Efficient Light Field Image Super-Resolution via Progressive Disentangling

Gaosheng Liu, Huanjing Yue, Jingyu Yang*
School of Electrical and Information Engineering, Tianjin University
{gaoshengliu, huanjing.yue, yjy}@tju.edu.cn

Abstract

The performance of light field (LF) image super-resolution (SR) has been significantly improved with the development of deep learning techniques. In recent state-of-the-art methods, increasingly deeper and wider networks with a massive number of layers are employed to improve SR performance. However, these approaches often incur heavy computational costs, hindering efficient inference and practical applications. In this paper, we address the problem by introducing an efficient network for LF image SR. Specifically, we propose an efficient progressive disentangling block (PDistgB), where the intermediate LF feature is progressively channel-wise split and selectively domain-specific disentangled. The PDistgB can well incorporate the LF structure prior while requiring fewer computational costs compared with existing disentangling strategies. In addition, we apply Transformer on the angular domain to incorporate angular correlations from all views for further improving the SR accuracy. Experimental results on public datasets demonstrate that our method achieves state-of-the-art performance with high efficiency. Codes and models are available at https://github.com/GaoshengLiu/PDistgNet.

1. Introduction

Light field (LF) photography captures not only the intensities of light but also the directions of the rays across a scene. The additional dimension offers valuable 3D cues, facilitating various downstream applications such as depth prediction [39], digital post-focusing [6], and 3D observation [26]. However, as the 4D LF structure is multiplexed on the 2D image sensor, existing LF imaging devices encounter difficulties in achieving dense sampling in the angular domain while maintaining satisfactory spatial size. In recent years, leveraging computational imaging techniques to address the challenges has garnered significant attention [15, 23, 24, 40, 47, 50].

*Corresponding author. This work was supported in part by the National Natural Science Foundation of China under Grant 62231018 and Grant 62072331.

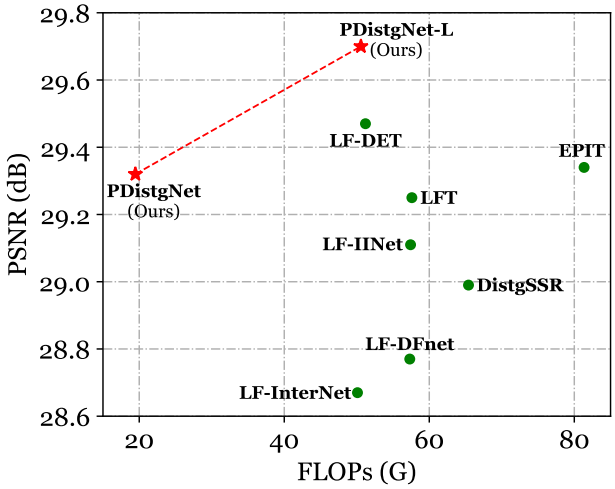

Figure 1. FLOPs *vs.* PSNR of different methods. The PSNR results are achieved on EPFL [29] for 4× LF image SR. The FLOPs are calculated with a $5 \times 5 \times 32 \times 32$ input.

In this work, we focus on enhancing the spatial resolution of LF images. To tackle this problem, various approaches [5, 12, 40, 44, 47, 50] have been introduced. The conventional methods [2, 27, 44] formulate LF image super-resolution (SR) as an optimization problem. Recently, the learning-based methods have dominated the area of LF image SR [12, 40, 47, 50]. These methods capitalize on different perspectives of the 4D LF structure to incorporate the complementary angular correlations for improving the SR performance. Among them, the Transformer-based methods [4, 18, 19] are leading the advancements. In addition, there has also been a trend toward designing deeper and wider networks [13, 31] to improve SR accuracy. However, despite the remarkable performance achieved by these approaches, they often require significant computational costs, limiting their flexibility and practical applications. For example, the floating-point operations (FLOPs) of the winner approach [13] in the NTIRE 2023 LF image SR challenge [41] reached up to 397.2 G measured using a

$5 \times 5 \times 32 \times 32$ input. Moreover, testing the entire LF image with resolutions like $5 \times 5 \times 374 \times 540$ demands even more substantial computing resources and inference time. Therefore, there is a pressing need for more efficient solutions in LF image SR.

In this paper, we present a simple yet efficient method to mitigate the obstacle. Specifically, we introduce an efficient progressive disentangling block (PDistgB) to leverage the inherent LF structure priors. In PDistgB, the LF feature is progressively split along the channel dimension into two independent parts. One part undergoes specific subspace disentanglement, while the other part is further split for subsequent subspace disentanglement. After multiple splits and disentanglement, we concatenate the disentangled features and fuse them to enhance the feature representation. Additionally, considering that the macro-pixel patterns (*i.e.*, angular domain) reflect the angular correlations via recording the directions of light rays, and given that the angular resolution is typically small, *e.g.*, 5×5, we also deploy the Transformer on angular domain to incorporate complementary information from all angular views while conserving GPU memory.

Based upon PDistgB and angular Transformer block (AngTB), we develop PDistgNet for efficient LF image SR. Our method achieves superior performance while requiring significantly lower computational costs. As shown in Figure 1, our PDistgNet and PDistgNet-L (the large version of PDistgNet) achieve better PSNR and FLOPs trade-off compared with state-of-the-art methods.

In summary, the contributions of this paper are as follows. **(1)** We address the problem of efficient LF image SR with a simple yet efficient network, PDistgNet. **(2)** We propose an efficient progressive disentangling block (PDistgB) to leverage the LF structure prior by progressively performing multiple subspace-specific disentanglement. **(3)** Extensive experimental results on public datasets demonstrate the superiority of our method over the state-of-the-art in terms of both SR accuracy and efficiency.

