Light Field Super-Resolution with Zero-Shot Learning
Supplementary Material

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

This supplementary document is organized as follows:

Sec. 1 provides numerical results of SSIM [13], VGG distance [14] and Ma's score [7].

Sec. 2 provides analysis on the angular consistency when SR is performed on the whole light field.

Sec. 3 provides additional visual comparisons.

Sec. 4 provides the details of networks (Bic-Res, SPconv, and SPconv-Res) used for investigating the divide-and-conquer strategy.

Sec. 5 provides the details of networks (AlignNet and AggreNet) used in our zero-shot light field SR framework.

Sec. 6 provides additional implementation details of our method.

1. More numerical results

In addition to the PSNR metric discussed in the paper, we further use SSIM [13], VGG distance [14] and Ma's score [7] for quantitative evaluation. As shown in Table 1, comparison results on these metrics are basically consistent with PSNR as we analyzed in the paper. The proposed method shows distinct advantage for light field SR in the wild, especially when the domain gap is large between source and target.

Method	Source	Stanford†	EPFL†	HCI1§	HCI2§
BIC	–	0.9281/ 0.2286/ 5.7024	0.9044/ 0.2809/ 5.2326	0.9309/ 0.2029/ 5.3897	0.9039/ 0.3143/ 5.7085
VDSR [4]	–	0.9551/ 0.1063/ 7.9528	0.9341/ 0.1432/ 7.4976	0.9542/ 0.0933/ 7.3667	0.9328/ 0.1174/ 7.7669
ZSSR [10]	–	0.9516/ 0.1110/ 7.8935	0.9313/ 0.1346/ 7.3973	0.9526/ 0.0912/ 7.0649	0.9296/ 0.1271/ 7.7009
GBSQ [8]	–	0.9389/ 0.2117/ 7.2242	0.9276/ 0.1877/ 6.5462	0.9411/ 0.1590/ 5.9527	0.9378/ 0.1738/ 6.8015
BM5D [1]	–	0.9521/ 0.1179/ 7.6752	0.9322/ 0.1376/ 7.2373	0.9586/ 0.0809/ 6.8968	0.9363/ 0.1280/ 7.5263
Ours-ZS	–	**0.9610/ 0.0824/ 8.0565**	**0.9432/ 0.1064/ 7.5660**	**0.9591/ 0.0675/ 7.3980**	**0.9402/ 0.0927/ 7.8337**
ResLF [15]	SAE§	0.9555/ 0.0992/ 7.8947	0.9380/ 0.1206/ 7.4571	0.9602/ **0.0663**/ 7.3139	0.9401/ 0.0935/ 7.7154
ATO [3]	SAE§	0.9586/ 0.0935/ 7.9097	0.9376/ 0.1285/ 7.4748	0.9561/ 0.0777/ 7.2510	0.9371/ 0.0973/ 7.6929
InterNet [12]	SAE§	0.9593/ 0.0986/ 7.9343	0.9417/ 0.1185/ 7.4885	**0.9619**/ 0.0703/ 7.2639	**0.9439**/ 0.0959/ 7.7326
Ours-Pre	SAE§	0.9571/ 0.1000/ 7.8887	0.9374/ 0.1298/ 7.3915	0.9584/ 0.0829/ 7.2122	0.9384/ 0.1122/ 7.6282
Ours-FT	SAE§	**0.9603/ 0.0907/ 8.0410**	**0.9427/ 0.1137/ 7.5695**	0.9605/ 0.0681/ **7.4140**	0.9414/ **0.0920/ 7.8418**
ResLF [15]	HFUT†	0.9550/ 0.0841/ 7.8709	0.9413/ 0.0998/ 7.2923	0.9452/ 0.1178/ 6.8222	0.9241/ 0.1554/ 7.3478
ATO [3]	HFUT†	0.9600/ 0.0796/ 7.8224	0.9456/ 0.0964/ 7.3113	0.9581/ 0.0696/ 7.0127	0.9356/ 0.1103/ 7.4996
InterNet [12]	HFUT†	0.9611/ 0.0771/ 7.8627	0.9469/ **0.0803**/ 7.4068	0.9613/ 0.0719/ 7.0392	0.9393/ 0.1068/ 7.5428
Ours-Pre	HFUT†	0.9588/ 0.0925/ 7.8064	0.9426/ 0.1011/ 7.3031	0.9565/ 0.0838/ 7.0422	0.9351/ 0.1221/ 7.3894
Ours-FT	HFUT†	**0.9635/ 0.0753/ 8.0575**	**0.9490**/ 0.0849/ **7.6133**	**0.9647/ 0.0543/ 7.3913**	**0.9473/ 0.0768/ 7.8571**

