Light Field Super-Resolution with Zero-Shot Learning

Zhen Cheng Zhiwei Xiong* Chang Chen Dong Liu Zheng-Jun Zha
University of Science and Technology of China

Abstract

Deep learning provides a new avenue for light field super-resolution (SR). However, the domain gap caused by drastically different light field acquisition conditions poses a main obstacle in practice. To fill this gap, we propose a zero-shot learning framework for light field SR, which learns a mapping to super-resolve the reference view with examples extracted solely from the input low-resolution light field itself. Given highly limited training data under the zero-shot setting, however, we observe that it is difficult to train an end-to-end network successfully. Instead, we divide this challenging task into three sub-tasks, i.e., pre-upsampling, view alignment, and multi-view aggregation, and then conquer them separately with simple yet efficient CNNs. Moreover, the proposed framework can be readily extended to finetune the pre-trained model on a source dataset to better adapt to the target input, which further boosts the performance of light field SR in the wild. Experimental results validate that our method not only outperforms classic non-learning-based methods, but also generalizes better to unseen light fields than state-of-the-art deep-learning-based methods when the domain gap is large.

1. Introduction

The 4D light field that records both angular and spatial information of light has been playing an increasing role in computer vision [33, 36, 43]. The commercialized light field cameras generally adopt micro-lens-array in front of the sensor, which poses an essential trade-off between the angular and spatial resolutions [16, 24]. The limited spatial resolution restricts the capability of light field in practical applications. Therefore, light field super-resolution (SR) has been an important and popular topic in the research community and attracts a lot of attention since the emergence of light field cameras [1, 2]. Recently, due to the prosperity of deep learning techniques, convolutional neural network (CNN) based methods have demonstrated promising performance for light field SR [10, 13, 23, 38, 39, 44, 45, 46, 49, 52], and the state-of-the-art methods exceed classic non-learning-based methods [4, 18, 30] with notable gains. Such performance boost is obtained by training well-engineered CNNs on a large external dataset to explore the 4D light field structures. However, these deep-learning-based methods inevitably face the domain shift problem [6], which hinders their capability of generalizing to unseen light fields with a large domain gap from the training set.

Domain shift, which means the performance drop when a deep neural network is trained on one dataset (source) but tested on another dataset (target), is much more severe in light field SR than that in single image SR. The underlying reason is that, light field SR exploits not only the 2D spatial correlation within each view, but also the 2D angular correspondence among different views (also called across-view redundancy). Such a 4D spatio-angular structure varies a lot between light fields captured by different acquisition systems and configurations. Take cameras using micro-lens-array for example, the baselines between the micro lens can be quite different in different types of cameras, resulting in distinct angular correspondences. Therefore, the network trained with a certain light field dataset could easily overfit to the spatio-angular structure within the given dataset and thus may not perform well on light fields in the wild.

To address this problem, we propose a zero-shot learning framework for light field SR, which learns a mapping to super-resolve the reference view with examples extracted solely from the input low-resolution (LR) light field itself. This work is inspired by the recently proposed zero-shot single image SR (ZSSR) method [34], which exploits across-scale recurrence within a single image and trains an SR network with paired examples extracted from the input LR image and its downscaled version. In this way, the input-specific model can generalize well on real images with unknown acquisition process, where abundant data for external training are not available. However, given highly limited training data under the zero-shot setting, we observe that it is difficult to train end-to-end SR networks successfully. Through a comparative study, we then find that a divide-and-conquer strategy, which explicitly divides the SR task into several sub-tasks and conquers them separately, can facilitate the learning of the SR mapping.

*Correspondence should be addressed to zwxiong@ustc.edu.cn

Specifically, our proposed zero-shot light field SR framework consists of three sub-tasks, *i.e.*, pre-upsampling, view alignment, and multi-view aggregation. We select the VDSR [15] network pre-trained on a 2D image dataset for preliminary upsampling. For view alignment, we design an alignment-oriented disparity estimation network following a plane-sweep volume generator, which can be readily trained in an unsupervised manner. After disparity-guided warping, an aggregation network is designed to aggregate the aligned views for high-frequency detail restoration, which can be trained with light field patch pairs extracted from the input LR light field and its downscaled version in a self-supervised manner. In this way, an input-specific SR model can be trained given an LR light field without any external light field dataset. Thanks to the divide-and-conquer strategy, the obtained model produces impressive high-resolution (HR) results even with highly limited training data, which outperforms state-of-the-art light field SR models when the domain gap is large.

The proposed zero-shot framework paves a way for light field SR in the wild, where abundant external training data that match the target input are not available. Moreover, this framework can be readily extended to finetune the pre-trained model on a source dataset to better adapt to the target input. Specifically, we propose an error-guided finetuning algorithm to handle the regions where the pre-trained model is less effective by selecting complementary training samples from the target input. In this way, new state-of-the-art results are generated for light field SR in the wild, which again validates the effectiveness of our method in closing the domain gap. We believe the zero-shot framework introduced in this paper could also inspire other inverse problems where high dimensional data acquiring with customized hardware are involved.

2. Related work

Classic light field SR. Light field SR aims to enhance the spatial resolution of the reference view from an LR light field by exploiting redundant information across different views. Classic non-learning-based methods utilize projection and optimization techniques to super-resolve the reference view, relying on geometric [18, 30] and mathematical [4, 41] modeling of the 4D light field structure. All of these classic light field SR methods are input-specific and will not have the domain shift problem. As a new input-specific solution, however, the proposed zero-shot framework achieves notably improved light field SR performance with deep internal learning.