2. Related Works

2.1. Efficient Image SR

Image SR is a long-standing problem in image restoration. The efficient image SR technique aims at reducing the computational cost of image SR methods for fast inference. To achieve this, a variety of strategies have been proposed. For example, Aha *et al*. [1] proposed a cascaded residual network with point-wise and depth-wise convolutions, which largely reduced the parameters compared with the original residual block [7]. Hui *et al*. [10] proposed a feature distillation strategy, which splits the intermediate feature along the channel dimension and applies convolutions on these separate features to reduce the number of parameters. In their later study [11], an information multi-distillation network (IMDN) is introduced, in which the intermediate feature is progressively split and processed with convolutions. Liu *et al*. [25] proposed a residual feature distillation block with 3×3 and 1×1 convolutions. Wang *et al*. [33, 34] explored the sparsity in image SR to improve the inference efficiency of SR networks.

2.2. LF Image SR

LF image SR aims at enhancing the resolution of each view of an LF and simultaneously preserving the angular consistency. Yoon *et al*. [48] stacked adjacent views and applied convolutions to leverage the complementary information. Yeung *et al*. [47] utilized alternating 2D convolution on spatial and angular domains to replace the 4D convolution, which achieves better performance and has lower computational costs. Wang *et al*. [36] introduced a bidirectional recurrent network to propagate the axial views in a recurrent manner, where angular correlations are implicitly incorporated. Instead of considering only the axial views, Zhang *et al*. [50, 51] proposed to model the multi-directional epipolar correlations with a multi-branch network. Jin *et al*. [12] developed an all-to-one method for LF image SR and performed structural consistency regularization to preserve the LF parallax structure. Wang *et al*. [37] decoupled the LF image into spatial and angular features and designed a feature interaction strategy to incorporate the spatial-angular correlations. In their extended study, spatial-angular decoupling is advanced to a disentangling mechanism [40] to further leverage the structure prior for improving the SR performance. Wang *et al*. [38] addressed the disparity variations by using deformable convolution [53] for view alignment. Liu *et al*. [22] proposed an intra- and inter-view interaction strategy to incorporate the angular correlations. Later, they further proposed to explicitly utilize the disparity information to guide the SR process [21]. Mo *et al*. [28] introduced a dense dual-attention network to incorporate long-term information from shallow to deep layers. Apart from these CNN-based methods, the Transformer techniques work on spatial-angular domains [4, 19] and epipolar-plane domain [19, 35] have been exploited, leading the recent advancements. Very recently, super-resolving the LF images under real-world degradation [43, 46] has also been explored.

However, increasing computational costs of recent methods limit their practical applications. In this paper, we aim to explore efficient solutions for LF image SR.

3. Method

3.1. Preliminary

The LF image can be represented as a 4D function $L(u, v, h, w) \in \mathbb{R}^{U \times V \times H \times W}$, where $U \times V$ and $H \times W$

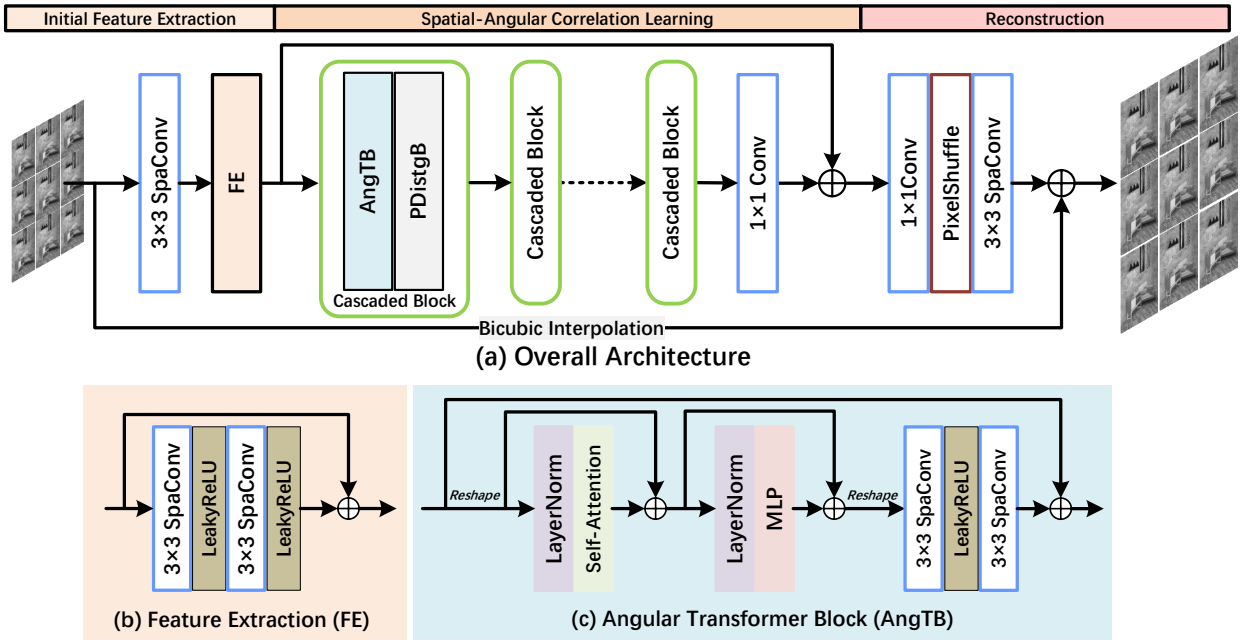

Figure 2. The framework of our proposed method. A 3×3 LF image is shown as an example. The SpaConv indicates the convolution works on the spatial, *i.e.*, $H \times W$ domain and shares weights for different views. To be consistent with previous works, our method performs on only the Y-channel component of LF images.

represent angular and spatial resolutions. The (u, v) and (h, w) denote angular and spatial locations. This work aims at enhancing the spatial resolution of L and reconstructing $L_{\text{SR}}(u, v, h, w) \in \mathbb{R}^{U \times V \times rH \times rW}$, where r is the upsampling factor.

