Cross-MPI: Cross-scale Stereo for Image Super-Resolution using Multiplane Images

Yuemei Zhou¹, Gaochang Wu², Ying Fu³, Kun Li⁴, Yebin Liu¹
¹Tsinghua University, ²Northeastern University, ³Beijing Institute of Technology, ⁴Tianjin University
zyml8@mails.tsinghua.edu.cn, ahwgc2009@163.com, fuying@bit.edu.cn, lik@tju.edu.cn, liuyebin@mail.tsinghua.edu.cn

Abstract

The combination of various cameras is enriching the way of computational photography, among which reference-based super-resolution (RefSR) plays a critical role in multiscale imaging systems. However, existing RefSR approaches fail to accomplish high-fidelity super-resolution under a large resolution gap, e.g., 8× upscaling, due to the less consideration of underneath scene structure. In this paper, we aim to solve the RefSR problem (in actual multiscale camera systems) inspired by multiplane images (MPI) representation. Specifically, we propose Cross-MPI, an end-to-end RefSR network composed of a novel plane-aware attention-based MPI mechanism, a multiscale guided upsampling module as well as a super-resolution (SR) synthesis and fusion module. Instead of using a direct and exhaustive matching between the cross-scale stereo, the proposed plane-aware attention mechanism fully utilizes the concealed scene structure for efficient attention-based correspondence searching. Further combined with a gentle coarse-to-fine guided upsampling strategy, the proposed Cross-MPI is able to achieve a robust and accurate detail transmission. Experimental results on both digital synthesized and optical zoomed cross-scale data show the Cross-MPI framework can achieve superior performance against the existing RefSR methods, and is a real fit for actual multiscale camera systems even with large scale differences.

1. Introduction

With computational photography prospering over the past few decades, higher quality and fidelity imaging is in sustained demand for both industry and daily life. Hybrid multiscale imaging camera systems [27, 29, 9, 19] stand out for their capability to compose different camera characteristics and obtain a much higher spatial bandwidth. Hybrid multiscale imaging has been successfully applied in various imaging systems including hybrid light field camera [27], gigapixel video system [29], and mobile phones with multi-

Figure 1: Top: Our Cross-MPI network takes cross-scale stereo images as input and generates a super-resolved MPI through a delicate plane-aware attention mechanism and an effective multiscale guided upsampling module. Bottom: A 8× SR comparison between Cross-MPI and the state-of-art RefSR work, CrossNet [37] and TTSR [28], along with our intermediate depth estimation result.

camera system [9, 19]. In order to realize hybrid multiscale imaging, one of the key issues is how to match and fuse cross-resolution images from different cameras under certain perspective parallax.

Reference-based image super-resolution (RefSR) has been proposed to solve cross-scale reconstruction problem, and applied to gigapixel video system [29] or hybrid light field imaging systems [1, 27]. RefSR aims to recover high-resolution (HR) information damaged in imaging and digitization process by transferring and fusing high-quality details from known HR references. Unlike data-driven single image super-resolution (SISR) [16, 33, 32, 14, 17] that only utilizes prior knowledge learned from massive data to build an upsampling model, RefSR can produce much more satisfying results thanks to the additional HR references, espe-
currently for upsampling factors larger than $4 \times$.

Currently, most RefSR methods typically follow a "warp-and-synthesis" pipeline which can be further divided into two branches based on different detail transferring schemes including pixel-alignment and patch-matching RefSR. For the pixel-alignment RefSR, the warping process is conducted based on a certain pixel-to-pixel mapping between two images, e.g., optical flow [37] and depth map [27, 35]. Nevertheless, pixel-alignment RefSR would fail when dealing with large-parallax cases or image pairs under a large resolution gap, due to the lack of ability to capture long-distance correspondences. For the patch-matching RefSR [1, 36, 27, 34, 28], image patches are extracted to locate correspondences and blended accordingly for HR detail transmission. However, it is limited by complex textures or textureless areas and inherently generates grid artifacts due to the averaging window blurs. To summarize, existing RefSR methods neglect the underneath scene structure of cross-scale views, and cannot capture correspondences and transfer high-fidelity details with large resolution gaps (e.g., $8 \times$) and large parallax.

In this paper, we set our sights on the emerging scene representation and view synthesis framework, i.e., Multiplane Images (MPI), to solve the cross-scale RefSR problem in actual multiscale camera systems. Thanks to the MPI, the performance of image-based novel view synthesis [38, 23, 22, 5, 18] has drastically improved in the last few years. MPI has many good properties, including the ability to capture concealed surfaces from the input views and the adaptability to discontinuous edges and transparent scenes. However, the MPI estimation requires image pairs to have the same resolution with slight viewpoint changes, which is not available in RefSR tasks, especially when the resolution gap of the input pair is up to $8 \times$.