CNN-based light field SR. CNN-based methods now dominate light field SR due to their promising performance. Yoon *et al.* [45] proposed the first light field SR network LFCNN, by reusing the SRCNN architecture [8] with multiple channels. After that, a number of CNNs have been designed to exploit across-view redundancy in the 4D light field, either explicitly [7, 13, 38, 49] or implicitly [23, 39, 44, 46, 52]. Although these well-designed CNNs outperform classic methods by a large margin, they always rely on an external training dataset, which inevitably suffer from the domain shift problem [6] for light field SR in the wild. In contrast, the proposed zero-shot framework uses the input light field only for training, or adapts to the input by finetuning a pre-trained model. In either way, the domain shift problem can be well addressed.

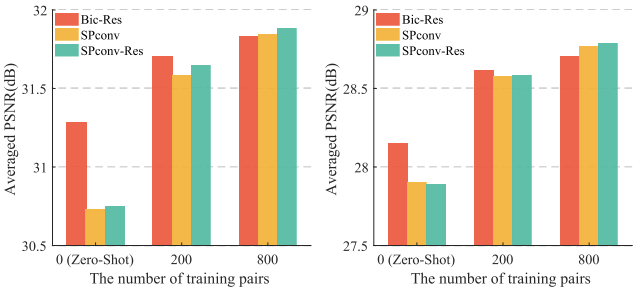
Domain adaptation. A popular solution for solving the domain shift problem is domain adaptation. By extracting shared features between source and target datasets with adversarial training at certain layers of the CNNs, domain adaptation has demonstrated promising performance for a number of tasks, *e.g.*, image classification [9, 20], semantic segmentation [17] and person re-identification [19]. However, such a technique needs sufficient training data for both source and target domains, which is not always feasible for light fields due to the acquisition cost in the unknown target domain. Instead, our zero-shot framework can be regarded as an efficient adaptation.

3. Divide-and-conquer strategy

Zero-shot SR uses the input LR image and its downscaled version as training pairs, which makes the amount of training data highly limited in nature. A question is brought out here: with such limited amount of training data, how to learn a better SR mapping?

To simplify the discussion, we start from the single image SR task. Specifically, we investigate three representative network architectures that are generally used for single image SR (network details are provided in the supplementary document). Bic-Res explicitly divides the SR task into two sub-tasks, *i.e.*, low-frequency information preservation and high-frequency detail restoration, and conquers them with bicubic interpolation and a residual learning CNN, respectively. SPconv denotes the network built with the popular learnable upsampling module, *i.e.*, sub-pixel convolution [32], in the tail of the network. SPconv is a typical end-to-end network, which directly takes the LR image as input and predicts the HR image as output. Compared with Bic-Res, SPconv enjoys higher efficiency due to learning in the LR feature space. SPconv-Res can be regarded as an updated version of SPconv, which further introduces residual learning by predicting the difference between the bicubic interpolated result and the groundtruth. Note that, however, SPconv-Res still takes the LR image as input and is thus an end-to-end network. The number of parameters in the three SR networks are set the same.

We then conduct a comparative study by training the above SR networks with different amounts of external training data: from zero (only the input LR image itself, *i.e.*,



(a) Scale 2 (b) Scale 3

Figure 1. PSNR results on the BSD100 dataset of different single image SR networks *w.r.t* the number of external training samples from the DIV2K dataset.

zero-shot learning) to the whole DIV2K dataset (800 images in total) [3]. The averaged testing PSNR over the BSD100 dataset (100 images in total) [22] against the number of images used for training is shown at three representative points in Fig. 1, at two typical scaling factors. As can be seen, Bic-Res performs notably better ($>$0.5 dB at the scaling factor of 2) in comparison with SPconv and SPconv-Res under the zero-shot setting. Meanwhile, this performance superiority diminishes with the increasing amount of training data. When a total of 800 external images are used for training, Bic-Res loses its advantage and is less competitive to SPconv and SPconv-Res.

Through the above study, we can observe that, when the amount of external training data is large, end-to-end networks (SPconv and SPconv-Res) do provide more favorable SR results than the network requiring pre-upsampling (Bic-Res), in addition to their higher efficiency. This is in line with the established experience, as end-to-end networks are now the mainstream for single image SR. However, the observation reverses when abundant data for external training are not available, especially under the zero-shot setting. In this case, it is better to divide the SR task into a few easier sub-tasks (*i.e.*, low-frequency information preservation and high-frequency detail restoration here). Such a strategy, named as divide-and-conquer, may narrow the space of parameter search, and thus facilitates the learning of the SR mapping compared to end-to-end networks.

The task of light field SR is more complicated than single image SR since it needs to take across-view redundancy into account. Therefore, we divide the light field SR task into three sub-tasks: pre-upsampling, view alignment, and multi-view aggregation. The three sub-tasks, with respective CNN backbones and corresponding operations, are unified to form a zero-shot learning framework. It is worth mentioning that, however, the CNN backbones we deploy in the three sub-tasks as in this paper are not the only possible embodiments. That is to say, they could be replaced by more advanced structures. Instead of paying attention to explore complicated networks, our focus is to realize a feasible zero-shot SR framework with highly limited training data from the input light field itself. To this end, we prefer simple structures that are more friendly to zero-shot learning in implementation.

4. Zero-shot light field SR

Fig. 2 illustrates our proposed zero-shot light field SR framework. Without loss of generality, we take SR on the central view as an example, which can be readily applied to other reference views. $Z^{LR} \in \mathbb{R}^{U\times V\times X\times Y}$ denotes the input LR light field with angular resolution of $U \times V$ and spatial resolution of $X \times Y$. The superscript *LR* can be replaced by *LLR* to denote the downscaled version of the LR input. $\mathcal{U} = \{\boldsymbol{u}|\boldsymbol{u} = [u, v], 1 \leq u \leq U, 1 \leq v \leq V\}$ denotes the set of 2D angular coordinates in Z^{LR} and $\boldsymbol{u_c} = [u_c, v_c]$ indexes the central view. α denotes the scaling factor. In the testing phase, the input is Z^{LR} and the expected output is the super-resolved central view $S^{LR}[\boldsymbol{u_c}] \in \mathbb{R}^{\alpha X\times \alpha Y}$. As aforementioned, the SR process consists of three sub-tasks: pre-upsampling, view alignment, and multi-view aggregation. The implementation details of each sub-task are given below.