3.2. Network Architecture

The architecture of our proposed method is depicted in Figure 2. Concretely, our method contains three steps, *i.e.*, initial feature extraction, spatial-angular correlation learning, and reconstruction. The initial feature extraction consists of a 1×1 convolution and a feature extraction (FE) module to extract the intra-view correlations. Then we introduce cascaded block, which is composed of angular transformer block (AngTB) and progressive disentangling block (PDistgB), to leverage the inherent spatial-angular correlations of LF. After N cascaded blocks and a skip connection, we up-sample the spatial resolution of LF feature to generate high-resolution (HR) LF image. A global residual connection with bicubic interpolation is deployed to feed low-frequency information to the output.

3.2.1 Initial Feature Extraction

The initial feature extraction aims at mapping input low-resolution (LR) LF image to the feature domain and exploring spatial correlations. To achieve this, a 1×1 convolution is first applied for channel expansion. Then two 3×3 convolutions with LeakyReLU activation are formed as an FE module, generating LF feature $\mathcal{F} \in \mathbb{R}^{U \times V \times C \times H \times W}$, where C is the channel dimension of $\mathcal{F}$.

3.2.2 Spatial-Angular Correlation Learning

To leverage spatial and angular correlations for enhancing SR accuracy, we deploy cascaded AngTBs and PDistgBs for deep spatial and angular correlation learning.

Angular Transformer Block. The macro-pixel pattern of LF image is a set of pixels with the same point but captured at different angular locations, in which angular correlations are reflected. Meanwhile, the macro-pixel pattern usually has a relatively small size, *e.g.*, 5×5. Therefore, applying Transformer on macro-pixel patterns is an efficient choice for incorporating the angular correlations. In this work, we extend the basic framework in [18] to formulate our AngTB, whose structure is depicted in Figure 2 (c). Specifically, the input feature $\mathcal{F} \in \mathbb{R}^{U \times V \times C \times H \times W}$ is converted into a sequence of "angular token", $T^i \in \mathbb{R}^{UV \times C}, i \in \{1, \cdots, HW\}$. In practice, the T^i are stacked as $T \in \mathbb{R}^{HW \times UV \times C}$, and HW is regarded as the batch dimension. We adopt the sinusoidal position encoding [18] to encode the angular locations to P. Then the *query* Q and

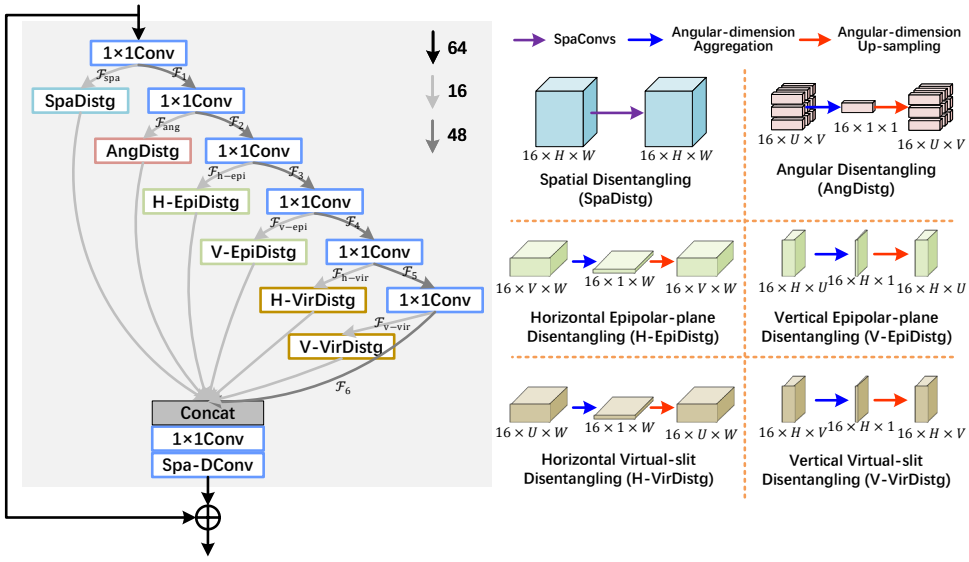

Figure 3. The architecture of PDistgB. The 64, 48, and 16 all represent the output channels of the convolution layer. Each 1×1 convolution is followed by a LeakyReLU activation (which is omitted for simplicity). The Spa-DConv is the depth-wise convolution that works on the spatial domain. The angular dimension indicates U and/or V.

key K is generated by performing layer normalization (LN) on $(T + P)$, while the *value* V is assigned as T. Then the Q, K, and V are linearly projected to a embedding dimension of D by multiplying projection matrix, $W_Q \in \mathbb{R}^{C\times D}$, $W_K \in \mathbb{R}^{C\times D}$ and $W_V \in \mathbb{R}^{C\times D}$, respectively. We follow the setting of multi-head self-attention (MHSA) [32] to evenly divide the Q, K, and V into M segments along the D dimension, and performing self-attention on corresponding segment, Q_j, K_j, and V_j, $j \in \{1, \cdots, M\}$, which is given by:

$$\hat{T}_j = \text{Softmax}(\frac{Q_j \otimes K_j^{\text{T}}}{\sqrt{d}}) \otimes V_j, j \in \{1, \cdots, M\}, \quad (1)$$

where $\otimes$ denotes matrix multiplication, and $d = D/M$. Then the $\hat{T}_j \in \mathbb{R}^{HW\times UV\times d}, j \in \{1, \cdots, N\}$ are concatenated to generate $\hat{T} \in \mathbb{R}^{HW\times UV\times D}$.