To fully understand the scene structure, we propose an end-to-end MPI-based approach that fully explores the properties of cross-scale image pairs by leveraging vanilla MPI in an efficient multiscale guided coarse-to-fine pipeline. Firstly, we introduce a novel MPI representation established on a lightweight plane-aware attention module. Compared with the original MPI estimation network with simple concatenation and convolution, our plane-aware attention mechanism explicitly captures correspondences along depth planes. Moreover, compared with popular non-local attention mechanisms [25, 6, 24, 28] which have to maintain a huge matrix for exhaustive pixel-to-pixel comparison, it takes much less computation and storage space. Secondly, in order to estimate robust and detailed correspondences in a space with a considerable resolution difference, e.g., $8 \times$, we design a multiscale guided upsampling module. This module takes the correspondences captured in low-resolution (LR) space as input, and conducts a gentle correspondence upsampling with guidance from the HR reference view in a coarse-to-fine manner, producing a super-resolved MPI. Finally, accurate transmission and fusion of high-quality details can be accomplished utilizing the obtained super-resolved MPI representation. The entire pipeline is designed in an end-to-end fashion, and is able to infer high fidelity depth information in an unsupervised manner. Experimental results show our method is visually and quantitatively superior to existing methods with high computational efficiency. We also build a real hybrid cross-scale stereo system and collect zooming data to validate the superior performance of the proposed algorithm. The main contributions of this paper are:

- We take a close look at the RefSR problem through the lens of the MPI representation, achieving super-resolution up to $8 \times$ on real hybrid imaging systems.
- We propose a novel plane-aware attention mechanism for MPI estimation which can achieve more explicit and efficient correspondence estimation compared with the original direct concatenation-and-convolution operation.
- We propose a novel multiscale guided upsampling module for cross-scale multiplane image synthesis which can well solve the matching problem under large resolution differences. A fine-detailed depth map which encodes the scene structure can also be inferred.

2. Related Work

In this section we review previous work on single image super-resolution (SISR) and reference-based image super-resolution (RefSR) that are the most relevant to our work.

2.1. Single Image Super-Resolution

In recent years, learning based methods have dominated the SISR research community due to the pioneering work of super-resolution convolutional neural network (SRCNN) [3] proposed by Dong et al., which maps LR images to HR ground truth by a 3-layer CNN. Kim et al. [11, 12] further improve the performance by equipping deeper networks. A series of work including SRRResNet [14], EDSR [16] and RDN [33] leverage residual learning [8] to boost the development of SISR. Zhang et al. [32] propose to improve residual learning by channel attention mechanism and build a deep model called RCAN to further improve the performance. However, these methods simply focus on raw pixel comparison, e.g., mean square error (MSE) for model optimization, which tends to produce blurry images. To mitigate this problem, Johnson et al. [10] first introduce the high-level perceptual loss into SR tasks to mimic the human perception system. After that, SRGAN [14] adopts generative adversarial networks (GANs) [7] to generate photo-realistic SR results. Sajjadi et al. [20] propose EnhanceNet that utilizes Gram matrix based texture matching loss to enhance local similar textures. Further, ESRGAN [26] enhances SRGAN by introducing Residual-in-Residual Dense Block (RRDB) and
relativistic adversarial loss. Moreover, RSRGAN [31] improves SR visual quality by a Ranker that can learn the behavior of perceptual metrics and a novel rank-content loss to optimize the perceptual quality. More recently, Ma et al. [17] propose SPSR that fully utilize gradient priors and structural information of images to achieve a better visual quality.

2.2. Referenced-based Image Super-Resolution

Compared with SISR that directly hallucinates missing HR details from the LR input, RefSR benefits from additional referable HR inputs. Since the decisive factor for RefSR to achieve good performance is the accuracy of matching and fusion between the LR input and HR reference, we can classify RefSR work into pixel-alignment related and patch-matching related. For the pixel-alignment RefSR, a pixel-to-pixel matching and warping is estimated and operated for the HR reference registration. Early work like Landmark [30], is proposed to retrieve correlated web images as reference, and a global registration and an energy minimizing problem are solved to recover the desired SR image. Multiscale gigapixel video [29] proposes to achieve cross-scale view alignment in the gigapixel video system through mesh-based warping and optical flow warping. However, mesh-based warping limits the deformation when large parallax occurs. Except for warping based on matched feature points, the work of reconstructing light field with extra HR reference like [27, 35] further introduce depth map estimation for transferring finer high-frequency details. But these methods focus on light field depth estimation and cannot be used to robustly find cross-scale correspondences for unstructured camera arrays like [29]. Recently, CrossNet [37] proposes a learning based method to conduct a cross-scale warping by predicting optical flow, which achieves high accuracy on light field datasets but fails when dealing with large-parallax cases. These pixel-alignment RefSR methods heavily depend on the pixel-level alignment quality between LR and HR reference, where optical flow estimation and feature point matching are limited to small-parallax conditions.

