Learning Non-Local Spatial-Angular Correlation for Light Field Image Super-Resolution

Zhengyu Liang¹, Yingqian Wang¹, Longguang Wang², Jungang Yang¹*, Shilin Zhou¹, Yulan Guo¹
¹National University of Defense Technology, ²Aviation University of Air Force
{zyliang, yangjungang}@nudt.edu.cn

Abstract

Exploiting spatial-angular correlation is crucial to light field (LF) image super-resolution (SR), but is highly challenging due to its non-local property caused by the disparities among LF images. Although many deep neural networks (DNNs) have been developed for LF image SR and achieved continuously improved performance, existing methods cannot well leverage the long-range spatial-angular correlation and thus suffer a significant performance drop when handling scenes with large disparity variations. In this paper, we propose a simple yet effective method to learn the non-local spatial-angular correlation for LF image SR. In our method, we adopt the epipolar plane image (EPI) representation to project the 4D spatial-angular correlation onto multiple 2D EPI planes, and then develop a Transformer network with repetitive self-attention operations to learn the spatial-angular correlation by modeling the dependencies between each pair of EPI pixels. Our method can fully incorporate the information from all angular views while achieving a global receptive field along the epipolar line. We conduct extensive experiments with insightful visualizations to validate the effectiveness of our method. Comparative results on five public datasets show that our method not only achieves state-of-the-art SR performance but also performs robust to disparity variations. Code is publicly available at https://github.com/ZhengyuLiang24/EPIT.

1. Introduction

Light field (LF) cameras record both intensity and directions of light rays, and enable various applications such as depth perception [25, 29, 32], view rendering [3, 52, 66], virtual reality [11, 74], and 3D reconstruction [6, 77]. However, due to the inherent spatial-angular trade-off [82], an LF camera can either provide dense angular samplings with low-resolution (LR) sub-aperture images (SAIs), or capture high-resolution (HR) SAIs with sparse angular sampling. To handle this problem, many methods have been proposed to enhance the angular resolution via novel view synthesis [28, 43, 67], or enhance the spatial resolution by performing LF image super-resolution (SR) [10, 20]. In this paper, we focus on the latter task, i.e., generating HR LF images from their LR counterparts.

Recently, convolutional neural networks (CNNs) have been widely applied to LF image SR and demonstrated superior performance over traditional paradigms [1, 34, 44, 49, 64]. To incorporate the complementary information (i.e., angular information) from different views, existing CNNs adopted various mechanisms such as adjacent-view combination [73], view-stack integration [26, 78, 79], bidirectional recurrent fusion [59], spatial-angular disentanglement [36, 60, 61, 72, 9], and 4D convolutions [42, 43]. However, as illustrated in both Fig. 1 and Sec. 4.3, existing methods achieve promising results on LFs with small baselines, but suffer a notable performance drop when handling scenes with large disparity variations.
We attribute this performance drop to the contradictions between the local receptive field of CNNs and the requirement of incorporating non-local spatial-angular correlation in LF image SR. That is, LF images provide multiple observations of a scene from a number of regularly distributed viewpoints, and a scene point is projected onto different but correlated spatial locations on different angular views, which is termed as the spatial-angular correlation. Note that, the spatial-angular correlation has the non-local property since the difference between the spatial locations of two views (i.e., disparity value) depends on several factors (e.g., angular coordinates of the selected views, the depth value of the scene point, the baseline length of the LF camera, and the resolution of LF images), and can be very large in some situations. Consequently, it is appealing for LF image SR methods to incorporate complementary information from different views by exploiting the spatial-angular correlation under large disparity variations.

In this paper, we propose a simple yet effective method to learn the non-local spatial-angular correlation for LF image SR. In our method, we re-organize 4D LFs as multiple 2D epipolar plane images (EPIs) to manifest the spatial-angular correlation to the line patterns with different slopes. Then, we develop a Transformer-based network called EPIT to learn the spatial-angular correlation by modeling the dependencies between each pair of pixels on EPIs. Specifically, we design a basic Transformer block to alternately process horizontal and vertical EPIs, and thus progressively incorporate the complementary information from all angular views. Compared to existing LF image SR methods, our method can achieve a global receptive field along the epipolar line, and thus performs robust to disparity variations.