Following previous Transformer [32], $\hat{T}$ is added with input tokens T and sequentially processed by a LN and a multi-layer perception (MLP), which is given by:

$$\tilde{T} = \text{MLP}(\text{LN}(\hat{T} + T)) + (\hat{T} + T). \quad (2)$$

After that, the angular tokens $\tilde{T}$ are reshaped back to LF features $\tilde{\mathcal{F}} \in \mathbb{R}^{UV\times C\times H\times W}$ and processed by two spatial-wise 3×3 convolutions. Finally, the residual learning is applied to generate the output of AngTB.

Progressive Disentangling Block. The disentangling mechanism is designed to incorporate the LF structure prior by performing domain-specific disentanglement, which has shown its effectiveness in both LF SR and depth prediction [40, 43]. In previous works [14, 40, 43], the input LF feature is directly separately subspace-specific disentangled. Then the disentangled outputs are concatenated and fused using convolutions. However, the separate disentangling requires large computational costs. To mitigate the obstacle, we introduce a more efficient progressive disentangling strategy to reduce the computational costs.

The structure of our PDistgB is depicted in Figure 3. The input feature is first processed by a 1×1 convolution, whose output is channel-wise split, obtaining two components, *i.e.*, $\mathcal{F}_{\text{spa}} \in \mathbb{R}^{U\times V\times C_1\times H\times W}$ and $\mathcal{F}_1 \in \mathbb{R}^{U\times V\times C_2\times H\times W}$. The $\mathcal{F}_{\text{spa}}$ is sent for spatial disentangling using spatial-wise convolutions. $\mathcal{F}_1$ is preserved and processed by a 1×1 convolution to expand the channel numbers, whose output is further divided into $\mathcal{F}_{\text{ang}} \in \mathbb{R}^{U\times V\times C_1\times H\times W}$ and $\mathcal{F}_2 \in \mathbb{R}^{U\times V\times C_2\times H\times W}$. $\mathcal{F}_{\text{ang}}$ is reshaped into macro-pixel pattern features with resolution of $HW \times C_1 \times U \times V$ and sent for angular disentangling. We follow previous work [40] to adopt an angular-domain *aggregation&up-sampling* strategy for angular disentangling. Concretely, we first apply convolution to aggregate each macro-pixel pattern feature from the resolution of $C_1 \times U \times V$ to that of $C_1 \times 1 \times 1$ and then utilize the *PixelShuffle* layer to up-sample it to the resolution of $C_1 \times U \times V$. Similarly, $\mathcal{F}_2$ is fed into the following layers for channel splits and epipolar-plane disentanglement. For the split epipolar-plane feature, $\mathcal{F}_{\text{h-epi}} \in \mathbb{R}^{UH\times C_1\times V\times W}$ and $\mathcal{F}_{\text{v-epi}} \in \mathbb{R}^{VW\times C_1\times U\times H}$ (after reshaping), the disentanglement includes an *aggregation* on V and U dimension, respectively, and a corresponding up-sampling operation, as shown on the right side of Figure 3. Recently, one more 2D representation of LF has

been investigated [9, 14], namely virtual-slit image representation [14], which also reflects the sub-pixel correlations. Therefore, we also perform disentanglement in virtual-slit image representation. Similar to epipolar-plane disentanglement, for virtual-slit feature, $\mathcal{F}_{\text{h-vir}} \in \mathbb{R}^{VH \times C_1 \times U \times W}$ and $\mathcal{F}_{\text{v-vir}} \in \mathbb{R}^{UW \times C_1 \times V \times H}$ (after reshaping), the *aggregation* is performed on U and V dimension, respectively. We set the channel numbers of disentangled feature C_1 to 16 and preserved feature C_2 to 48.

After the progressive feature splits and disentanglement, the disentangled outputs and the preserved split feature are concatenated and fused via a point-wise (1×1) and a depth-wise (on the spatial domain) convolution. We also deploy residual learning in PDistgB.

3.2.3 Reconstruction

To reconstruct the high-resolution LF image, we utilize a 1×1 convolution to expand the channel dimension and then adopt the *PixelShuffle* layer to rearrange the pixels from channel dimension to spatial dimension. Finally, a 3×3 spatial-wise convolution is applied to reduce the channel numbers and generate the final output. To train our network, we use the ℓ_1 loss function to constrain the network output to be similar to ground truth.

4. Experiments

4.1. Implementation Details

To be consistent with previous studies, we use five public LF datasets, including three real-world datasets (EPFL [29], STFgantry [30], and INRIA [17]) and two synthetic datasets (HCInew [8] and HCIold [45]) in the experiments. Following *BasicLFSR* [41], a total number of 144 LF images from the five datasets are used for training and 23 LF images for testing. For experiments, we extract the central 5×5 views of each scene and cropped them into patches with a spatial resolution 128×128. Then we apply $4\times$ bicubic downsampling to generate the LR inputs.

In our PDistgNet, the number of cascaded blocks, *i.e.*, N is set to four. We also provide a large version of our method, PDistgNet-L, by deploying 12 cascaded blocks. The head number M in AngTB is set to eight. For training, the learning rate is initially set to $2e-4$ and weighted decayed by a factor of 0.5 for every 15 epochs. The total training epoch is set to 80. To evaluate the quantitative performance of each scene, we calculate the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) metrics on each view and then average the scores of all views. The final score on each test set is by averaging the scores of all the scenes in this test set.

Table 1. PSNR results by deploying different subspace disentangling in PDistgB. Note that, the number of parameters (#Prm.) of different variants are adjusted.

Spa	Ang	Epi	Vir	#Prm.	EPFL	HCInew	HCIold	INRIA
✓	✓			0.86M	29.24	31.30	37.53	31.18
✓	✓	✓		0.84M	29.21	31.33	37.50	31.13
✓	✓	✓	✓	0.83M	31.32	31.33	37.55	31.24

Table 2. The number of parameters (#Prm.) and FLOPs (computed with input size of $5\times5\times32\times32$) comparison between our PDistgB and previous disentangling strategies.