4.1. Pre-upsampling

Pre-upsampling is used to upsample the LR light field in the first place to preserve the low-frequency information and provide an initial super-resolved light field that matches the target resolution. The up-sampled light field is further used for view alignment and multi-view aggregation. While bicubic interpolation is a straightforward operation for pre-upsampling, we find that a simple single image SR network, *e.g.*, VDSR [15], is a better choice. On the one hand, this network can be pre-trained on a 2D image dataset, which will not suffer from domain shift caused by light field acquisition. On the other hand, the single image SR result provides a more favorable initialization than the interpolated ones. While we use VDSR for pre-upsampling throughout this paper, we also conduct ablation study when this sub-task is realized by bicubic interpolation (Sec. 7).

4.2. View alignment

View alignment in either feature or image space has been exploited for multi-view image SR, which can be divided into two categories: implicit methods and explicit methods. The former regard the redundancy as latent features, *e.g.*, attention [35] and deformable offsets [37], while the latter cast the redundancy as geometry correspondences, *e.g.*, disparity [12] and optical flow [31]. To make the learning easier, we follow the explicit way and introduce an alignment-oriented disparity estimation network (AlignNet) with a trainable parameter set $\boldsymbol{\Theta_1}$ for view alignment.

As shown in Fig. 2, given an LR light field Z^{LR} and a training patch size M, we extract 4D light field patches

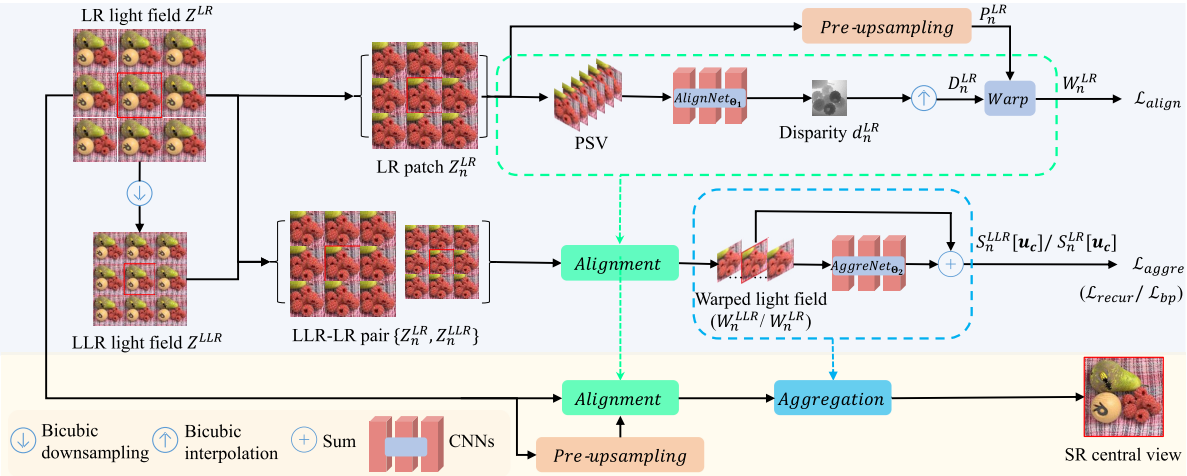


Figure 2. Our proposed zero-shot light field SR framework. During the training phase, we extract patches solely from the input LR light field to train both AlignNet and AggreNet using $\mathcal{L}_{align}$ and $\mathcal{L}_{aggre}$, respectively. Then in the testing phase, we use the trained networks to inference an HR central view from the input LR light field. Details of the CNN backbones can be found in the supplementary document.

$Z_n^{LR} \in \mathbb{R}^{U\times V\times M\times M}, n = 1, 2, \cdots, N$ for training AlignNet. Instead of feeding the light field into the network directly, we generate a plane-sweep volume (PSV), which is proved to be efficient for scene geometry inference [25, 26, 27, 51], as the input of AlignNet. As the output, we get an estimated disparity map via

$$d_n^{LR} = AlignNet_{\mathbf{\Theta_1}}(PSV(Z_n^{LR})). \tag{1}$$

The backbone of AlignNet follows the one proposed in [27] (details are provided in the supplementary document), which is quite simple and can be trained in an unsupervised manner. For further operations in the HR space, we then upsample the LR disparity map to the target resolution as $D_n^{LR} = d_n^{LR} \uparrow_\alpha$, where $\uparrow_\alpha$ denotes the bicubic interpolation with the scaling factor of α.

On the other hand, we already obtain the pre-upsampled LR light field as $P_n^{LR} = VDSR(Z_n^{LR}, \alpha)$. Having D_n^{LR}, we can warp each view in P_n^{LR} to the central view and get an aligned light field at the target resolution. This process can be represented as

$$W_n^{LR} = Warp(P_n^{LR}, D_n^{LR}). \tag{2}$$

Each view in this warped light field W_n^{LR} should be as similar as the pre-upsampled central view $P_n^{LR}[\boldsymbol{u_c}]$. Therefore, the loss function for training AlignNet is defined as

$$\mathcal{L}_{align} = \frac{1}{N}\sum_{n=1}^{N}\sum_{\boldsymbol{u}\in\mathcal{U}} \|W_n^{LR}[\boldsymbol{u}] - P_n^{LR}[\boldsymbol{u_c}]\|_2^2. \tag{3}$$

Since neither HR light field nor HR disparity is involved in $\mathcal{L}_{align}$, AlignNet can be trained with patches Z_n^{LR} that are extracted solely from the input LR light field. Meanwhile, the trained AlignNet naturally adapts to the geometric structure of the input light field, which provides an efficient and domain-shift-free solution for the view alignment sub-task. As demonstrated in the ablation study (Sec. 7), the view alignment plays an essential role in our zero-shot learning framework.

