | 0.955 | 0.963 |

The best and second-best results are colored in red and blue, respectively.
due to the limitation of the estimation method itself. Although the errors introduced by the LF depth estimation method are inevitable, our method achieves top-2 accuracy on the majority of scenes, which is sufficient to demonstrate the advantage of our method.

3) Ablation Study: We conducted experiments to demonstrate that our method is able to leverage multiple LR images and benefit the reconstruction quality of each side view. Specifically, we trained two SR models for hybrid inputs with $3 \times 3$ and $5 \times 5$ LR side views, respectively, both containing a central HR image as a reference. Then we took the super-resolved results of the eight views around the HR central view and compared their average quantitative results of the two SR models in Table V, where we can see that the same eight views can be reconstructed with higher quality when more side views are provided, demonstrating that our model can effectively leverage the complementary information of multiple views to improve the reconstruction quality.

Besides, to validate the effectiveness of the newly proposed modifications and architectures in SR-Net and Warp-Net, i.e., using a stack of LR side images in SR-Net, and multi-scale structure in Warp-Net. We qualitatively compared the performance of SR-Net and Warp-Net in HybridLF-Net and our method. We denote the results of SR-Net and Warp-Net in HybridLF-Net as HybridLF-Net-S and HybridLF-Net-W, respectively, and denote the results of SR-Net and Warp-Net in our method as Ours-S and Ours-W (multi-scale), respectively. To directly verify the advantage of the multi-scale Warp-Net over the single-scale one, we also set a baseline, named Ours-W (single-scale), by modifying the Warp-Net in our method with a single-scale structure while leaving other modules unchanged. As shown in Fig. 11, it can be observed that Ours-S and Ours-W (multi-scale) can reconstruct sharper edges and clear textures than HybridLF-Net-S and HybridLF-Net-W, respectively, validating the advantage of using a stack of LR side images for SR and multi-scale Warp-Net over those of Hybrid-Net; and Ours-W (multi-scale) can reconstruct sharp edges at the occlusion boundaries that are closer to the ground truth than Ours-W (single-scale), directly validating the advantage of the multi-scale structure.

We also quantitatively compared the performance of the above models over the simulated hybrid dataset. As shown in Table VI, it can be observed that Ours-S (resp. Ours-W) achieves higher average PSNR and SSIM values than HybridLF-Net-S (resp. HybridLF-Net-W), and Ours-W (multi-scale) produces higher average PSNR and SSIM values than Ours-W (single-scale).

4) More Quantitative Analysis: To investigate the performance of our method with respect to very low-resolution side views, we re-trained our model on the $16 \times$ task, where the resolution of input LR side views is $32 \times 32$. Additionally, we also provided the results of the $2 \times$ task to have a comprehensive understanding of the performance of our method. As listed in Table VII, we can see that the PSNR/SSIM values indeed decrease rapidly when the input images have very low resolution but are still within an acceptable (viewable) range. The possible reason is that the low-resolution side views cannot provide accurate geometric information to facilitate the propagation of the high-frequency details from the HR central view.

### Table V

| Scale | Ours-S x 3 | Ours-x 5 | Ours-x 5
|-------|------------|----------|----------|
| 8 x | 38.65/0.983 | 39.51/0.986 |

We refer readers to the supplementary file, available online, for the quantitative result of each test LF.
Besides, to investigate the generalization capacity of our network in terms of the camera layouts (i.e., putting the HR image at different views), we further conducted experiments on data with the HR view set as the top-left corner view, denoted as **Ours-Corner**. As compared in Table VIII, we can see that putting the HR image as the central view, denoted as **Ours-Central**, improves the reconstruction quality significantly. Besides, from Fig. 13, it can be seen that putting the HR image at the central view balances the quality of all views better.

### D. Comparison With NeRF-Based Reconstruction

Recently, the popular view synthesis model NeRF [71] has drawn much attention in the computer vision/graphics community. We also conducted comprehensive experiments to investigate the advantage of our explicit formulation of transferring the information of an HR image to LR ones over the NeRF-based implicit modeling.