	PDistgB	DistgB [40]	C_4^2FE [14]
#Prm. (M)	0.06	0.21	0.27
FLOPs (G)	1.06	3.68	4.36

4.2. Ablation Investigation

Progressive Disentangling Block. In our method, the PDistgB is designed with a small number of parameters but is effective for high performance. To verify this, we conduct several experiments. Firstly, we study the influence of different numbers of subspace disentangling. Specifically, we sequentially integrate angular, epipolar-plane, and virtual-slit disentangling on a baseline with spatial disentangling. As listed in Table 1, the performance achieved by integrating all subspace disentangling is consistently better on different datasets. Secondly, we compare our PDistgB with disentangling block (DistgB) [40] and C_4^2 feature extractor (C_4^2FE) [14]. A computational cost comparison is listed in Table 2. We then design two variants by separately replacing the PDistgBs in our PDistgNet with DistgBs and C_4^2FEs, termed as *w_DistgB* and *w_*C_4^2*FE*, respectively. For a fair comparison, we reduce the number of parameters in these two variants. As listed in Table 3, the performance of *w_DistgB* suffers from a degradation of 0.19 dB on INRIA [17]. The *w_*C_4^2*FE* is also inferior to our method by 0.15 dB on EPFL [29]. Thirdly, we directly remove the PDistgBs in our method, termed as *w/o_PDistgB*. From Table 3, the performance of *w/o_PDistgB* is degraded by 0.17 dB on HCIold [45].

In summary, these experimental results demonstrate the high effectiveness of our proposed PDistgB.

Angular Transformer Block. In our method, we deploy Transformer on the angular domain and introduce AngTB to effectively leverage the angular correlations from all views. To prove its effectiveness, we remove the AngTBs in PDistgNet, termed as *w/o_AngTB*. As listed in Table 3, this variant also suffers from a drop of 0.42 dB on HCIold [45].

4.3. Comparisons with State-of-the-art Methods

We compare our method (two versions) with state-of-the-art approaches, including four single image SR meth-

Table 3. Ablation study results for PDistgB and AngTB. For PSNR/SSIM, larger is better. The FLOPs are computed with input size of $5 \times 5 \times 32 \times 32$. Note that, the number of parameters (#Prm.) of different variants are adjusted.

Variant	#Prm. (M)	FLOPs (G)	Datasets				
			EPFL	HCInew	HCIold	INRIA	STFgantry
w_DistgB	0.85	18.37	29.15/0.9189	31.29/0.9199	37.58/0.9731	31.05/0.9513	31.40/0.9512
w_C_4^2FE	0.93	18.48	29.17/0.9191	31.30/0.9200	37.56/0.9730	31.18/0.9517	31.42/0.9513
w/o_AngTB	0.88	16.39	28.87/0.9145	31.04/0.9166	37.13/0.9708	30.99/0.9495	30.76/0.9452
w/o_PDistgB	0.83	21.87	29.14/0.9178	31.21/0.9185	37.38/0.9723	31.22/0.9508	31.34/0.9501
PDistgNet (**Ours**)	0.83	19.47	29.32/0.9195	31.33/0.9205	37.55/0.9730	31.24/0.9517	31.43/0.9514

Table 4. Comparison of PSNR/SSIM with different methods. The FLOPs are computed with input size of $5 \times 5 \times 32 \times 32$. The inference time is measured by super-resolving a $5 \times 5 \times 192 \times 192$ LR input. The best results are highlighted in red and the second results in blue.

Method	#Prm. (M)	FLOPs (G)	Time (s)	Datasets				
				EPFL	HCInew	HCIold	INRIA	STFgantry
Bicubic	-	-	-	25.14/0.8311	27.61/0.8507	32.42/0.9335	26.82/0.8860	25.93/0.8431
VDSR [16]	0.66	272.26	3.98	27.25/0.8777	29.31/0.8823	34.81/0.9515	29.19/0.9204	28.51/0.9009
EDSR [20]	38.89	1016	11.44	27.84/0.8858	29.60/0.8874	35.18/0.9538	29.66/0.9259	28.70/0.9075
RCAN [52]	15.36	390.12	7.52	27.88/0.8863	29.63/0.8886	35.20/0.9548	29.76/0.9276	28.90/0.9131
resLF [50]	6.79	39.70	2.32	27.46/0.8899	29.92/0.9011	36.12/0.9651	29.64/0.9339	28.99/0.9214
LFSSR [47]	1.61	128.24	5.11	28.27/0.9080	30.72/0.9124	36.70/0.9690	30.31/0.9446	30.15/0.9385
LF-ATO [12]	1.36	1898	16.84	28.64/0.9130	30.97/0.9150	37.06/0.9703	30.79/0.9490	30.79/0.9448
LF-InterNet [37]	5.23	50.10	2.11	28.67/0.9143	30.98/0.9165	37.11/ 0.9715	30.64/0.9486	30.53/0.9426
LF-DFnet [38]	3.99	57.31	2.54	28.77/0.9165	31.23/0.9196	37.32/0.9718	30.83/0.9503	31.15/0.9494
MEG-Net [51]	1.77	102.18	2.49	28.74/0.9160	31.10/0.9177	37.28/0.9716	30.66/0.9490	30.77/0.9453
LF-IINet [22]	4.88	57.42	1.55	29.11/0.9194	31.36/0.9211	37.62/0.9737	31.08/0.9516	31.21/0.9494
DPT [35]	3.78	66.55	11.96	28.93/0.9170	31.19/0.9188	37.39/0.9721	30.96/0.9503	31.14/0.9488
LF-DGNet [21]	4.70	50.20	1.16	29.06/0.9191	31.39/0.9215	37.48/0.9728	31.13/0.9521	31.35/0.9516
DistgSSR [40]	3.58	65.41	2.52	28.99/0.9195	31.38/0.9217	37.56/0.9732	30.99/0.9519	31.65/0.9535
PDistgNet (**Ours**)	0.83	19.47	2.31	29.32/0.9195	31.33/0.9205	37.55/0.9730	31.24/0.9516	31.43/0.9514
LFT [18]	1.16	57.60	11.73	29.25/0.9210	31.46/0.9218	37.63/0.9735	31.20/0.9524	31.86/0.9548
LF-SAV [3]	1.54	115.80	6.96	29.37/0.9223	31.45/0.9217	37.50/0.9721	31.27/0.9531	31.36/0.9505
HLFSR [31]	13.87	182.52	11.81	29.20/0.9222	31.57/0.9238	37.78/0.9742	31.24/0.9534	31.64/0.9537
EPIT [19]	1.47	81.35	4.98	29.34/0.9197	31.51/0.9231	37.68/0.9737	31.37/0.9526	32.18/0.9571
LF-DET [4]	1.69	51.20	7.82	29.47/0.9230	31.55/0.9235	37.84/0.9744	31.39/0.9534	32.14/0.9573
PDistgNet-L (**Ours**)	2.19	50.57	7.34	29.70/0.9245	31.61/0.9241	37.82/0.9744	31.54/0.9542	32.19/0.9578