4.3. Multi-view aggregation

In the aligned light field W_n^{LR}, pixels in different views sampled from neighboring scene points become much closer in the spatial dimension due to the disparity-guided warping. It thus facilitates the exploitation of complementary information from these pixels and their local neighbors to enrich the high-frequency details in the pre-upsampled result. This sub-task is fulfilled by a multi-view aggregation network (AggreNet) with a trainable parameter set $\mathbf{\Theta_2}$, which can be trained in a self-supervised manner.

As shown in Fig. 2, we extract 4D light field patches from both Z^{LLR} and Z^{LR} to form the LLR-LR pairs $\{Z_n^{LLR}, Z_n^{LR}\}, n = 1, 2, \cdots, N$ for training AggreNet. We first feed Z_n^{LLR} into the pretrained AlignNet, which generates the aligned light field W_n^{LLR}. Take W_n^{LLR} as input, AggreNet predicts the residual between the pre-upsampled LLR central view $P^{LLR}[\boldsymbol{u_c}]$ and the LR central view $Z^{LR}[\boldsymbol{u_c}]$. Here we adopt a simple structure previously used in [14] as the backbone of AggreNet (details are provided in the supplementary document). The SR result of the LLR central view can be represented as

$$S_n^{LLR}[\boldsymbol{u_c}] = AggreNet_{\mathbf{\Theta_2}}(W_n^{LLR}) + P_n^{LLR}[\boldsymbol{u_c}]. \tag{4}$$

Following zero-shot single image SR [34], an L_1 loss between the SR results of LLR inputs and the LR labels is adopted for training AggreNet as

$$\mathcal{L}_{recur} = \frac{1}{N}\sum_{n=1}^{N} \|S_n^{LLR}[\boldsymbol{u_c}] - Z_n^{LR}[\boldsymbol{u_c}]\|_1. \tag{5}$$

In this way, AggreNet can be trained without the need of HR light field. However, as analyzed in [53], in a natural image, the more gradient contents are within a patch, the less this patch recurs across scales. For regions with abundant gradient contents, which are always high-frequency textures

or edges, there may not be enough LLR-LR pairs to train the network. Therefore, AggreNet trained with only the recurrence loss $\mathcal{L}_{recur}$ could overfit to the regions with less gradient contents, resulting missing high-frequency details in the super-resolved image. To alleviate this problem, we propose an additional back-projection loss.

This time, we feed Z_n^{LR} into the pretrained AlignNet, which generates the aligned light field W_n^{LR}. Taking W_n^{LR} as input, AggreNet generates the super-resolved LR central view $S_n^{LR}[\boldsymbol{u_c}]$. Since the ground truth HR central view is not available under the zero-shot setting, we then calculate the L_1 distance between the downsampled SR result and the LR central view as

$$\mathcal{L}_{bp} = \frac{1}{N}\sum_{n=1}^{N} \|S_n^{LR}[\boldsymbol{u_c}] \downarrow_\alpha - Z_n^{LR}[\boldsymbol{u_c}]\|_1, \tag{6}$$

where $\downarrow_\alpha$ denotes the bicubic downsampling operation. As demonstrated in the ablation study (Sec. 7), this back-projection loss improves the SR performance when HR labels are not available. The complete loss function for training AggreNet is $\mathcal{L}_{aggre} = \mathcal{L}_{recur} + \gamma_1 \mathcal{L}_{bp}$, where γ_1 is a weighting factor.

As discussed above, AlignNet and AggreNet are both trained with patches extracted from the LR light field as well as its downscaled version, either in an unsupervised or in a self-supervised manner. Therefore, the whole framework can be regarded zero-shot as no external light field dataset is needed for training. For more efficient training of the whole framework, we introduce a three-stage training strategy. At the first stage, we train AlignNet only. Then, we fix the parameters of AlignNet and train AggreNet only. Finally, to avoid possible error accumulation, we jointly train all the parameters of AlignNet and AggreNet with a combined loss function $\mathcal{L}_{joint} = \mathcal{L}_{aggre} + \gamma_2 \mathcal{L}_{align}$, where γ_2 is a weighting factor. Ablation study (Sec. 7) validates the effectiveness of this three-stage training strategy.

5. Zero-shot learning as finetuning

The above zero-shot SR framework adapts to the specific input light field and thus gets rid of domain shift in nature. However, it still has certain limitations. On the one hand, for regions that are difficult to align, such as non-Lambertian surfaces, the learning of AlignNet could be ineffective. On the other hand, as an inherent shortcoming of zero-shot SR, for regions without across-scale recurrence, there lacks training data for the learning of AggreNet. These circumstances hinder the performance of zero-shot learning to a certain extent. Given a source light field dataset with large domain gap with the target input, is it possible for the proposed framework to exploit the useful information in the source while adapting to the target? The answer is YES: the proposed framework can be used to pre-train a model with the source dataset firstly, and the parameters of this pre-trained model can be then finetuned with the target light field by using our framework again.