Generally, we trained an NeRF model, named NeRF-Hybrid, supervised by the hybrid LF image (i.e., an HR central view surrounded by eight LR side views), to render the LR side views at the same resolution as the HR central view. It is expected that NeRF-Hybrid could implicitly transfer the information of the HR central view to the LR side views during training. However, as demonstrated in a recent work [72], an NeRF trained with LR images usually generates blurring effects when used for rendering images with higher resolution. To promote the NeRF trained with LR images to reconstruct high-quality HR images, [72] proposes a super-sampling strategy, which splits a pixel of the LR image into multiple sub-pixels and draws a ray for each sub-pixel. Super-sampling performs supervision by minimizing the loss between the split pixel and the average of the radiances rendered from its corresponding sub-pixels. Therefore, to generate high-quality HR side views with only the supervision of their LR counterparts, following [72], we supervised the side views with the super-sampling strategy. More specifically, we supervised the side view by calculating the L2 loss between a typical pixel of the side view and the average of a grid of $s \times s$ radiances rendered from its corresponding sub-pixels, where $s$ is the scale factor. We supervised the HR central view by calculating the L2 loss between the input pixel and its rendered radiance. In addition, we trained NeRF models with all views being LR images, named NeRF-LR, as the baseline to verify the

| Scale | Ours-Corner | Ours-Central |
|-------|-------------|--------------|
| 4×    | 38.29/0.982 | **39.91/0.987** |
| 8×    | 35.07/0.966 | **37.34/0.979** |

The PSNR/SSIM values refer to the average of the simulated hybrid data. We refer readers to the supplementary file, available online, for the result of each test LF.
NeRF-Hybrid consistently performs better than NeRF-LR. Our method achieves much higher PSNR than NeRF-LR for sharing the visual information and ability of NeRF-Hybrid in transferring the information of the HR central view to the LR side views.

As the camera parameters of the Inria dataset used in our simulated hybrid LF datasets are not available, we conducted experiments in terms of the $2 \times$, $4 \times$, and $8 \times$ reconstruction tasks only on the HCI dataset including four scenes. The working mechanism of NeRF only allows us to train NeRF-LR and NeRF-Hybrid, for each scene per task. Fig. 12 visually shows the average PSNR at each angular position of reconstructed LFs from NeRF-LR, NeRF-Hybrid, and our method, where it can be observed that

- NeRF-Hybrid consistently performs better than NeRF-LR under all scenarios, indicating that NeRF can implicitly transfer the information of the HR central view to LR side views; and
- our method achieves much higher PSNR than NeRF-Hybrid in all scenarios, validating the advantage of explicitly transferring information from the HR central view to LR side views adopted in our method. Although NeRF-Hybrid regresses a model for each scene separately, it cannot supervise the LR side views with the ground-truth HR counterparts containing detailed information (or high-frequency components), resulting in the trained model having a weak ability to infer this kind of information during rendering HR images, thus limiting performance. This is consistent with the observation of the work [73], i.e., deep networks are biased towards learning lower frequency functions.

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. The elegant and innovative network architecture enables 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 imaging system very well. Extensive experimental results demonstrate that our framework not only reconstructs HR LF images with higher quality and better LF parallax structure, but also runs at a relatively high speed, when compared with state-of-the-art approaches.

ACKNOWLEDGMENTS

We would like to thank the authors of [19] for sharing the datasets captured by their hybrid LF imaging system.

REFERENCES

[1] M. Levoy and P. Hanrahan, “Light field rendering,” in Proc. 23rd Annu. Conf. Comput. Graph. Interactive Techn., 1996, pp. 31–42.
[2] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross, “Scene reconstruction from high spatio-angular resolution light fields,” ACM Trans. Graph., vol. 32, no. 4, 2013, Art. no. 73.
[3] J. Fiss, B. Curless, and R. Szeliski, “Refocusing plenoptic images using depth-adaptive splatting,” in Proc. IEEE Int. Conf. Comput. Photogr., 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 Proc. Eur. Conf. Comput. Vis., 2016, pp. 121–138.
[5] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, “Saliency detection on light field,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 2806–2813.
[6] M. Gao, J. Jin, H. Liu, and J. Hou, “Learning dynamic interpolation for extremely sparse light fields with wide baselines,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 2450–2459.
[7] P. P. Srinivasan, R. Ng, and R. Ramamoorthi, “Light field blind motion deblurring,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2012, pp. 2354–2362.
[8] 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 Trans. Graph., vol. 34, no. 4, 2015, Art. no. 60.
[9] J. Yu, “A light-field journey to virtual reality,” IEEE Multimedia, vol. 24, no. 2, pp. 104–112, Second Quarter 2017.
W. Shi et al., “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1874–1883.

W-S. Lai, J-B. Huang, N. Ahuja, and M-H. Yang, “Deep laplacian pyramid networks for fast and accurate super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 624–632.

Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 2472–2481.

B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, “Enhanced deep residual networks for single image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2017, pp. 136–144.

J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 694–711.

C. Ledig et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 105–114.

Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 286–301.

T. Dai, J. Cai, Y. Zhang, S-T. Xia, and L. Zhang, “Second-order attention network for single image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 11065–11074.