Table 1. SSIM(↑)/ VGG×100(↓)/ Ma's score(↑) comparisons between different SR methods on different light field datasets at the scaling factor of 2 (↑ indicates that higher is better while ↓ indicates that lower is better). The subscript † denotes real-world datasets while the subscript § denotes synthetic datasets. Gray background indicates large domain gap.

2. Angular consistency analysis

As mentioned in the paper, the proposed method can be readily applied to any reference view in the light field other than the central view. To analyze the angular consistency of the SR result, we super-resolve the whole light field of 10 scenes randomly selected from the HCI2 dataset. Since this synthetic dataset contains ground truth depth map for each scene, we are able to evaluate the angular consistency of the SR result by conducting depth estimation on the super-resolved light field. Due to the fact that depth information is highly sensitive to the angular consistency, the difference between the estimated depth map and the ground truth one reflects the SR performance on the other side. Specifically, we adopt a representative non-learning-based light field depth estimation algorithm OCC [11] for the evaluation. VDSR [4] (single-image), RCAN [16] (single-image), BM5D [1] (classic), InterNet [12] (deep-learning, with the HFUT dataset as source), VDSR-ZS (ours with VDSR as the pre-upsampler), and RCAN-ZS (ours with RCAN as the pre-upsampler) are compared in terms of both PSNR across the whole light field and MSE of estimated disparity. As shown in Fig. 1, our VDSR-ZS dominates other methods except RCAN, while our RCAN-ZS dominates RCAN, indicating the superiority of our zero-shot light field SR method in terms of both angular consistency and reconstruction fidelity.

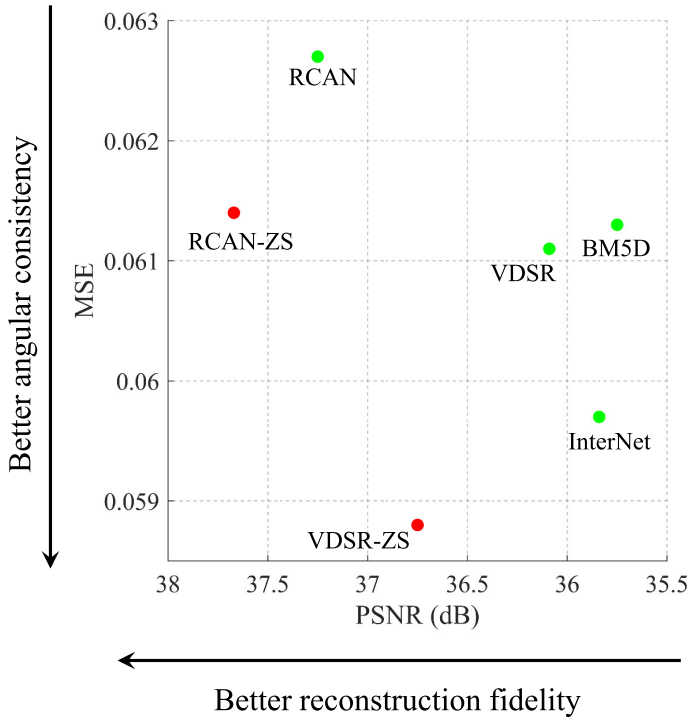

Figure 1. Comparisons on reconstruction fidelity and angular consistency between light fields super-resolved through different methods.