In summary, the contributions of this work are as follows: (1) We address the importance of exploiting non-local spatial-angular correlation in LF image SR, and propose a simple yet effective method to handle this problem. (2) We develop a Transformer-based network to learn the non-local spatial-angular correlation from horizontal and vertical EPIs, and validate the effectiveness of our method through extensive experiments and visualizations. (3) Compared to existing state-of-the-art LF image SR methods, our method achieves superior performance on public LF datasets, and is much more robust to disparity variations.

2. Related Work

2.1. LF Image SR

LFCNN [73] is the first method to adopt CNNs to learn the correspondence among stacked SAIs. Then, it is a common practice for LF image SR networks to aggregate the complementary information from adjacent views to model the correlation in LFs. Yeung et al. [72] designed a spatial-angular separable (SAS) convolution to approximate the 4D convolution to characterize the sub-pixel relationship of LF 4D structures. Wang et al. [59] proposed a bidirectional recurrent network to model the spatial correlation among views on horizontal and vertical baselines. Meng et al. [42] proposed a densely-connected network with 4D convolutions to explicitly learn the spatial-angular correlation encoded in 4D LF data. To further learn inherent correspondence relations in SAIs, Zhang et al. [78, 79] grouped LFs into four different branches according to the specific angular directions, and used four sub-networks to model the multi-directional spatial-angular correlation.

The aforementioned networks use part of input views to super-resolve each view, and the inherent spatial-angular correlation in LF images cannot be well incorporated. To address this issue, Jin et al. [26] proposed an All-to-One framework for LF image SR, and each SAI can be individually super-resolved by combining the information from all views. Wang et al. [61, 60] organized LF images into macro-pixels, and designed a disentangling mechanism to fully incorporate the angular information. Liu et al. [38] introduced 3D convolutions based multi-view context block to exploit the correlations among all views. In addition, Wang et al. [62] adopted deformable convolutions to achieve long-range information exploitation from all SAIs. Existing methods generally learn the local correspondence across SAIs, and ignore the importance of non-local spatial-angular correlation in LF images. However, due to the limited receptive field of CNNs, existing methods generally learn the local correspondence across SAIs, and ignore the importance of non-local spatial-angular correlation in LF images.

Recently, Liang et al. [36] applied Transformers to LF image SR and developed an angular Transformer and a spatial Transformer to incorporate angular information and model long-range spatial dependencies, respectively. However, since 4D LFs were organized into 2D angular patches to form the input of angular Transformers, the non-local property of spatial-angular correlations reduces the robustness of LFT to large disparity variations.

2.2. Non-Local Correlation Modeling

Non-local means [5] is a classical algorithm that computes the weighted mean of pixels in an image according to the self-similarity measure, and a number of studies on such non-local priors have been proposed for image restoration [12, 51, 19, 4], image and video SR [16, 76, 14, 71, 23]. Then, the attention mechanism is developed as a tool to bias the most informative components of an input signal, and achieves significant performance in various computer vision tasks [22, 8, 58, 15]. Huang et al. [24] proposed novel criss-cross attention to capture contextual information from full-image dependencies in an efficient way. Wang et al. [56, 55] proposed a parallax attention mechanism to handle
the varying disparities problem of stereo images. Wu et al. [69] applied attention mechanisms to 3D LF reconstruction and developed a spatial-angular attention module to learn the first-order correlation on EPIs.

Recently, the attention mechanism is further generalized as Transformers [54] with multi-head structures and feed-forward networks. Transformers have inspired lots of works [39, 35, 7, 13] to further investigate the power of attention mechanisms in visions. Liu et al. [40] presented a new understanding of Transformers to process spatial-temporal locality of videos for action recognition. Naseer et al. [45] investigated the robustness and generalizability of Transformers, and demonstrated favorable merits of Transformers over CNNs for occlusion handling. Shi et al. [50] observed that Transformers can implicitly make accurate connections for misaligned pixels, and presented a new understanding of Transformers to process spatially unaligned images.

3. Method

3.1. Preliminary

Based on the two-plane LF parameterization model [33], an LF image is commonly formulated as a 4D function $L(u, v, h, w) \in R^{U \times V \times H \times W}$, where $U$ and $V$ represent angular dimensions, $H$ and $W$ represent spatial dimensions. The EPI sample of 4D LF is acquired with a fixed angular coordinate and a fixed spatial coordinate. Specifically, the horizontal EPI is obtained with constant $u$ and $h$, and the vertical EPI is obtained with constant $v$ and $w$.