Furthermore, some work [1, 36, 27] utilize image patches to locate correspondences and synthesize HR details from references. Despite the favorable performance achieved by these methods, they are limited to the huge searching space of patch-based methods and tend to induce the problem of blending blur as a result of averaging the transferred HR patches. Further expanding references to HR images without view constraints like light field applications, SRNTT [34] and TTSSR [28] combine patch-matching with advanced deep learning architectures to achieve visually pleasant SR results. The former work reconsiders RefSR as a texture transfer problem and synthesizes swapped pre-trained VGG features in a style transfer way, while the latter designs a texture transformer fully utilizing attention mechanism to accomplish matching and synthesis. Both work require exhausting comparison between the extracted patches and are unable to handle image areas with less or repeating textures.

3. Approach

Given a cross-scale stereo camera system, our goal is to use the HR reference image \( I_{Ref} \) to super-resolve the corresponding LR image \( I_{LR} \). We formulate the training tuple as \( \{I_{LR}, I_{Ref}, c_{LR}, c_{Ref}, I_{GT}\} \), where \( I_{LR} \in \mathbb{R}^{h \times w \times c} \), \( I_{Ref} \in \mathbb{R}^{\beta h \times \beta w \times c} \), \( \beta \) is the resolution difference factor, \( c_{LR} \) and \( c_{Ref} \) are calibration parameters (including intrinsics and extrinsics) of the input cross-scale stereo pair, and \( I_{GT} \in \mathbb{R}^{\beta h \times \beta w \times c} \) is the super-resolution ground truth of the LR view. To decode the captured scene components into planes, we also calculate a plane sweep volume \( PSV_{Ref} = \{PSI_{Ref}^i\}, i = 1, ..., d \) from \( \{I_{Ref}, c_{LR}, c_{Ref}\} \), where \( d \) is the number of depth planes.

Different from traditional MPI estimation, our stereo input suffers from a large resolution difference, \( 8 \times \) for instance. To match the frequency band of \( I_{LR} \) and obtain alpha maps with target resolution and fine details, we first calculate rough alpha maps at low resolution using plane-aware attention as depicted in Sect. 3.1 and gradually add details through the proposed multiscale guided upsampling module described in Sect. 3.2. Then, the following SR synthesis and fusion module depicted in Sect. 3.3 is used to adequately transfer details from HR reference and generate the final SR result. Besides, loss functions for optimizing the proposed Cross-MPI network as well as the implementation details are discussed in Sect. 3.4 and Sect. 3.5. Please see Fig. 3 for an overview.

3.1. Plane-Aware Attention-based MPI

Here, we introduce our innovative plane-aware attention mechanism for efficient feature matching. The scene representation we adopt, multiplane images (MPI) [38], is composed of multiple fronto-parallel planes at a fixed depth range relative to a selected view. The MPI representation is a collection of RGBA layers \( \{(C_1, \alpha_1), ..., (C_d, \alpha_d)\} \), where \( C \) indicates RGB color, \( \alpha \) indicates occupancy of each plane and \( d \) is the number of depth planes. Please see
Fig. 2 for more details. As we can know from this representation, the set of alpha maps, $\mathcal{A} = \{\alpha_i\}, i = 1, ..., d$, indicates the occupancy of colors on each depth planes and also reflects the softness of the decoded scene layers especially for boundaries and reflective or transparent parts. In our RefSR task, the alpha maps are perfect for transferring HR details from the reference view to the LR view.

However, MPI is normally estimated by a CNN whose input is the concatenation of one dominant view and the plane sweep volume of the other view. Other than learning alpha maps simply through the stack of network layers with implicit and weak comparisons, we develop a plane-aware attention-based MPI mechanism to achieve an explicit matching, as depicted in Fig. 4. Inspired by the strong capability of attention to capture long-range dependencies in computer vision tasks, such as image segmentation [6], stereoSR [24] and RefSR [28], our plane-aware attention further considers the inherent principle of MPI estimation procedure from stereo input, and explicitly conduct an adequate and efficient comparison on each depth plane to estimate multi-plane correspondences.