3. More visual results

We show more visual results in Fig. 2.

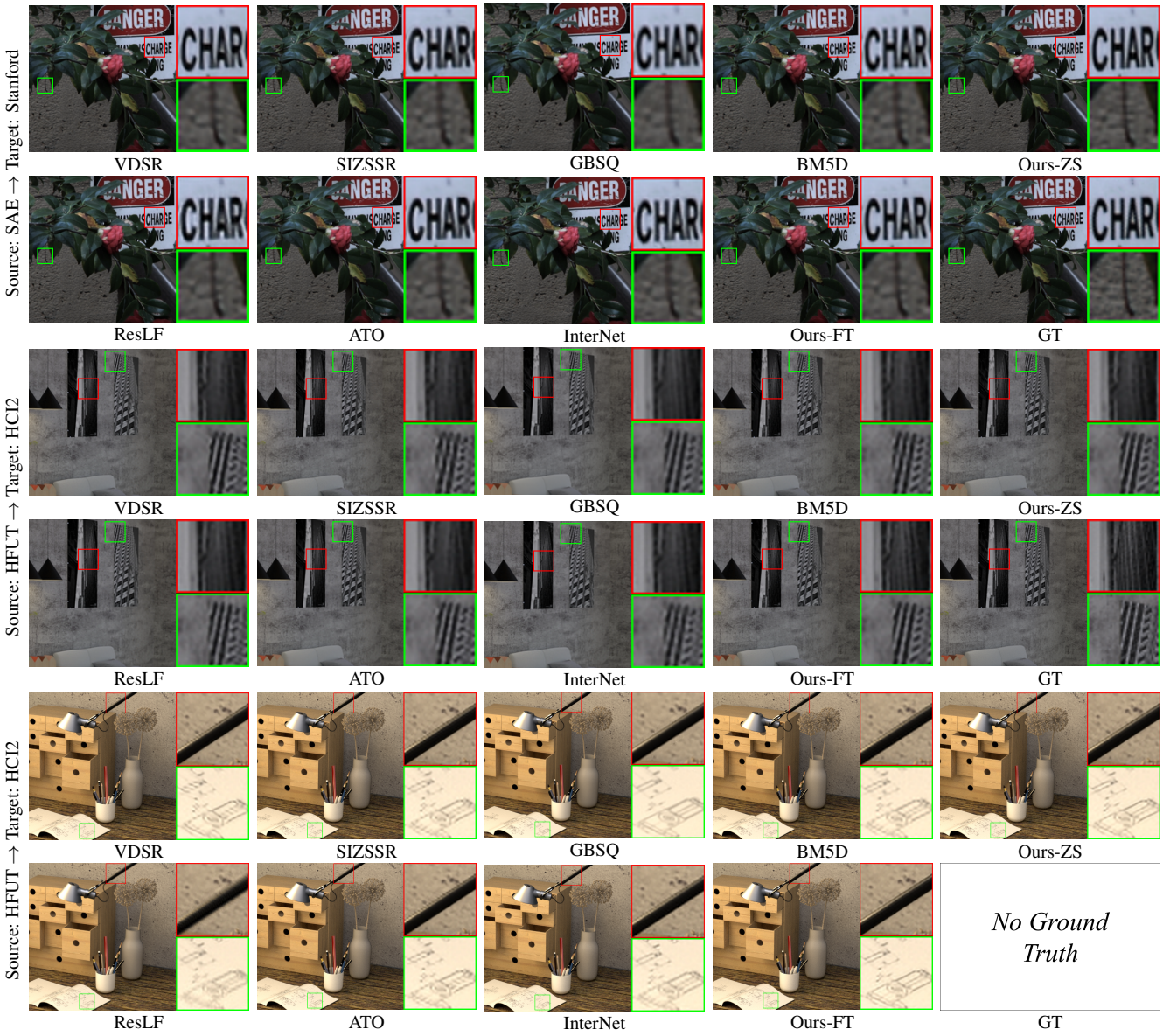

Figure 2. Visual comparisons of super-resolved central view (cropped for better visualization) through different methods together with the ground truth (GT) at the scaling factor of 2. The inputs of the first scene *Reflective12* and the second scene *Bedroom* are the downsampled light fields while the input of the third scene *Table* is the original light field. Zoom in the figure for a better visual experience.

4. Details of Bic-Res, SPconv and SPconv-Res