As shown in Fig. 2, the EPIs not only record spatial structures at edges or textures, but also reflect the disparity information via line patterns of different slopes. Specifically, due to large disparities, the EPIs contain abundant spatial-angular correlation of LFs in a long-range way. Therefore, we propose to explore the non-local spatial-angular correlation from horizontal and vertical EPIs for LF image SR.

3.2. Network Design

As shown in Fig. 3(a), our network takes an LR LF $L_{LR} \in R^{U \times V \times H \times W}$ as its input, and produces an HR LF $L_{SR} \in R^{U' \times V' \times H' \times W'}$, where $\alpha$ presents the upscaling factor. Our network consists of three stages including initial feature extraction, deep spatial-angular correlation learning, and feature upsampling.

3.2.1 Initial Feature Extraction

As shown in Fig. 3(b), we follow most existing works [7, 35, 75] to use three $3 \times 3$ convolutions with LeakyReLU [41] as a SpatialConv layer to map each SAI to a high-dimensional feature. The initially extracted feature can be represented as $F \in R^{U \times V \times H \times W \times C}$, where $C$ denotes the channel dimension.

3.2.2 Deep Spatial-Angular Correlation Learning

Non-Local Cascading Block. The basic module for spatial-angular correlation learning is the Non-Local Cascading block. As shown in Fig. 3(a), each block consists of two cascaded Basic-Transformer units to separately incorporate the complementary information along the horizontal and vertical epipolar lines. In our method, we employed five Non-Local Cascading blocks to achieve a global perception of all angular views, and followed SwinIR [35] to adopt spatial convolutions to enhance the local feature representation. The effectiveness of this inter-block spatial convolution is validated in Sec. 4.4. Note that, the weights of the two Basic-Transformer units in each block are shared to jointly learn the intrinsic properties of LFs, which is demonstrated effective in Sec. 4.4.

As shown in Fig. 3(c), the initial features $F$ can be first separately reshaped to the horizontal EPI pattern $F_{hor} \in R^{U' \times V \times H \times C}$ and the vertical EPI pattern $F_{ver} \in R^{W' \times V \times H \times C}$. Next, $F_{hor}$ (or $F_{ver}$) is fed to a Basic-Transformer unit to integrate the long-range information along the horizontal (or vertical) epipolar line. Then, the enhanced feature $F_{hor}$ (or $F_{ver}$) is reshaped into the size of $UV \times H \times W \times C$, and passes through a SpatialConv layer to incorporate the spatial context information within each SAI. Without loss of generality, we take the vertical Basic-Transformer as an example to introduce the detail of our Basic-Transformer unit in the following texts.

Basic-Transformer Unit. The objective of this unit is to capture long-range dependencies along the epipolar line via Transformers. To leverage the powerful sequence modeling capability of Transformers, we convert the vertical EPI features $F_{ver}$ to the sequences of ‘tokens” for capturing spatial-angular correlation in $U$ and $H$ dimensions. As shown in Fig. 3(d), the vertical EPI features are passed through a linear projection matrix $W_{ln} \in R^{C \times D}$, where $D$...
denotes the embedding dimension of each token. The projected EPI features are a sequence of tokens with a length of \( UH \), i.e., \( T_{\text{ver}} \in \mathbb{R}^{UH \times D} \). Following the PreNorm operation in [70], we also apply Layer Normalization (LN) before the attention calculation, and obtain the normalized tokens \( \tilde{T}_{\text{ver}} = \text{LN}(T_{\text{ver}}) \).

Afterwards, tokens \( T_{\text{ver}} \) are passed through the Self-Attention layer and transformed into the deep tokens involving non-local spatial-angular information along the vertical epipolar line. Specifically, \( T_{\text{ver}} \) need to be separately multiplied by \( W_Q \in \mathbb{R}^{D \times D} \), \( W_K \in \mathbb{R}^{D \times D} \) and \( W_V \in \mathbb{R}^{D \times D} \) to generate corresponding query, key and value components for self-attention calculation, i.e., \( Q_{\text{ver}} = T_{\text{ver}} W_Q \), \( K_{\text{ver}} = T_{\text{ver}} W_K \) and \( V_{\text{ver}} = T_{\text{ver}} W_V \).