**Shared Feature Extractor.** We first extract dense features from the LR inputs ($I_{LR}, PSI_{Ref}^i$) through a shared feature extractor (SFE), which is a Residual Atrous Spatial Pyramid Pooling (ResASPP) module [24]. Given $I_{LR} \in \mathbb{R}^{h \times w \times c_x}$ and $PSI_{Ref}^i \in \mathbb{R}^{h \times w \times c_x \times d} = \{PSI_{Ref}^i\}, i = 1, ..., d$ where $PSI_{Ref}^i$ is calculated by $1/\beta \times$ bicubic downsampling to match the LR domain, the feature extraction process can be expressed as:

$$F_{LR} = \text{SFE}(I_{LR}), \quad FV_{Ref}^i = \text{Concat}((\text{SFE}(PSI_{Ref}^i))_{i=1}^d),$$

where $F_{LR} \in \mathbb{R}^{h \times w \times c_x}$, $FV_{Ref}^i \in \mathbb{R}^{1 \times c_e \times d}$, $c_x$ is the channel size of the feature maps, $\text{SFE}(\cdot)$ is the shared feature extractor and $\text{Concat}(\cdot)$ is the channel-wise concatenation. Besides, the receptive field of this feature extractor

![Figure 3: The Cross-MPI pipeline. We first estimate the initial alpha maps through a plane-aware attention-based MPI module. A novel multiscale guided upsampling module is then designed for generating super-resolved alpha maps. To transfer HR details of reference view and generating the final SR result, the pipeline ends with a SR synthesis and fusion module. The whole pipeline is elaborately designed for cross-scale stereo RefSR problem by fully considering camera relationships as well as the underneath scene structure.](image)

![Figure 4: Our proposed novel plane-aware attention mechanism, which explicitly calculates the correlation of the input view and the corresponding plane sweep image on each image plane.](image)

is large enough for further plane-aware attention formation based on the structure of ResASPP module.

**Plane-Aware Attention.** Once the rich features of the LR images to be matched are extracted, a novel plane-aware attention is proposed to accomplish a delicate matching. As illustrated in Fig. 4, given the features at the LR view $F_{LR}$ and plane sweep features of the reference view $FV_{Ref}^i$, we reshape them into $\tilde{F}_{LR} \in \mathbb{R}^{n \times c_e}$ and $\tilde{FV}_{Ref}^i \in \mathbb{R}^{n \times c_e \times d}$ where $n = h \times w$ is the number of pixels. After that, we perform a batch-wise matrix multiplication between $\tilde{F}_{LR}$ and $\tilde{FV}_{Ref}^i$ and apply a softmax layer to obtain the plane-aware attention $\tilde{A}_{init} \in \mathbb{R}^{n \times 1 \times d}$, which is then reshaped to be the initial estimation of the alpha maps $\tilde{A}_{init} \in \mathbb{R}^{h \times w \times d}$. The plane-aware attention estimation procedure can be expressed as:

$$\tilde{A}_{init} = \text{softmax}(\tilde{F}_{LR} \otimes \tilde{FV}_{Ref}^i),$$

$$A_{init} = \text{reshape}(\tilde{A}_{init}),$$

where $\otimes$ denotes batch-wise matrix multiplication. The initial alpha maps $A_{init}$ represent the similarities between $I_{LR}$ and $PSI_{Ref}^i$ on each depth plane. The proposed plane-aware attention effectively decomposes scene components into predefined planes.
3.2. Multiscale Guided Upsampling Module

Benefiting from plane-aware attention-based MPI in LR space, our method can efficiently obtain the initial alpha maps. To further upsample the alpha maps into the target spatial size, we design a multiscale guided upsampling module, which can add more details in a coarse-to-fine manner for a more precise mapping. We concatenate \( I_{LR1} \) and \( PSV_{Ref} \) as guidance while gradually upsampling the initial alpha maps \( A_{init} \in \mathbb{R}^{h\times w\times d} \), where \( I_{LR1} \) is calculated by \( \beta \times \) bicubic upsampling the \( I_{LR} \) to match the HR domain. As illustrated in Fig. 5, we use a multi-scale guided filter within the decoder of \( A_{init} \), and features are fused into the up-sampling procedure progressively. The multiscale guided upsampling module can be represented as:

\[
G_{l-1} = \left[ \text{Res}(G_l) \right] \Downarrow 2x,
\]

\[
A_l = \left[ \text{Res(Concat}(A_{l-1}, G_{l-1})) \right] \Uparrow 2x, \tag{3}
\]

where \( G_l \) is the guidance map at spatial level \( l \) with the maximum level guidance map indicating \( \text{Concat}(I_{LR1}, PSV_{Ref}) \). \( A_l \) is the attention feature map at level \( l \) with the minimum level feature map as input \( A_{init} \) and \( \text{Res}(\cdot) \) denotes residual blocks. Notation \( \text{Concat}(\cdot) \) denotes channel-wise concatenation, \( \Downarrow 2x \) and \( \Uparrow 2x \) denote nearest-neighbour resize with \( 2 \times \) scale. In the output layer of the multiscale guided upsampling module, a \( 1 \times 1 \) convolution is applied to recover the channel size of the final alpha maps as:

\[
A = \text{softmax(Conv}_{1\times 1}(\text{Res(Concat}(A_l, G_l)))}, \tag{4}
\]

where the output alpha maps \( A \) have a spatial size of \( \beta_h \times \beta_w \times d \), and the softmax operation ensures that the depth probabilities sum to one on the emitting light of each pixel unit.