We adopt different single image SR networks for investigating the divide-and-conquer strategy, the details of which are illustrated in Fig. 3.

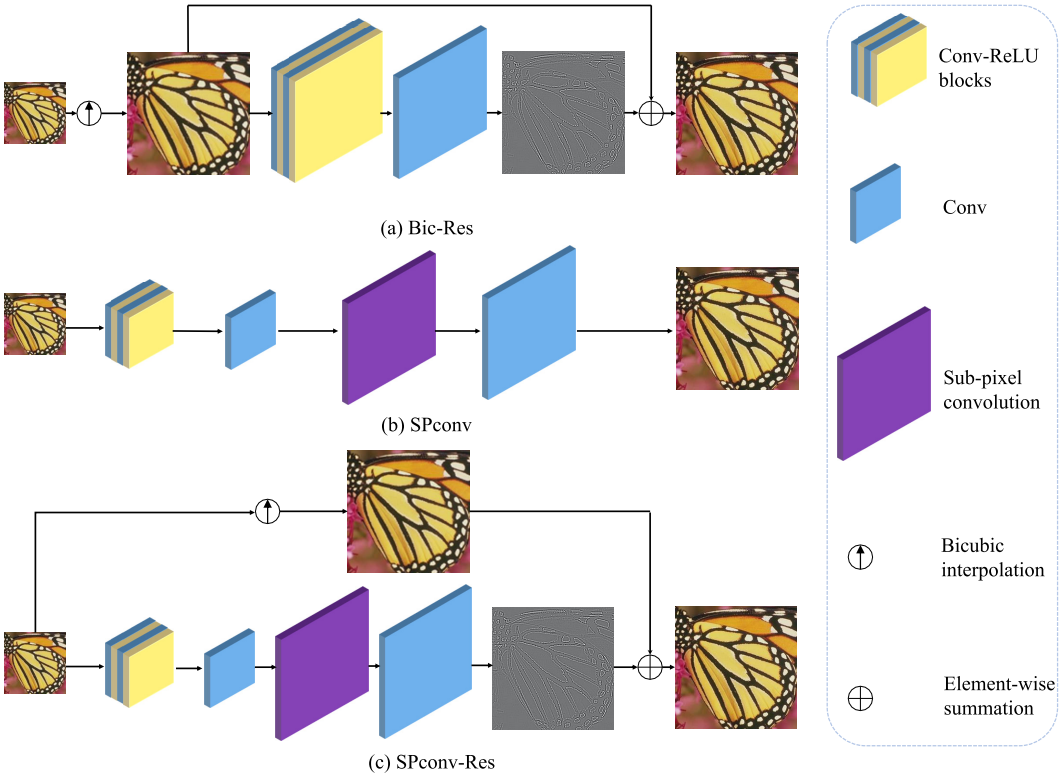

Figure 3. Details of networks used for investigation on single image SR task. (a) illustrates network Bic-Res while (b) and (c) illustrate SPconv and SPconv-Res, respectively. We keep the parameters of these three networks the same during the investigation by adjusting the numbers of Conv-ReLU blocks and channels of each convolution layer.