Algorithm 1 Error-guided finetuning algorithm

Require: LLR light field Z^{LLR} and LR light field Z^{LR},
Pre-trained network parameters $\boldsymbol{\Theta_1}$ (AlignNet) and $\boldsymbol{\Theta_2}$ (AggreNet).
Initialize AlignNet and AggreNet with $\boldsymbol{\Theta_1}$ and $\boldsymbol{\Theta_2}$, respectively;
Input Z^{LR} and get aligned pre-upsampled light field W^{LR};
Calculate averaged alignment error map by
$E_{align} = \frac{1}{|\mathcal{U}|}\sum_{\boldsymbol{u}\in\mathcal{U}} |W^{LR}[\boldsymbol{u}] - P^{LR}[\boldsymbol{u_c}]|$;
Downsample E_{align} to the resolution of Z^{LR} and get E_{align}^{LR};
Normalize E_{align}^{LR} by $p_{align} = E_{align}^{LR}/(sum(E_{align}^{LR}))$;
while *AlignNet* does not converge **do**
  Choose patch Z_n^{LR} with probability map p_{align};
  Feed forward patch Z_n^{LR} and update $\boldsymbol{\Theta_1}$ with $\mathcal{L}_{align}$.
end while
Fix the finetuned AlignNet;
Input Z^{LLR} and get the SR result $S^{LLR}[\boldsymbol{u_c}]$;
Calculate SR error by
$E_{recur} = |S^{LLR}[\boldsymbol{u_c}] - Z^{LR}[\boldsymbol{u_c}]|$;
Normalize E_{recur} by $p_{recur} = E_{recur}/sum(E_{recur})$;
while *AggreNet* does not converge **do**
  Choose patch pair $\{Z_n^{LLR}, Z_n^{LR}\}$ with probability map p_{recur};
  Feed forward the selected patch pair and update $\boldsymbol{\Theta_2}$ with $\mathcal{L}_{aggre}$.
end while

Specifically, when training an initial model with a source dataset, we do not need to fall back on LLR-LR light field pairs for training AggreNet since LR-HR pairs are available now. Instead, the L_1 distance between the super-resolved LR central view and the ground truth HR central view is calculated as the loss function $\mathcal{L}_{aggre}$. Once the pre-trained model is obtained, we apply an error-guided finetuning algorithm to the target input Z^{LR} as summarized in Algorithm 1. First, by using the pre-trained AlignNet, we generate the aligned light field W^{LR} from Z^{LR} and calculate the averaged absolute error map $E_{align} \in \mathbb{R}^{\alpha X \times \alpha Y}$ between each view of W^{LR} and the pre-upsampled LR central view $P^{LR}[\boldsymbol{u_c}]$. This error map is downsampled to the input resolution and normalized to range $[0, 1]$, resulting in a probability map p_{align}. According to this probability map, we randomly select patches from L^{LR} for the finetuning of AlignNet. Due to the fact that larger probability values indicate larger alignment errors, the finetuning of AlignNet would pay more attention to the regions that are not well-aligned by the pre-trained model.

Then, by fixing the finetuned AlignNet and using the pre-trained AggreNet, we generate super-resolved LLR central view $S^{LLR}[\boldsymbol{u_c}]$, and calculate the absolute error map $E_{recur} \in \mathbb{R}^{X \times Y}$ between $S^{LLR}[\boldsymbol{u_c}]$ and the LR central view $Z^{LR}[\boldsymbol{u_c}]$. This error map is normalized to range $[0, 1]$, resulting in a probability map p_{recur}. According to this probability map, we randomly select patch from Z^{LLR} and Z^{LR} in pair for the finetuning of AggreNet. Due to the fact that larger probability values indicate larger aggregation errors, the finetuning of AggreNet will also pay more attention to the regions that are not well-aggregated by the pre-trained model. As demonstrated in the ablation study

(Sec. 7), our proposed error-guided finetuning algorithm is more effective compared with a plain finetuning process.

6. Experimental results

Datasets and evaluation metrics. To validate the effectiveness of the proposed method, we use a large real-world light field dataset HFUT (640 scenes) [47] and a large synthetic dataset SAE (180 scenes) [5] as source datasets and other four relatively small datasets as target datasets. The target datasets include Stanford Lytro Archive (Stan) [28], EPFL [29], HCI old (HCI1) [42], and HCI new (HCI2) [11]. The former two are real-world datasets and the latter two are synthetic datasets. There exists obvious domain gap between real-world and synthetic datasets due to their distinct imaging models, and different real-world/synthetic datasets also have domain gap due to different acquisition conditions. The number of scenes in the target datasets are 20, 20, 10, and 20, respectively. In addition, for real-world datasets, we extract central 9×9 views to avoid the vignetting effect. For evaluation, we use two distortion metrics PSNR (dB) and SSIM [40] and two perception metrics VGG distance [48] and Ma's score [21].

Comparison methods. We compare the SR results of the central view through a number of representative methods including three categories: 1) single image SR: bicubic interpolation (BIC), VDSR [15], and ZSSR [34], 2) classic light field SR: the method using graph-based regularization based on geometric modeling (GBSQ [30]) and the method iteratively alternating between LFBM5D filter and back-projection (BM5D [4]), 3) deep-learning-based light field SR (SoTA): the multi-stream residual network (ResLF [49]), the all-to-one network using combinatorial geometry embedding and structural consistency (ATO [13]), and the latest network using spatial-angular interaction modules (InterNet [39]). We implement all comparison methods with the public code provided by the authors[1]. We train VDSR with the DIV2K dataset [3] and ZSSR with the LR central view from the input light field. As for deep-learning-based light field SR methods, we train them using the source datasets and test them on the target datasets. More implementation details are provided in the supplementary document.

Results without source dataset. The first set of rows in Table 1 shows the comparison results of methods without using any source light field dataset. In other words, these methods will not have the domain shift problem when testing on the target datasets. As can be seen, our zero-shot method (Ours-ZS) achieves the best performance on all target datasets at different scaling factors. Specifically, Ours-ZS has a notable improvement over our pre-upsampler VDSR [15] and ZSSR [34] which only exploits across-scale recurrence. These results validate the effectiveness of alignment and aggregation in our zero-shot learning framework. Through these two sub-tasks, across-view redundancy in the light field is fully exploited for high-frequency detail restoration. Such improvement on high-frequency details can also be observed in the visual results in Fig. 3. On the other hand, compared with classic methods GBSQ [30] and BM5D [4], Ours-ZS also has obvious performance superiority. Such superiority reflects that, with effective zero-shot learning, the internal correspondence within a 4D light field can be better exploited.