ods (*i.e.*, bicubic interpolation, VDSR [16], EDSR [20], and RCAN [52]) and 15 LF image SR methods (*i.e.*, resLF [50], LFSSR [47], LF-ATO [12], LF-InterNet [37], LF-DFnet [38], MEG-Net [51], LF-IINet [22], DPT [35], LF-DGNet [21], DistgSSR [40], LFT [18], LF-SAV [3], HLFSR [31], EPIT [19], and LF-DET [4]). These learning-based methods are trained using the same datasets as ours. We conduct the comparison on $4\times$ LF image SR.

Quantitative Results. To be consistent with the *BasicLFSR* [41], we report the results of all methods (including ours) by cropping input LR LF images into patches with zero padding for testing. The results are listed in Table 4, from which we can observe that:

- Recent state-of-the-art methods, such as LF-SAV, HLFSR, and EPIT achieve much higher performance than early resLF and LF-InterNet but require much higher computational costs, *e.g.*, FLOPs.
- Our PDistgNet achieves higher PSNR scores than DistgSSR on EPFL [29] and INRIA [17] with 76% parameters (#Prm.) and 70% FLOPs reduction, which demon-

Figure 4. Visual Comparison on different methods for 4× LF image SR. The enlarged patches cut from the green box of central view and epipolar-plane image cut along the blue line are presented for comparison.

strates the high efficiency of our PDistgNet.

- The larger version of our method, PDistgNet-L, achieves the highest PSNR scores on EPFL [29], HCInew [8], INRIA [17], and STFgantry [30] with an acceptable computational cost. Specifically, PDistgNet-L outperforms EPIT and LD-DET by 0.17 dB and 0.15 dB on INRIA [17], respectively.

In Figure 1, we visualize the trade-off between FLOPs and PSNR scores on EPFL [29]. It can be observed that our PDistgNet and PDistgNet-L achieve higher efficiency compared with state-of-the-art methods.

Inference Time Evaluation. We also report the inference time of different methods to further demonstrate the efficiency. The results are listed in Table 4. The inferences of different methods are conducted on the same desktop with an NVIDIA RTX 3090 GPU and Xeon Platinum 8369B CPU @2.90 GHz. We report the average time over five runs. It can be observed that our PDistgNet is faster than DistgSSR, MEG-Net, and LF-DFnet. Our PDistgNet-L is also faster than LF-DET, and LFT.

Visual Results. The visual comparison results are presented in Figure 4. We provide the results for 4× LF im-

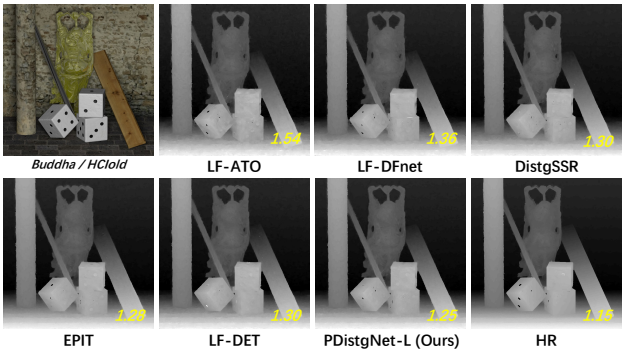

Figure 5. Depth prediction results using SPO [49]. The MSE (marked in yellow on each depth map) between the predicted depth map and ground truth is utilized for objective comparison.

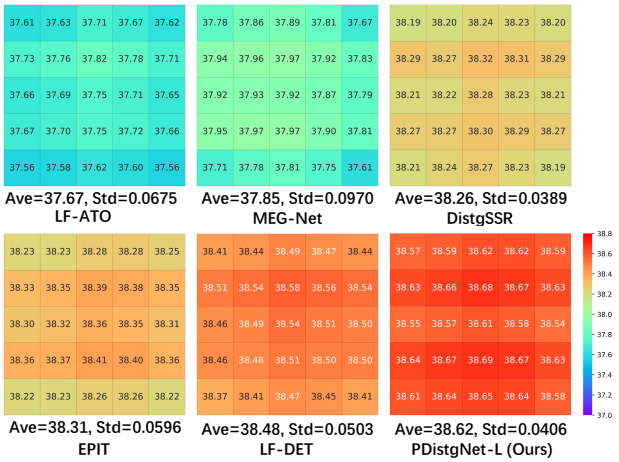

Figure 6. PSNR distribution among different SAIs achieved by different methods on scene *MonasRoom* from HCIold [45]. The **Ave** denotes the average PSNR over 25 views and **Std** is the standard deviation to show the uniformity.

age SR on Lytro-captured scene from EPFL [29], moving-gantry captured scene from STFgantry [30], and synthetic scene from HCInew [8]. It can be observed that our PDistgNet-L reconstructs more visually pleasing results with fine-granular details and clear edges.