5. Details of AlignNet and AggreNet

The network details of AlignNet and AggreNet used in our zero-shot light field SR framework are provided in Table 2 and Table 3, respectively.

Layer Type	ReLU	Filter Size	Stride	Zero Padding	Output	Output Size
Input	–	–	–	–	$PSV(Z^{LR})$	$B \times L \times A \times X \times Y$
Level-wise feature extraction						
2-D convolution	✓	1×1	1×1	0×0	–	$B \times L \times 2 \times X \times Y$
reshape	–	–	–	–	–	$B \times 2L \times X \times Y$
Disparity estimation						
2-D convolution	✓	7×7	1×1	3×3	–	$B \times 100 \times X \times Y$
2-D convolution	✓	5×5	1×1	2×2	–	$B \times 100 \times X \times Y$
2-D convolution	✓	3×3	1×1	1×1	–	$B \times 50 \times X \times Y$
2-D convolution	–	1×1	1×1	0×0	d^{LR}	$B \times 1 \times X \times Y$

Table 2. Network details of AlignNet. For an input low-resolution (LR) light field $Z^{LR} \in \mathbb{R}^{U \times V \times X \times Y}$, given a disparity level number L for PSV generator, the output of AlignNet is an LR disparity map d^{LR}. Note that $A = U \times V$ and B denotes the batch size.

Layer Type	ReLU	Filter Size	Stride	Zero Padding	Output	Output Size
Input	–	–	–	–	W^{LR}	$B \times A \times \alpha X \times \alpha Y$
Residual prediction						
2-D convolution	✓	7×7	1×1	3×3	–	$B \times 32 \times \alpha X \times \alpha Y$
2-D convolution	✓	7×7	1×1	3×3	–	$B \times 64 \times \alpha X \times \alpha Y$
2-D convolution	✓	7×7	1×1	3×3	–	$B \times 32 \times \alpha X \times \alpha Y$
2-D convolution	✓	7×7	1×1	3×3	–	$B \times 16 \times \alpha X \times \alpha Y$
2-D convolution	–	7×7	1×1	3×3	$AggreNet_{\Theta_2}(W^{LR})$	$B \times 1 \times \alpha X \times \alpha Y$

Table 3. Network details of AggreNet. For an aligned pre-upsampled LR light field $W^{LR} \in \mathbb{R}^{U \times V \times \alpha X \times \alpha Y}$, the output of AggreNet is an HR residual map between the pre-upsampled LR central view and the ideal HR central view. Note that $A = U \times V$ and B denotes the batch size.

6. Implementation details

To make full use of the training data, we randomly crop patches with size 64×64 for scale 2 and size 72×72 for scale 3 and exploit 4D flipping and rotation proposed in [9] for data augmentation.

During training, we set the weighting factors γ_1 and γ_2 as 0.5 and 0.1 empirically. The batch size is set as 1. We use the ADAM optimizer [5] with $\alpha = 0.9$ and $\beta = 0.999$. For the training from scratch (Ours-ZS), the initial learning rate for each stage is 10^{-4} and for the training of finetune (Ours-FT), the initial learning rate for each stage is 10^{-5}.

After training, we extend the geometric self-ensemble [6] to 4D light field with 4D flipping and rotation. The iterative back-projection refinement [2] used in ZSSR [10] is also adopted for post-processing. We implement our method using Pytorch and run all experiments on a GTX 1080Ti GPU.

References

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

[2] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, 53(3):231–239, 1991. 6

[3] 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

[4] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 1, 2

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[6] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 6

[7] 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. 1

[8] Mattia Rossi and Pascal Frossard. Graph-based light field super-resolution. In *MMSP*, 2017. 1

[9] 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. 6

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

[11] Ting-Chun Wang, Alexei A Efros, and Ravi Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *ICCV*, 2015. 2

[12] 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

[13] 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. 1

[14] 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. 1

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

[16] 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. 2