J. Tian and K-K. Ma, “A survey on super-resolution imaging,” Sign. Image Video Process., vol. 5, no. 3, pp. 329–342, 2011.

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

H. Zheng, M. Ji, H. Wang, Y. Liu, and L. Fung, “CrossNet: An end-to-end reference-based super-resolution network using cross-scale warping,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 87–104.

Y. Tan et al., “CrossNet: Cross-scale large-parallax warping for reference-based super-resolution,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 12, pp. 4291–4305, Dec. 2021.

Z. Zhang, Z. Wang, Z. Lin, and H. Qi, “Image super-resolution by neural texture transfer,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 7982–7991.

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 Proc. Eur. Conf. Comput. Vis., 2020, pp. 230–245.

F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, “Learning texture transformer network for image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5791–5800.

G. Shim, J. Park, and I. S. Kweon, “Robust reference-based super-resolution with similarity-aware deformable convolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 8425–8434.

R. Shao, G. Wu, Y. Zhou, Y. Fu, L. Fang, and Y. Liu, “LocalTrans: A multiscale local transformer network for cross-resolution homography estimation,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 14890–14899.

Y. Zhou, G. Wu, Y. Fu, K. Li, and Y. Liu, “Cross-MPI: Cross-scale stereo for image super-resolution using multiframe images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 14842–14851.

K. Mitra and A. Veeraraghavan, “Light field denoising, light field super-resolution and stereo camera based refocusing using a GMM light field patch prior,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2012, pp. 22–28.

S. Heber and T. Pock, “Shape from light field meets robust PCA,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 751–767.

R. A. Farruggia, C. Galea, and C. Guillen, “Super resolution of light field images using linear subspace projection of patch-volumes,” IEEE J. Sel. Topics Signal Process., vol. 11, no. 7, pp. 1058–1071, Oct. 2017.

Y. Yuan, Z. Cao, and L. Su, “Light-field image superresolution using a combined deep CNN based on EPI,” IEEE Signal Process. Lett., vol. 25, no. 9, pp. 1359–1363, Sep. 2018.

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

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 Trans. Image Process., vol. 28, no. 5, pp. 2319–2330, May 2019.

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

J. Jin, J. Hou, J. Chen, and S. Kwong, “Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 2260–2269.
[58] C.-H. Lu, S. Muenzel, and J. W. Fleischer, “High-resolution light-field microscopy,” in Proc. Imag. Appl. Opt., 2013, p. CTh3B.2. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=COSI-2013-CTh3B.2

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

[60] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4700–4708.

[61] M. Mathieu, C. Couprie, and Y. LeCun, “Deep multi-scale video prediction beyond mean square error,” in Proc. Int. Conf. Learn. Representations, 2016, pp. 1–11.

[62] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assit. Intervention., 2015, pp. 234–241.

[63] M. Jaderberg et al., “Spatial transformer networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 2017–2025.

[64] J. Shi, X. Jiang, and C. Guilleminot, “A framework for learning depth from a flexible subset of dense and sparse light field views,” IEEE Trans. Image Process., vol. 28, no. 12, pp. 5867–5880, Dec. 2019.

[65] R. Jonschkowski, A. Stone, J. T. Barron, A. Gordon, K. Konolige, and A. Angelova, “What matters in unsupervised optical flow,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 557–572.

[66] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, “Learning-based view synthesis for light field cameras,” ACM Trans. Graph., vol. 35, no. 6, 2016, Art. no. 193.

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

[68] J. Shi, X. Jiang, and C. Guilleminot, “A framework for learning depth from a flexible subset of dense and sparse light field views,” IEEE Trans. Image Process., vol. 28, no. 12, pp. 5867–5880, Dec. 2019.

[69] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, “A dataset and evaluation methodology for depth estimation on 4D light fields,” in Proc. Asian Conf. Comput. Vis., 2016, pp. 19–34.

[70] J. Chen, J. Hou, Y. Ni, and L.-P. Chau, “Accurate light field depth estimation with superpixel regularization over partially occluded regions,” IEEE Trans. Pattern Recognit. Comput., vol. 27, no. 10, pp. 4889–4900, Oct. 2018.

[71] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “NeRF: Representing scenes as neural radiance fields for view synthesis,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 405–421.

[72] C. Wang, X. Wu, Y.-C. Guo, S.-H. Zhang, Y.-W. Tai, and S.-M. Hu, “NeRF-SR: High-quality neural radiance fields using super-sampling,” 2021, arXiv:2112.01759.

[73] N. Rahaman et al., “On the spectral bias of neural networks,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 5301–5310.