Given a query position \( q = \{1, 2, \ldots, UH\} \) in \( Q_{\text{ver}} \) and a key position \( k = \{1, 2, \ldots, UH\} \) in \( K_{\text{ver}} \), the corresponding response \( A_{\text{ver}}(q, k) \in \mathbb{R} \) measures the mutual similarity of the pairs by the dot-product operation, followed by a Softmax function to obtain the attention scores on the vertical EPI tokens. That is,

\[
A_{\text{ver}}(q,k) = \text{Softmax}(Q_{\text{ver}}(q) \cdot K_{\text{ver}}(k) \cdot \sqrt{D}).
\]  

(1)

Based on the attention scores, the output of self-attention \( T'_{\text{ver}} \) can be calculated as the weighted sum of value. In summary, the calculation process of Self-Attention layer can be formulated as:

\[
T'_{\text{ver}} = A_{\text{ver}} V_{\text{ver}} + T_{\text{ver}}.
\]  

(2)

To further incorporate the spatial-angular correlation, following [54], our Basic-Transformer unit also contains the multi-layer perceptron (MLP) and LN, and generates the enhanced \( \tilde{T_{\text{ver}}} \) as:

\[
\tilde{T_{\text{ver}}} = \text{MLP}(\text{LN}(T'_{\text{ver}})) + T'_{\text{ver}}.
\]  

(3)

At the end of the Basic-Transformer unit, the enhanced \( \tilde{T_{\text{ver}}} \) is fed to another linear projection \( W_{\text{out}} \in \mathbb{R}^{D \times C} \), and reshaped into the size of \( UV \times H \times W \times C \) for the subsequent SpatialConv layer.

**Cross-View Similarity Analysis.** Note that, the setting \( [A_{\text{ver}}(q, 1), \ldots, A_{\text{ver}}(q, UH)] \in \mathbb{R}^{1 \times UH} \) represents the similarity scores of \( q \) with all \( k \) in \( K_{\text{ver}} \), and thus can be re-organized as a slice of cross-view attention map according to the angular coordinate. Inspired by this, we visualized the cross-view attention maps on an example scene in Fig. 4. The regions marked by the red stripe in Fig. 4(a) are set as the query tokens, and the self-similarity (i.e.,
Table 1. Quantitative comparison of different SR methods in terms of the number of parameters (#Prm.) and PSNR/SSIM. Larger PSNR and SSIM values indicate higher SR quality. We mark the best results in red and the second best results in blue.

<table>
<thead>
<tr>
<th>Methods</th>
<th>#Prm.(M)</th>
<th>EPFL</th>
<th>HCInew</th>
<th>HCold</th>
<th>INRIA</th>
<th>STFgantry</th>
<th>EFLL</th>
<th>HColn</th>
<th>INRIA</th>
<th>STFgantry</th>
</tr>
</thead>
</table>

4.2. Comparisons on Benchmark Datasets

We compare our method to 14 state-of-the-art methods, including 3 single image SR methods [30, 37, 81] and 11 LF image SR methods [79, 72, 26, 61, 62, 78, 38, 79, 60, 9]. Quantitative Results. A quantitative comparison among different methods is shown in Table 1. Our EPIT with a small model size (i.e., 1.42M/1.47M for 2×/4× SR) achieves state-of-the-art PSNR and SSIM scores on almost all the datasets for both 2× and 4× SR. It is worth noting that LFs in the STFgantry dataset [53] have larger disparity variations, and are thus more challenging. Our EPIT significantly outperforms all the compared methods and achieves 1.66dB/0.32dB PSNR improvements over the second-best-performing method LFT for 2×/4× SR, respectively, which demonstrates the powerful capacity of our EPIT in non-local correlation modeling.
Figure 5. Qualitative comparison of different SR methods for $2 \times 4 \times$ SR.

Figure 6. Quantitative and qualitative (MSE) comparisons of disparity estimation results achieved by SPO [80] using different SR results. The MSE (\(\times\)) is the mean square error multiplied by 100.

Table 2. PSNR values achieved by DistgSSR [60] and our EPIT with different angular resolution for $4 \times$ SR.