[1] ATO [13] does not provide official implementation for scale 3.

Table 1. PSNR results of different methods. In the second and third sets of rows, (S) denotes that the source dataset is SAE and (H) denotes HFUT. The subscript † denotes real-world datasets while § denotes synthetic datasets. Gray background indicates large domain gap. The results of SSIM [40], VGG distance [48] and Ma's score [21] are provided in the supplementary document, along with angular consistency analysis.

Method	Scale 2				Scale 3			
	Stan†	EPFL†	HCI1§	HCI2§	Stan†	EPFL†	HCI1§	HCI2§
BIC	33.83	31.66	36.30	34.67	30.42	28.99	32.72	31.92
VDSR	37.03	34.00	38.86	37.08	32.82	30.88	34.58	33.78
ZSSR	36.17	33.34	38.62	36.60	31.90	30.11	34.02	33.05
GBSQ	34.77	32.82	37.26	36.95	30.35	29.96	34.48	33.38
BM5D	36.44	33.39	39.43	36.99	32.38	30.66	35.02	33.99
Ours-ZS	**37.91**	**34.51**	**39.74**	**37.75**	**33.27**	**31.09**	**35.15**	**34.04**
ResLF(S§)	37.16	33.87	39.39	37.63	33.16	31.18	35.72	34.74
ATO(S§)	37.55	34.45	39.24	37.52	–	–	–	–
InterNet(S§)	37.56	34.05	39.51	37.82	33.56	31.34	35.71	**34.80**
Ours-Pre(S§)	37.36	34.29	39.42	37.56	33.36	31.33	35.53	34.58
Ours-FT(S§)	**37.97**	**34.64**	**39.98**	**37.85**	**33.77**	**31.48**	**35.90**	34.69
ResLF(H†)	37.13	34.11	38.58	35.82	33.50	31.35	35.66	34.01
ATO(H†)	37.87	34.51	39.60	37.16	–	–	–	–
InterNet(H†)	37.92	34.41	39.69	37.17	**34.05**	31.39	35.41	33.89
Ours-Pre(H†)	37.55	34.55	39.42	37.22	33.70	31.53	35.55	34.19
Ours-FT(H†)	**38.27**	**35.21**	**40.38**	**38.22**	33.85	**31.57**	**35.71**	**34.47**

Results with source dataset. The second and third sets of rows of Table 1 show the comparison results with methods using a source light field dataset (SAE or HFUT). Existing deep-learning-based methods inevitably face domain shift in this case, when testing on the target datasets with a large domain gap. As can be seen, at the scaling factor of 2, when the domain gap is large (*i.e.*, source is synthetic and target is real-world, or the opposite), Ours-ZS gives superior performance compared with three SoTA methods. For example, when the source is HFUT and the target is HCI2, Ours-ZS has 1.83/0.59/0.58 dB gains over ResLF, ATO, and InterNet, respectively. Even when the domain gap is not that large (*i.e.*, source and target are both synthetic or real-world), Ours-ZS still has better or comparable performance with these SoTA methods. Such results validate the adaptation ability of our zero-shot method without the usage of a source dataset.

Performance boost with finetuning. Furthermore, when our zero-shot learning framework is used to finetune a pre-trained model obtained with this framework (Ours-

Figure 3. Visual comparisons of super-resolved central view (cropped for a better visualization) through different methods at the scaling factor of 2. The input of the first scene (*Sideboard*) is the downsampled light field while the input of the second scene (*Origami*) is the original light field. Zoom in the figure for a better visual experience. More visual results are provided in the supplementary document.

Pre), the performance of the resulting model (Ours-FT) can be boosted by a large margin. For example, when the source is HFUT and the target is HCI2 (domain gap is large), Ours-FT gains over ResLF, ATO, and InterNet with 2.40/1.06/1.05 dB. Such improvement validates the effectiveness of combining our zero-shot framework and error-guided finetuning algorithm together.

More challenging case. Under the scaling factor of 3, Ours-ZS does not hold its performance superiority due to extremely limited training data. For instance, the spatial resolution of an LLR light field from the Stanford dataset is only 40×60 during the training of AggreNet. With such limited training data, although Ours-ZS still outperforms classic light field SR methods, it loses advantage compared with the SoTA methods using a large source dataset. However, also using the source dataset, Ours-FT makes up the shortcomings of Ours-ZS and gives much better results. It again outperforms the SoTA methods when the domain gap is large. For example, when the source is SAE and the target is Stan, Ours-FT gains over ResLF and InterNet with 0.51/0.21 dB; when the source is HFUT and the target is HCI2, Ours-FT gains over ResLF and InterNet with 0.46/0.59 dB. On the other hand, when the domain gap is not large, Ours-FT may not keep the best performance, *e.g.*, from SAE to HCI2, and from HFUT to Stan. It is worth mentioning that, however, the results with respect to different scaling factors give additional information. That is, the performance of our zero-shot method highly depends on the resolution of the input light field. In the above experiments, for the purpose of calculating quantitative metrics, the input light field is actually downsampled from the original one, making it challenging for the zero-shot learning. In other words, the resolution of the input light field could be much higher in practical applications, which indicates more sufficient data for training our zero-shot model and the superiority of our method will be highlighted.