Angular Consistency Evaluation. To evaluate the angular consistency of our super-resolved results, we provide the epipolar-plane image (EPI), as shown under each enlarged patch in Figure 4. We can observe that EPIs generated by our method show more linear structures and fewer artifacts. Furthermore, we perform depth map prediction using the reconstructed HR LF images by different methods. Following previous methods [19, 38, 40], we adopt the SPO [49] for depth estimation and mean-square error (MSE) for quantitative comparison. As reported in Figure 5, the MSE achieved by our results is lower than that of other methods, which suggests the high angular consistency of our results.

Perspective Evaluation. To further investigate the reconstruction quality with respect to different perspectives, we follow previous methods [22, 37, 38] to report the PSNR distribution of the reconstructed 5×5 views. As visualized in Figure 6, our method can achieve a relatively balanced distribution and high reconstruction quality compared with state-of-the-art methods.

4.4. NTIRE 2024 LF Image SR Challenge

We participated in the NTIRE 2024 LF image SR challenge [42] with the proposed approach. This challenge adopts the developed validation and test sets from [41] and has two tracks. For Track 2 (Fidelity & Efficiency), we submitted the results achieved by our PDistgNet and we conducted full-resolution testing to avoid the zero-padding in overlapped patch-based testing [13]. For Track 1 (Fidelity only), we increased the number of cascaded blocks in our framework to 16, termed as PDistgNet-L*, and further adopted the ensemble strategy [20] for higher performance. The results achieved by our methods and three baseline methods are listed in Table 5. Our PDistgNet-L* and PDistgNet ranked 4th and 3rd place in Track 1 (Fidelity only) and Track 2 (Fidelity & Efficiency), respectively.

Table 5. PSNR and SSIM scores on validation and test sets of NTIRE 2024 LF image SR Challenge.

Method	#Prm.	FLOPs	Validation	Test
LF-InterNet [37]	5.23	50.10	31.33/0.9381	29.45/0.9198
DistgSSR [40]	3.58	65.41	31.75/0.9424	29.64/0.9244
EPIT [19]	1.47	81.35	32.04/0.9447	29.87/0.9259
PDistgNet	0.83	19.47	31.72/0.9411	29.96/0.9238
PDistgNet-L*	2.87	66.13	32.25/0.9462	30.44/0.9288

5. Conclusion and Future Work

In this paper, we address the problem of efficient LF image SR. Specifically, we proposed a progressive disentangling block, in which the LF feature is sequentially channel-wise split and selectively domain-specific disentangling, to leverage the structure prior. We also applied Transformer on the angular domain to incorporate the angular correlations from all views. Extensive experimental results demonstrated that our method achieves state-of-the-art performance and high efficiency. The EPIs and predicted depth map further demonstrate the high angular consistency of our reconstructed results.

In future work, we plan to explore the sparsity [34] in the angular domain during angular correlations modeling for efficient inference.

References

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*, pages 252–268, 2018. 2

[2] Tom E Bishop and Paolo Favaro. The light field camera: Extended depth of field, aliasing, and superresolution. *IEEE transactions on pattern analysis and machine intelligence*, 34(5):972–986, 2011. 1

[3] Zhen Cheng, Yutong Liu, and Zhiwei Xiong. Spatial-angular versatile convolution for light field reconstruction. *IEEE Transactions on Computational Imaging*, 8:1131–1144, 2022. 6

[4] Ruixuan Cong, Hao Sheng, Da Yang, Zhenglong Cui, and Rongshan Chen. Exploiting spatial and angular correlations with deep efficient transformers for light field image super-resolution. *IEEE Transactions on Multimedia*, 2023. 1, 2, 6

[5] Reuben A Farrugia, Christian Galea, and Christine Guillemot. Super resolution of light field images using linear subspace projection of patch-volumes. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):1058–1071, 2017. 1

[6] Juliet Fiss, Brian Curless, and Richard Szeliski. Refocusing plenoptic images using depth-adaptive splatting. In *IEEE international conference on computational photography (ICCP)*, pages 1–9. IEEE, 2014. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[8] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *ACCV*, pages 19–34. Springer, 2016. 5, 7, 8

[9] Zexi Hu, Xiaoming Chen, Henry Wing Fung Yeung, Yuk Ying Chung, and Zhibo Chen. Texture-enhanced light field super-resolution with spatio-angular decomposition kernels. *IEEE Transactions on Instrumentation and Measurement*, 71:1–16, 2022. 5

[10] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *CVPR*, pages 723–731, 2018. 2

[11] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *ACM international conference on multimedia*, pages 2024–2032, 2019. 2

[12] 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*, pages 2260–2269, 2020. 1, 2, 6

[13] Kai Jin, Angulia Yang, Zeqiang Wei, Sha Guo, Mingzhi Gao, and Xiuzhuang Zhou. Distgepit: Enhanced disparity learning for light field image super-resolution. In *CVPRW*, pages 1373–1383, 2023. 1, 8

[14] Manchang Jin, Gaosheng Liu, Kunshu Hu, Xin Luo, Kun Li, and Jingyu Yang. Physics-informed ensemble representation for light-field image super-resolution. *arXiv preprint arXiv:2305.20006*, 2023. 4, 5

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

[16] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016. 6

[17] Mikael Le Pendu, Xiaoran Jiang, and Christine Guillemot. Light field inpainting propagation via low rank matrix completion. *IEEE Transactions on Image Processing*, 27(4): 1981–1993, 2018. 5, 6, 7