<table>
<thead>
<tr>
<th>Input</th>
<th>EFPI</th>
<th>HC/new</th>
<th>HC/old</th>
<th>INRIA</th>
<th>STF/lytro</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 x 2</td>
<td>Ours</td>
<td>30.88/0.919</td>
<td>30.79/0.919</td>
<td>30.69/0.919</td>
<td>30.59/0.919</td>
</tr>
<tr>
<td>3 x 3</td>
<td>Ours</td>
<td>31.07/0.920</td>
<td>31.18/0.920</td>
<td>31.29/0.920</td>
<td>31.40/0.920</td>
</tr>
<tr>
<td>3 x 4</td>
<td>Ours</td>
<td>31.25/0.921</td>
<td>31.36/0.921</td>
<td>31.47/0.921</td>
<td>31.58/0.921</td>
</tr>
<tr>
<td>5 x 5</td>
<td>Ours</td>
<td>31.38/0.922</td>
<td>31.50/0.922</td>
<td>31.61/0.922</td>
<td>31.72/0.922</td>
</tr>
<tr>
<td>6 x 6</td>
<td>Ours</td>
<td>31.40/0.923</td>
<td>31.52/0.923</td>
<td>31.63/0.923</td>
<td>31.74/0.923</td>
</tr>
<tr>
<td>7 x 7</td>
<td>Ours</td>
<td>31.52/0.924</td>
<td>31.64/0.924</td>
<td>31.76/0.924</td>
<td>31.87/0.924</td>
</tr>
<tr>
<td>8 x 8</td>
<td>Ours</td>
<td>31.64/0.925</td>
<td>31.76/0.925</td>
<td>31.88/0.925</td>
<td>31.99/0.925</td>
</tr>
<tr>
<td>9 x 9</td>
<td>Ours</td>
<td>31.76/0.926</td>
<td>31.88/0.926</td>
<td>32.00/0.926</td>
<td>32.11/0.926</td>
</tr>
</tbody>
</table>

Qualitative Results. Figure 5 shows the qualitative results achieved by different methods for $2 \times 4 \times$ SR. It can be observed from the zoom-in regions that single image SR method RCAN [81] cannot recover the textures and details in the SR images. In contrast, our EPIT can incorporate sub-pixel correspondence among SAs and generate more faithful details with fewer artifacts. Compared to most LF image SR methods, our EPIT can generate superior visual results with high angular consistency. Please refer to the supplemental material for additional visual comparisons.

Angular Consistency. We evaluate the angular consistency by using the $4 \times$ SR results on several challenging scenes (e.g., Backgammon and Stripes) in 4D LF benchmark [21] for disparity estimation. As shown in Fig. 6, our EPIT achieves competitive MSE scores on these challenging scenes, which demonstrates the superiority of our EPIT on angular consistency.

Performance with Different Angular Resolution. Since the angular resolution of LR images can vary significantly with different LF devices, we compare our method to DistgSSR [60] on LFs with different angular resolutions. It can be observed from Table 2 that our method achieves higher PSNR values than DistgSSR on almost all the datasets with each angular resolution (except on the EPFL and INRIA datasets with $2 \times 2$ input LFs). The consistent performance improvements demonstrate that our EPIT can well model the spatial-angular correlation with various angular resolutions. More comparisons and discussions are provided in the supplemental material.
Performance on Real-World LF Scenes. We compare our method to state-of-the-art methods under real-world degradation by directly applying them to LFs in the STFlytro dataset [47]. Since no groundtruth HR images are available in this dataset, we present the LR input and their super-resolved results in Fig. 7. It can be observed that our method can recover more faithful details and generate more clear letters than other methods. Since the LF structure keeps unchanged under both bicubic and real-world degradation, our method can learn the spatial-angular correlation from bicubically downsampled training data, and well generalize to LF images under real degradation.

4.3. Robustness to Large Disparity Variations

Considering the parallax structure of LF images, we followed the shearing operation in existing works [67, 68] to linearly change the overall disparity range of LF datasets. Note that, the content of SAIs maintain unchanged after the shearing operation, and thus we can quantitatively investigate the performance of different SR methods with respect to the disparity variations.