3.3. SR Synthesis and Fusion

After we obtain the alpha maps \( A \) with the same SR spatial size, we can generate a SR MPI of the LR view. The SR MPI only contains HR colors from the HR reference view, \( i.e. \), HR plane sweep images element-wise multiplying the alpha maps on each depth planes. This procedure can be presented as:

\[
MPI_{SR} = PSV_{Ref} \circ A, \tag{5}
\]

where \( MPI_{SR} \in \mathbb{R}^{h\times w\times c\times d} \) and \( \circ \) denotes the Hadamard product.

By element-wise summarizing \( MPI_{SR} \) through the depth plane channel, a transferred HR image \( T_{Ref} \in \mathbb{R}^{h\times w\times c} \) at the LR view can be synthesized. To further fuse the synthesized fine details according to the LR input and produce a final SR result, we then design a SR fusion sub-network. The SR synthesis and fusion process can be expressed as:

\[
T_{Ref} = \sum_i MPI_{SR}^i, \tag{6}
\]

\[
I_{SR} = \text{FuseNet}(T_{Ref}, I_{LR1}),
\]

where \( MPI_{SR}^i \in \mathbb{R}^{h\times w\times c} \) is the \( i \)-th component on the depth plane channel, \( I_{LR1} \) is calculated by \( \beta \times \) bicubic upsampling, and \( \text{FuseNet(\cdot)} \) denotes the SR fusion sub-network which is a residual network with multiple cascaded sub-residual blocks. Above all, the SR synthesis and fusion module is designed to adequately transfer HR textures to align with the LR input.

The whole pipeline is trained in an end-to-end way, and we can additionally obtain the scene depth \( D_{SR} \in \mathbb{R}^{\beta_h \times \beta_w \times 1} \) from the alpha maps by simply applying an argmax function:

\[
D_{SR} = \arg \max_i (A^\beta_h \times \beta_w \times i). \tag{7}
\]

We also provide some visual results of our super-resolved depths in Fig. 8.

3.4. Loss Functions

The design of our overall loss considers maintaining the spatial characteristics of the SR image, gaining better visual quality and supervising within the multiscale structure to obtain precise alpha maps. Specifically, the loss is composed of the following three parts, reconstruction loss, perception loss and internal supervision loss:

\[
\mathcal{L}_{all} = \lambda_{rec} \mathcal{L}_{rec} + \lambda_{per} \mathcal{L}_{per} + \lambda_{is} \mathcal{L}_{is}. \tag{8}
\]

**Reconstruction Loss.** We firstly define a reconstruction loss to encourage the output \( I_{SR} \) to match the ground truth, and we choose L1 per-pixel loss:

\[
\mathcal{L}_{rec} = \frac{1}{\beta_h \times \beta_w \times c} \| I_{SR} - I_{GT} \|_1, \tag{9}
\]

where \( (\beta_h, \beta_w, c) \) is the spatial size of the SR. Moreover, we adopt \( L_1 \) loss to produce more sharp results and easier convergence.
Perceptual Loss. Perceptual loss has been proved to improve visual quality and has been successfully applied to image super-resolution and generation tasks [4, 10, 14, 34, 28]. Specifically, we adopt the normalized VGG-19 [21] referring to the layer matching from [2]:

$$L_{per} = \sum_i \lambda_i \| \phi_i(I_{SR}) - \phi_i(I_{GT}) \|_1,$$

where \(\{\phi_i\}\) is a set of appointed neural layers of VGG-19 and the weights \(\{\lambda_i\}\) are set as the inverse of neuron numbers in each layer.

Internal Supervision Loss. Since a precise transfer from HR reference is crucial in our RefSR task and we proposed to estimate alpha maps in LR space followed by gradually guided upsampling, it is significant to concern the internal estimation quality. Thus we propose an internal supervision loss to monitor the accuracy of initial estimate of the alpha maps \(A_{init}\). Follow the content transfer in Equation (6), we reshape \(PSV_{Ref}\in \mathbb{R}^{h\times w\times c}\times d\) into \(\tilde{PSV}_{Ref}\in \mathbb{R}^{n\times c\times d}\) and \(A_{init}\in \mathbb{R}^{h\times w\times c}\) into \(\tilde{A}_{init}\in \mathbb{R}^{n\times d\times 1}\) where \(n = h \times w\), the loss can be expressed as:

$$L_{is} = \frac{1}{h \times w \times c} \| \tilde{PSV}_{Ref}\otimes \tilde{A}_{init} - I_{LR} \|_1,$$

where \(\otimes\) denotes batch-wise matrix multiplication. This internal supervision loss encourages the initial alpha maps to produce an accurate content warping at the beginning, which helps our whole pipeline to generate more precise SR alpha maps for HR detail transmission.