Visual comparison. In addition to the numerical results, we also show some visual examples in Fig. 3. Among methods without source datasets, Ours-ZS recovers more detailed textures and cleaner edges than others. When training with source datasets, SoTA deep-learning-based methods suffer from blurring and aliasing artifacts due to the domain gap, while Ours-FT gets rid of these artifacts and recovers more realistic high-frequency details. Such comparisons demonstrate that our zero-shot framework provides an advanced solution for light field SR in the wild.

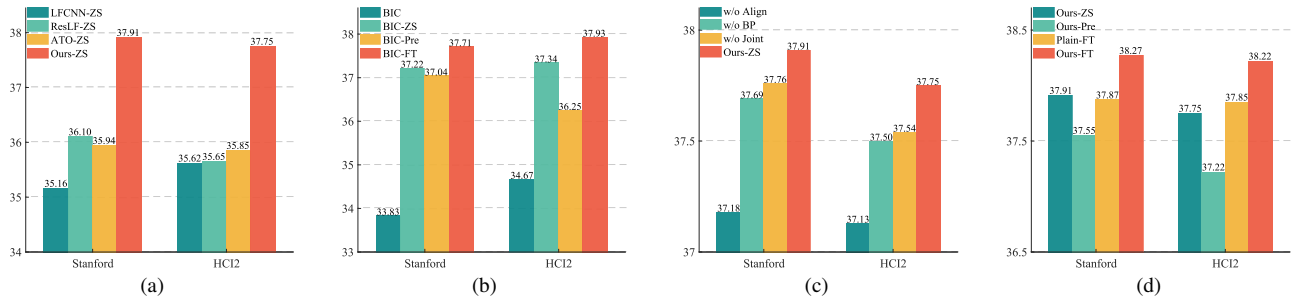


Figure 4. PSNR results for ablation study. (a) Divide-and-conquer *vs.* End-to-end. (b) Ablations on pre-upsampling. (c) Ablations on sub-tasks, training stages, and loss functions. (d) Ablations on finetuning methods.

7. Ablation study

To validate the effectiveness of different components of the proposed framework, we conduct comprehensive ablation studies on the Stanford and the HCI2 datasets at the scaling factor of 2.

Divide-and-conquer vs. End-to-end. To validate the superiority of the divide-and-conquer strategy over an end-to-end network, we train several end-to-end models under the zero-shot setting. Specifically, we choose LFCNN [45], ResLF [49] and the coarse stage of ATO [13]. These networks take LR light field as input and predict the HR reference view as output. As can be seen in Fig. 4(a), end-to-end networks perform much worse (*e.g.*, PSNR drops by 2.75/1.81/1.97 dB on the Stanford dataset) than our divide-and-conquer method under the zero-shot setting. These results reflect that, with highly limited training data, the divide-and-conquer strategy facilitates the learning of the SR mapping.

Pre-upsampling. To investigate the influence of pre-upsampling, we replace VDSR with bicubic interpolation and keep the other parts the same. As shown in Fig. 4(b), even using bicubic interpolation as pre-upsampler, the performance of zero-shot learning is also promising. These results not only validate the effectiveness of the alignment and aggregation sub-tasks, but also reflect the potential of our framework, *i.e.*, using an even powerful pre-upsampler than VDSR could further elevate its performance (see the results using RCAN [50] in the supplementary document).

Alignment. To validate the necessity of view alignment, we remove this sub-task in our framework and feed the pre-upsampled LR light field into AggreNet directly. The performance comparison is shown in Fig. 4(c). We can see that, without view alignment, PSNR drops severely by 0.73/0.62 dB on the two datasets, respectively. These results suggest that view alignment plays an important role for zero-shot light field SR, which facilitates the aggregation of useful information from different views.

Back-projection loss. As a complement to the recurrence loss $\mathcal{L}_{recur}$, we introduce the back-projection loss $\mathcal{L}_{bp}$ for training AggreNet. For ablation, we remove this loss term and find in Fig. 4(c) that, without this loss, PSNR drops by 0.22/0.25 dB on the two datasets. These results validate the role of the back-projection loss when no HR label is available under the zero-shot setting.

Joint training stage. For the zero-shot framework, we conduct a three-stage training in which the joint training stage is to avoid possible error accumulation. Fig. 4(c) shows the performance without this stage. As can be seen, this stage provides about 0.2 dB improvement. It suggests that the error accumulation does exist and can be alleviated by the joint training stage.

Error-guided finetuning. When our zero-shot framework is used to finetune a pre-trained model, we propose an error-guided finetuning algorithm instead of a plain finetuning process. Fig. 4(d) shows the performances of these two finetuning strategies. As can be seen, the plain finetuning leads to a PSNR drop of about 0.4 dB, which validates the superiority of our error-guided finetuning algorithm.

8. Conclusion

The main contributions of this work are summarized as the conclusion: 1) we propose the first zero-shot learning framework for light field SR, which learns an input-specific SR mapping with examples extracted solely from the input LR light field itself. 2) We analyze different learning strategies under the zero-shot setting, and propose a divide-and-conquer strategy for effective learning from highly limited training data. 3) We propose an error-guided finetuning algorithm to further extend our zero-shot framework for jointly using a source dataset and the target input. 4) We validate the superiority of the proposed framework against SoTA light field SR methods through comprehensive experiments, both quantitatively and qualitatively.

Acknowledgements

We acknowledge funding from National Key R&D Program of China under Grant 2017YFA0700800, and Natural Science Foundation of China under Grant U19B2038.