Quantitative & Qualitative Comparison. Figure 8 shows the quantitative and qualitative results of different SR methods with respect to sheared values, from which we can observe that: 1) Except for the single image SR method RCAN, all LF image SR methods suffer a performance drop when the absolute sheared value of LF images increases. That is because, large sheared values can result in more significant misalignment among LF images, and introduce difficulties in complementary information incorporation; 2) As the absolute sheared value increases, the performance of existing LF image SR methods is even inferior to RCAN. The possible reason is that, these methods do not make full use of local spatial information, but rather rely on local angular information from adjacent views. When the sheared value exceeds their receptive fields, the large disparities can make the spatial-angular correlation non-local and thus introduce challenges in complementary information incorporation; 3) Our EPIT performs much more robust to disparity variations and achieves the highest PSNR scores under all sheared values. More quantitative comparisons on the whole datasets can be referred to the supplemental material.

LAM Visualization. We used Local Attribution Map (LAM) [18] to visualize the input regions that contribute to the SR results of different methods. As shown in Fig. 8, we first specify the center of green stripes in HR images as the target regions, and then re-organize the corresponding attribution maps on LR images into the EPI patterns. It can be observed that RCAN achieves a larger receptive field along the spatial dimension than other compared methods, which supports the results in Figs. 8(b) and 8(e) that RCAN achieves a relatively stable SR performance with different
sheared values. It is worth noting that our EPIT can automatically incorporate the most relevant information from different views, and can learn the non-local spatial-angular correlation regardless of disparity variations.

**Perspective Comparison.** We compare the performance of MEG-Net, DistgSSR and our method with respect to different perspectives and sheared values (0 to 3). It can be observed in Fig. 9 that, both MEG-Net and DistgSSR suffer significant performance drops on all perspectives as the sheared value increases. In contrast, our EPIT can well handle the disparity variation problem, and achieve much higher PSNR values with a balanced distribution among different views regardless of the sheared values.

**4.4. Ablation Study**

In this subsection, we compare the performance of our EPIT with different variants to verify the effectiveness of our design choices, and additionally, investigate their robustness to large disparity variations.

**Horizontal/Vertical Basic-Transformer Units.** We demonstrated the effectiveness of the horizontal and vertical Basic-Transformer units in our EPIT by separately removing them from our network. Note that, without using horizontal or vertical Basic-Transformer unit, these variants cannot incorporate any information from the corresponding angular directions. As shown in Table 3, both variants w/o-Horizontal and w/o-Vertical suffer a decrease of 0.72dB in the INRIA dataset as compared to EPIT, which demonstrates the importance of exploiting spatial-angular correlations from all angular views.

**Weight Sharing in Non-Local Cascading Blocks.** We introduced the variant w/o-Share by removing the weight sharing between horizontal and vertical Basic-Transformer units. As shown in Table 3, the additional parameters in variant w/o-Share do not introduce further performance improvement. It demonstrates that the weight sharing strategy between two directional Basic-Transformer units is beneficial and efficient to regularize the network.

**SpatialConv in Non-Local Cascading Blocks.** We introduced the variant w/o-Local by removing the SpatialConv layers from our EPIT, and we adjusted the channel number to make the model size of this variant not smaller than the main model. As shown in Table 3, the SpatialConv has a significant influence on the SR performance, e.g., the variant w/o-Local suffers a 0.41dB PSNR drop on the EPFL dataset. It demonstrates that local context information is crucial to the SR performance, and the simple convolutions can fully incorporate the spatial information from each SAI.

**Basic-Transformer in Non-Local Cascading Blocks.** We introduced the variant w/o-Trans by replacing Basic-Transformer in Non-Local Blocks with cascaded convolutions. As shown in Table 3, w/o-Trans suffers a most significant performance drop as the sheared value increases, which demonstrates the effectiveness of the Basic-Transformer in incorporating global information on the EPIs.

**Basic-Transformer Number.** We introduced the variants with-n-Block (n=1,2,3) by retaining n Non-Local Blocks. Results in Table 3 show the effectiveness of our EPIT (having 5 Non-Local Blocks) with higher-order spatial-angular correlation modeling capability.

**5. Conclusion**

In this paper, we propose a Transformer-based network for LF image SR. By modeling the dependencies between each pair of pixels on EPIs, our method can learn the spatial-angular correlation while achieving a global receptive field along the epipolar line. Extensive experimental results demonstrated that our method can not only achieve state-of-the-art SR performance on benchmark datasets, but also perform robust to large disparity variations.

**Acknowledgment:** This work was supported in part by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61921001.
References