3.5. Implementation Details

The shared feature extractor contains a residual ASPP (Atrous Spatial Pyramid Pooling) block similar to [24] concatenated with a regular residual block. To balance the LR plane-ware attention and guided upsampling procedure, we calculate the plane-ware attention at scale \(2\times\) and gradually up-sample the initial alpha maps 4 times by our multiscale guided upsampling module to achieve \(8\times\) upsampling. For the multiscale guided upsampling module, we adopt residual blocks and nearest neighbour interpolation for gradually upsampling while the guidance is gradually downsampled by residual blocks and convolution with stride=2. As for the fusion module, we use cascaded residual blocks to extract HR features of warped \(I_{Ref}\) conditioned on \(I_{LR}\) and merge these features with the upsampled blurry input to produce the final SR result.

During the training, we set the spatial size \(h = 384, w = 768, c = 3, d = 32, \beta = 8\) and use the train set of the RealEstate10K [38] dataset. The weights for \(L_{rec}, L_{per}, L_{is}\) are 1, 1 and 0.2 respectively. We train the proposed network in three steps. Specifically, we firstly pre-train initial alpha maps with internal supervision loss, secondly add the multiscale guided upsampling module to warp the reference image and finally train the whole network together with the all losses. The network is trained with ADAM solver [13] for 816.2k, 326k and 472k iterations with learning rate 0.0002, \(\beta_1 = 0.9, \beta_2 = 0.999\) and batch size = 1 on a single TITAN X GPU.

4. Experiments

In this section, we introduce two kinds of data from digital synthesized cross-scale data to optical zoomed cross-scale data for quantitative and qualitative comparisons to verify the superior performance of the proposed Cross-MPI. The description of data is in Sect. 4.1, the comparison results of digital synthesized and optical zoomed cross-scale data is presented in Sect. 4.2 and Sect. 4.3 separately. The ablation study is then presented in Sect. 4.4.

4.1. Datasets

The first type of data is the digital synthesized cross-scale data where the LR images are generated by bicubic downsampling. We use RealEstate10K [38] dataset to generate cross-scale data for training (59637 sequences) and testing (6546 test sequences). We obtain the LR images through bicubic downsampling \((8\times)\). And the second type of data is the optical zoomed cross-scale data where the HR reference images are generated by optical zoom. Specifically, we collect a Optical Zoom-in dataset with a motorized zoom camera for training and testing, and a

| Algorithm    | frame diff. = 9 | frame diff. = 11 | frame diff. = 15 | frame diff. = 19 | frame diff. = 23 |
|--------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Bicubic      | 25.016 / 0.733  | 25.017 / 0.733  | 25.052 / 0.735  | 25.044 / 0.734  | 25.052 / 0.734  |
| MDSR [15]    | 27.424 / 0.808  | 27.434 / 0.808  | 27.490 / 0.810  | 27.480 / 0.810  | 27.486 / 0.810  |
| RCAN [32]    | 27.846 / 0.822  | 27.855 / 0.822  | 27.918 / 0.824  | 27.907 / 0.824  | 27.908 / 0.824  |
| ESRGAN [26]  | 25.555 / 0.743  | 25.563 / 0.743  | 25.623 / 0.746  | 25.634 / 0.746  | 25.633 / 0.746  |
| RankSRGAN [31]| 25.244 / 0.692  | 25.243 / 0.692  | 25.304 / 0.694  | 25.297 / 0.694  | 25.306 / 0.694  |
| SPSR [17]    | 25.200 / 0.708  | 25.201 / 0.708  | 25.263 / 0.710  | 25.252 / 0.710  | 25.258 / 0.710  |
| CrossNet-8× [37]| 32.914 / 0.933  | 31.509 / 0.911  | 30.610 / 0.893  | 29.928 / 0.878  | 29.441 / 0.865  |
| SRNTT-8× [34]| 24.485 / 0.758  | 24.306 / 0.749  | 24.254 / 0.745  | 24.170 / 0.740  | 24.115 / 0.737  |
| TTSR-8× [28] | 31.913 / 0.923  | 31.002 / 0.908  | 30.383 / 0.896  | 29.843 / 0.884  | 29.405 / 0.874  |
| Cross-MPI    | 32.878 / 0.937  | 31.736 / 0.921  | 30.993 / 0.909  | 30.360 / 0.896  | 29.852 / 0.885  |
Ground Truth | RCAN [32] | RankSRGAN [31] | CrossNet [37] | Ours
---|---|---|---|---
Parallax
SRNTT [34]
TTSR [28]
Ours

Figure 6: Visual evaluation (8×) on the RealEstate10K dataset [38] (top) and Optical Zoom-in dataset (down).