References

[1] *https://www.lytro.com/*. 1

[2] *https://www.raytrix.de/*. 1

[3] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017. 3, 6

[4] Martin Alain and Aljosa Smolic. Light field super-resolution via lfbm5d sparse coding. In *ICIP*, 2018. 1, 2, 6

[5] Anna Alperovich, Ole Johannsen, Michael Strecke, and Bastian Goldluecke. Light field intrinsics with a deep encoder-decoder network. In *CVPR*, 2018. 6

[6] Zhen Cheng, Zhiwei Xiong, Chang Chen, and Dong Liu. Light field super-resolution: A benchmark. In *CVPRW*, 2019. 1, 2

[7] Zhen Cheng, Zhiwei Xiong, and Dong Liu. Light field super-resolution by jointly exploiting internal and external similarities. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2604–2616, 2020. 2

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 2

[9] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011. 2

[10] M. Shahzeb Khan Gul and Bahadir K. Gunturk. Spatial and angular resolution enhancement of light fields using convolutional neural networks. *IEEE Transactions on Image Processing*, 27(5):2146–2159, 2018. 1

[11] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *ACCV*, 2016. 6

[12] Daniel S Jeon, Seung-Hwan Baek, Inchang Choi, and Min H Kim. Enhancing the spatial resolution of stereo images using a parallax prior. In *CVPR*, 2018. 3

[13] Jing Jin, Junhui Hou, Jie Chen, and Sam Kwong. Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In *CVPR*, 2020. 1, 2, 6, 8

[14] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics*, 35(6):1–10, 2016. 4

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 2, 3, 6

[16] Anat Levin, William T Freeman, and Frédo Durand. Understanding camera trade-offs through a bayesian analysis of light field projections. In *ECCV*, 2008. 1

[17] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, 2019. 2

[18] Chia-Kai Liang and Ravi Ramamoorthi. A light transport framework for lenslet light field cameras. *ACM Transactions on Graphics*, 34(2):16:1–16:19, 2015. 1, 2

[19] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *CVPR*, 2019. 2

[20] Yajing Liu, Xinmei Tian, Ya Li, Zhiwei Xiong, and Feng Wu. Compact feature learning for multi-domain image classification. In *CVPR*, 2019. 2

[21] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 6

[22] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001. 3

[23] Nan Meng, Hayden Kwok-Hay So, Xing Sun, and Edmund Lam. High-dimensional dense residual convolutional neural network for light field reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):873–886, 2021. 1, 2

[24] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report, Stanford University*, 2(11):1–11, 2005. 1

[25] Jürg Nievergelt and Franco P. Preparata. Plane-sweep algorithms for intersecting geometric figures. *Communications of the ACM*, 25(10):739–747, 1982. 4

[26] Jiayong Peng, Zhiwei Xiong, Dong Liu, and Xuejin Chen. Unsupervised depth estimation from light field using a convolutional neural network. In *3DV*, 2018. 4

[27] Jiayong Peng, Zhiwei Xiong, Yicheng Wang, Yueyi Zhang, and Dong Liu. Zero-shot depth estimation from light field using a convolutional neural network. *IEEE Transactions on Computational Imaging*, 6:682–696, 2020. 4

[28] Abhilash Sunder Raj, Michael Lowney, Raj Shah, and Gordon Wetzstein. Stanford lytro light field archive. `http://lightfields.stanford.edu/LF2016.html`. 6

[29] Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In *International Conference on Quality of Multimedia Experience*, 2016. 6

[30] M. Rossi and P. Frossard. Graph-based light field super-resolution. In *MMSP*, 2017. 1, 2, 6

[31] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, 2018. 3

[32] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 2

[33] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim. Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *CVPR*, 2018. 1

[34] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *CVPR*, 2018. 1, 4, 6

[35] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *CVPR*, 2019. 3

[36] Tiantian Wang, Yongri Piao, Xiao Li, Lihe Zhang, and Huchuan Lu. Deep learning for light field saliency detection. In *ICCV*, 2019. 1

[37] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019. 3

[38] Yunlong Wang, Fei Liu, Kunbo Zhang, Guangqi Hou, Zhenan Sun, and Tieniu Tan. Lfnet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution. *IEEE Transactions on Image Processing*, 27(9):4274–4286, 2018. 1, 2

[39] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Spatial-angular interaction for light field image super-resolution. In *ECCV*, 2020. 1, 2, 6

[40] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[41] Sven Wanner and Bastian Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):606–619, 2014. 2

[42] Sven Wanner, Stephan Meister, and Bastian Goldlücke. Datasets and benchmarks for densely sampled 4d light fields. In *International Symposium on Vision Modeling and Visualization*, 2013. 6

[43] Gaochang Wu, Belen Masia, Adrian Jarabo, Yuchen Zhang, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field image processing: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):926–954, 2017. 1

[44] Henry Wing Fung Yeung, Junhui Hou, Xiaoming Chen, Jie Chen, Zhibo Chen, and Yuk Ying Chung. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Transactions on Image Processing*, 28(5):2319–2330, 2018. 1, 2

[45] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Learning a deep convolutional network for light-field image super-resolution. In *ICCVW*, 2015. 1, 2, 8

[46] Yan Yuan, Ziqi Cao, and Lijuan Su. Light-field image super-resolution using a combined deep cnn based on epi. *IEEE Signal Processing Letters*, 25(9):1359–1363, 2018. 1, 2

[47] Miao Zhang, Wei Ji, Yongri Piao, Jingjing Li, Yu Zhang, Shuang Xu, and Huchuan Lu. Lfnet: Light field fusion network for salient object detection. *IEEE Transactions on Image Processing*, 29:6276–6287, 2020. 6

[48] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[49] Shuo Zhang, Youfang Lin, and Hao Sheng. Residual networks for light field image super-resolution. In *CVPR*, 2019. 1, 2, 6, 8

[50] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 8

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

[52] Hao Zhu, Mantang Guo, Hongdong Li, Qing Wang, and Antonio Robles-Kelly. Revisiting spatio-angular trade-off in light field cameras and extended applications in super-resolution. *IEEE Transactions on Visualization and Computer Graphics*, 2019. 1, 2

[53] Maria Zontak and Michal Irani. Internal statistics of a single natural image. In *CVPR*, 2011. 4