[18] Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, and Shilin Zhou. Light field image super-resolution with transformers. *IEEE Signal Processing Letters*, 29:563–567, 2022. 1, 3, 6

[19] Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, Shilin Zhou, and Yulan Guo. Learning non-local spatial-angular correlation for light field image super-resolution. In *ICCV*, pages 12376–12386, 2023. 1, 2, 6, 8

[20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, pages 136–144, 2017. 6, 8

[21] Gaosheng Liu, Huanjing Yue, Kun Li, and Jingyu Yang. Disparity-guided light field image super-resolution via feature modulation and recalibration. *IEEE Transactions on Broadcasting*, 69(3):740–752, 2023. 2, 6

[22] Gaosheng Liu, Huanjing Yue, Jiamin Wu, and Jingyu Yang. Intra-inter view interaction network for light field image super-resolution. *IEEE Transactions on Multimedia*, 25: 256–266, 2023. 2, 6, 8

[23] Gaosheng Liu, Huanjing Yue, Jiamin Wu, and Jingyu Yang. Efficient light field angular super-resolution with sub-aperture feature learning and macro-pixel upsampling. *IEEE Transactions on Multimedia*, 25:6588–6600, 2023. 1

[24] Gaosheng Liu, Huanjing Yue, Kun Li, and Jingyu Yang. Adaptive pixel aggregation for joint spatial and angular super-resolution of light field images. *Information Fusion*, 104:102183, 2024. 1

[25] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. In *ECCVW*, pages 41–55. Springer, 2020. 2

[26] Zhi Lu, Yu Liu, Manchang Jin, Xin Luo, Huanjing Yue, Zian Wang, Siqing Zuo, Yunmin Zeng, Jiaqi Fan, Yanwei Pang, et al. Virtual-scanning light-field microscopy for robust snapshot high-resolution volumetric imaging. *Nature Methods*, 20(5):735–746, 2023. 1

[27] Kaushik Mitra and Ashok Veeraraghavan. Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 22–28. IEEE, 2012. 1

[28] Yu Mo, Yingqian Wang, Chao Xiao, Jungang Yang, and Wei An. Dense dual-attention network for light field image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4431–4443, 2021. 2

[29] Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016. 1, 5, 6, 7, 8

[30] Vaibhav Vaish and Andrew Adams. The (new) stanford light field archive. *Computer Graphics Laboratory, Stanford University*, 6(7), 2008. 5, 7, 8

[31] Vinh Van Duong, Thuc Nguyen Huu, Jonghoon Yim, and Byeungwoo Jeon. Light field image super-resolution network via joint spatial-angular and epipolar information. *IEEE Transactions on Computational Imaging*, 9:350–366, 2023. 1, 6

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[33] Longguang Wang, Xiaoyu Dong, Yingqian Wang, Xinyi Ying, Zaiping Lin, Wei An, and Yulan Guo. Exploring sparsity in image super-resolution for efficient inference. In *CVPR*, pages 4917–4926, 2021. 2

[34] Longguang Wang, Yulan Guo, Xiaoyu Dong, Yingqian Wang, Xinyi Ying, Zaiping Lin, and Wei An. Exploring fine-grained sparsity in convolutional neural networks for efficient inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4474–4493, 2022. 2, 8

[35] Shunzhou Wang, Tianfei Zhou, Yao Lu, and Huijun Di. Detail-preserving transformer for light field image super-resolution. In *AAAI*, pages 2522–2530, 2022. 2, 6

[36] 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. 2

[37] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Spatial-angular interaction for light field image super-resolution. In *ECCV*, pages 290–308, 2020. 2, 6, 8

[38] Yingqian Wang, Jungang Yang, Longguang Wang, Xinyi Ying, Tianhao Wu, Wei An, and Yulan Guo. Light field image super-resolution using deformable convolution. *IEEE Transactions on Image Processing*, 30:1057–1071, 2020. 2, 6, 8

[39] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Wei An, and Yulan Guo. Occlusion-aware cost constructor for light field depth estimation. In *CVPR*, pages 19809–19818, 2022. 1

[40] Yingqian Wang, Longguang Wang, Gaochang Wu, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):425–443, 2022. 1, 2, 4, 5, 6, 8

[41] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Radu Timofte, et al. NTIRE 2023 challenge on light field image super-resolution: Dataset, methods and results. In *CVPRW*, 2023. 1, 5, 6, 8

[42] Yingqian Wang, Zhengyu Liang, Qianyu Chen, Longguang Wang, Jungang Yang, Radu Timofte, Yulan Guo, et al. NTIRE 2024 challenge on light field image super-resolution: Methods and results. In *CVPRW*, 2024. 8

[43] Yingqian Wang, Zhengyu Liang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Real-world light field image super-resolution via degradation modulation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 2, 4

[44] Sven Wanner and Bastian Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):606–619, 2013. 1

[45] Sven Wanner, Stephan Meister, and Bastian Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *VMV*, pages 225–226, 2013. 5, 8

[46] Zeyu Xiao, Ruisheng Gao, Yutong Liu, Yueyi Zhang, and Zhiwei Xiong. Toward real-world light field super-resolution. In *CVPRW*, pages 3407–3417, 2023. 2

[47] 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, 6

[48] 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*, pages 24–32, 2015. 2

[49] Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong. Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding*, 145:148–159, 2016. 8

[50] Shuo Zhang, Youfang Lin, and Hao Sheng. Residual networks for light field image super-resolution. In *CVPR*, pages 11046–11055, 2019. 1, 2, 6

[51] Shuo Zhang, Song Chang, and Youfang Lin. End-to-end light field spatial super-resolution network using multiple epipolar geometry. *IEEE Transactions on Image Processing*, 30:5956–5968, 2021. 2, 6

[52] 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*, pages 286–301, 2018. 6

[53] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019. 2