**Cross-Scale Stereo dataset** collected by a real cross-scale stereo camera setup. In the optical zoom-in dataset, we collected 15 scenes while each scene contains 30 to 60 content-matched image pairs (with resolution difference about 10×) along with their camera parameters and poses, which are further divided into a train set (13 scenes) and a test set (2 scenes). In the cross-scale stereo dataset, we obtained a set of stereo pairs with focal length “16mm-135mm” from different views of the scene, introducing not only resolution gaps but also parallaxes between stereo pairs. More details of the camera configurations and calibration procedures can be found in the supplementary.

### 4.2. Comparisons on Digital Synthesized Data

In this section, we quantitatively compare the performance of Cross-MPI against the latest RefSR [37, 34, 28] and SISR [15, 32, 26, 31, 17] methods on RealEstate10K [38] dataset. To be fair, we re-train Cross-Net, SRNTT, TTSR following the procedure according to their papers on the train set of RealEstate10K. Since SRNTT and TTSR do not have 8× models, we add a 2× subpixel upsampling layer at the end of main network. As for the SISR work [15, 26, 31, 17] without 8× models, we use 4× models twice and downsample 2× to achieve 8× magnification on test images. Moreover, the SR spatial resolution of train set and test set is set to (384,768) to be consistent with our training procedure.

Table. 1 shows quantitative comparisons between Cross-MPI against SISR and RefSR methods. We test cross-scale image pairs for different frame intervals between cross-scale image pairs. Cross-MPI achieves the highest score on RealEstate10K dataset for 8× magnification compared with the tested SISR methods and Methods are grouped by SISR approaches (top) and RefSR approaches (bottom) under large parallax (frame differences from 11 to 23). When the parallax decreases, the PSNR result is similar with CrossNet which is suitable for small parallax. However, our SSIM is higher because Cross-MPI can preserve the underneath scene structure and generate SR results with better structural performance. Moreover, Cross-MPI is also
Table 2: Test results on Optical Zoom-in dataset.

| Algorithm      | PSNR/SSIM     |
|----------------|--------------|
| CrossNet-8×    | 27.830 / 0.798 |
| SRNTT-8×       | 24.762 / 0.687 |
| TTSR-8×        | 27.114 / 0.787 |
| Cross-MPI      | **29.209 / 0.841** |

4.3. Comparisons on Optical Zoomed data

We also compare our model with the state-of-art RefSR models [37, 34, 28] on our Optical Zoom-in dataset. In detail, we fine-tune the pre-trained 8× models on the train set (13 scenes) for 2 epochs and test on the test set (2 scenes). Table 2 gives the numerical results and last two cases in Fig. 6 give the visual results. Fully utilizing the concealed scene structure, Cross-MPI performs the best numerically and visually on Optical Zoom-in dataset which has more complicated degradation model than digital synthesized cross-scale data. To further testify the effectiveness of Cross-MPI for real hybrid cross-scale data, we build a cross-scale stereo camera prototype, referring to Fig. 7.

4.4. Ablation Study

In this section, we will verify the effectiveness of different modules in our Cross-MPI network, including the plane-aware attention, the multiscale guided upsampling module as well as the internal supervision loss. We train models with different combinations of these modules. Randomly sampling stereo pairs with frame difference from 3 to 10 as the test set, we calculate the quantitative metrics recorded in Table 3 and give an example of the rendered depth of different models in Fig. 8.

**Plane-Aware Attention.** In order to testify that the novel plane-aware attention improves the correspondence matching in the whole pipeline, we replace the plane-aware attention with “concatenation + convolution” while the other parts of the network as well as the training procedure remain the same as our full model. As we can see from Table 3 and Fig. 8, the numerical result slightly descends and some incorrect estimation appears without the confident initialization of the plane-aware attention.

**Multiscale Guided Upsampling Module.** The design of multiscale guided upsampling effectively helps to superresolve the initial alpha maps. To verify the effectiveness of the guided upsampling, we remove the external guidance path (the lower branch in Fig. 5) and the upsampling procedure becomes a pure convolutional upsampling. From Fig. 8, we can see the depth becomes blurry without the delicately designed guidance.

**Internal Supervision Loss.** The Internal Supervision Loss helps the network to learn a better initial estimation of the alpha maps which is crucial for the latter SR synthesis and fusion procedure. The numerical and visual results decrease without the IS Loss as we can see from Table 3 and Fig. 8.

5. Conclusion

This paper explores the RefSR problem for real multi-scale camera systems. We propose Cross-MPI, an end-to-end network that takes the underneath scene structures as a clue to transfer HR details from one view to the other under a large resolution gap, e.g., 8×. The proposed Cross-MPI consists of a plane-aware attention-based MPI module to explicitly and effectively captures correspondences along depth planes, a multiscale guided upsampling module to estimate correspondences in a coarse-to-fine manner with elaborately designed LR correspondence estimation and guided upsampling mechanism, and a SR synthesis and fusion module to adequately transfer HR details and generate the final SR result. Extensive experiments show the superior performance of Cross-MPI over the state-of-art in both digital synthesized and optical zoomed cross-scale datasets. We will further explore the Cross-MPI structure for other applications in hybrid multiscale camera systems.
References

[1] Vivek Boominathan, Kaushik Mitra, and Ashok Veeraraghavan. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In Proceedings of the IEEE International Conference on Computational Photography, 2014. 1, 2, 3

[2] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In Proceedings of the IEEE International Conference on Computer Vision, 2017. 6

[3] Chao Dong, Chen Change Loy, Kaiming He, and Xiaou Tang. Learning a deep convolutional network for image super-resolution. In Proceedings of the European Conference on Computer Vision, 2014. 2

[4] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In Advances in Neural Information Processing Systems, 2016. 6

[5] Matthew DuVall, John Flynn, Michael Broxton, and Paul Debevec. Compositing light field video using multiple images. In SIGGRAPH, pages 1–2. Association for Computing Machinery, 2019. 2

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 2, 4

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, 2014. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2

[9] iPhone 11 Pro. https://www.apple.com/iphone-11-pro/. 1

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision, 2016. 2, 6

[11] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2

[12] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations, 2015. 6

[14] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4681–4690, 2017. 1, 2, 6

[15] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017. 6, 7

[16] Bee Lim, Sanghyun Son, Heeowon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017. 1, 2

[17] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 1, 3, 6, 7

[18] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (TOG), 38(4):1–14, 2019. 2

[19] HUAWEI P40 Pro+. https://consumer.huawei.com/en/phones/p40-pro-plus/. 1

[20] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In Proceedings of the IEEE International Conference on Computer Vision, 2017. 2

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. Proceedings of the International Conference on Learning Representations, 2015. 6

[22] Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiple plane images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 2

[23] Richard Tucker and Noah Snavely. Single-view view synthesis with multiple plane images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 2

[24] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 2, 4, 6

[25] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 2

[26] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision, 2018. 2, 6, 7

[27] Yuwang Wang, Yebin Liu, Wolfgang Heidrich, and Qionghai Dai. The light field attachment: Turning a dslr into a light field camera using a low budget camera ring. IEEE Transactions on Visualization and Computer Graphics, 23(10):2357–2364, 2016. 1, 2, 3
[28] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Bain- 
ing Guo. Learning texture transformer network for image super-resolution. In Proceedings of the IEEE Conference on 
Computer Vision and Pattern Recognition, 2020. 1, 2, 3, 4, 
6, 7, 8

[29] Xiaoyun Yuan, Lu Fang, Qionghai Dai, David J. Brady, and 
Yebin Liu. Multiscale gigapixel video: A cross resolution 
image matching and warping approach. In Proceedings of 
the IEEE International Conference on Computational Pho-
tography, 2017. 1, 3

[30] Huanjing Yue, Xiaoyan Sun, Jingyu Yang, and Feng Wu. 
Landmark image super-resolution by retrieving web images. 
IEEE Transactions on Image Processing, 22(12):4865–4878, 
2013. 3

[31] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. 
Ranksrgan: Generative adversarial networks with ranker for 
image super-resolution. In Proceedings of the IEEE Interna-
tional Conference on Computer Vision, 2019. 3, 6, 7

[32] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng 
Zhong, and Yun Fu. Image super-resolution using very deep 
residual channel attention networks. In Proceedings of the 
European Conference on Computer Vision, 2018. 1, 2, 6, 7

[33] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and 
Yun Fu. Residual dense network for image super-resolution. 
In Proceedings of the IEEE Conference on Computer Vision 
and Pattern Recognition, 2018. 1, 2

[34] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Im-

age super-resolution by neural texture transfer. In Proceed-
ings of the IEEE Conference on Computer Vision and Pattern 
Recognition, 2019. 2, 3, 6, 7, 8

[35] Mandan Zhao, Gaochang Wu, Yipeng Li, Xiangyang Hao, 
Lu Fang, and Yebin Liu. Cross-scale reference-based light 
field super-resolution. IEEE Transactions on Computational 
Imaging, 4(3):406–418, 2018. 2, 3

[36] Haitian Zheng, Mengqi Ji, Lei Han, Ziwei Xu, and Lu Fang. 
Learning cross-scale correspondence and patch-based syn-
thesis for reference-based super-resolution. In Proceed-
ings of the British Machine Vision Conference, 2017. 2, 3

[37] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu 
Fang. Crossnet: An end-to-end reference-based super reso-
lution network using cross-scale warping. In Proceedings of 
the European Conference on Computer Vision, 2018. 1, 2, 3, 
6, 7, 8

[38] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, 
and Noah Snavely. Stereo magnification: Learning view syn-
thesis using multiplane images. SIGGRAPH, 2018. 2, 3, 6, 
7