

# Learning Cross-scale Correspondence and Patch-based Synthesis for Reference-based Super-Resolution

Haitian Zheng<sup>1</sup>  
zheng.ht.ustc@gmail.com

Mengqi Ji<sup>3</sup>  
mji@ust.hk

Haoqian Wang<sup>12</sup>  
wanghaoqian@tsinghua.edu.cn

Yebin Liu<sup>4</sup>  
liuyebin@mail.tsinghua.edu.cn

Lu Fang<sup>1</sup>  
fanglu@sz.tsinghua.edu.cn

<sup>1</sup> Tsinghua University  
Graduate School at Shenzhen

<sup>2</sup> Shenzhen Institute of Future Media  
Technology

<sup>3</sup> Dept. of ECE  
Hong Kong Univ. of Science and Tech.

<sup>4</sup> Dept. of Automation  
Tsinghua University

## Abstract

In this paper, we explore the Reference-based Super-Resolution (RefSR) problem, which aims to super-resolve a low definition (LR) input to a high definition (HR) output, given another HR reference image that shares similar viewpoint or capture time with the LR input. We solve this problem by proposing a learning-based scheme, denoted as RefSR-Net. Specifically, we first design a Cross-scale Correspondence Network (CC-Net) to indicate the cross-scale patch matching between reference and LR image. The CC-Net is formulated as a classification problem which predicts the correct matches from the candidate patches within the search range. Using dilated convolution, the training and feature map generation are efficiently implemented. Given the reference patch selected via CC-Net, we further propose a Super-resolution image Synthesis Network (SS-Net) for the synthesis of the HR output, by fusing the LR patch and the reference patch at multiple scales. Experiments on MPI Sintel Dataset and Light-Field (LF) video dataset demonstrate our learned correspondence features outperform existing features, and our proposed RefSR-Net substantially outperforms conventional single image SR and exemplar-based SR approaches.

## 1 Introduction

Reference-based super-resolution (RefSR) is a new image super-resolution category appeared recently, which aims to super-resolve a low definition (LR) input image to a high definition (HR) output, with a given HR reference image which shares similar viewpoint or capture time with the LR input [0, 52, 41]. Because the given reference is very similar to the high definition ground-truth of the LR input, the scale difference between the high and low definition images is usually set to more than  $8\times$ . This kind of cross-scale image

super-resolution is crucial for compressive sensing of high definition visual contents, and has been successfully demonstrated in light field reconstruction [0, 5, 24, 54] and gigapixel video synthesis [40].

Even though a similar HR reference image is available, RefSR is still not a trivial task because of two reasons. First, a precise correspondence between the LR-HR patch pairs is highly desired for high quality super-resolution. In available RefSR [0, 55] or example-based super-resolution methods [0, 4, 50, 51], the HR patches in the dictionary are required to be down-sampled to match the features of LR patches. Such over-simplified feature extractor loses some valuable information for building the correspondence between the LR-HR patch pairs.

Second, based on the selected LR-HR correspondences, the RefSR should subsequently synthesize realist HR output. Available methods select multiple best candidates from the HR patch dictionary for linear combination. If quite a lot candidates are considered, the fusion result will be too blurry to be perceptually real. If only one candidate is used for SR, such method just simply use this HR patch as output prediction. That could lead to severe blocky artifact, since it rarely happens that the exact desired HR patch can be found in the dictionary.

Against this backdrop, this paper proposes a learning-based cross-scale correspondence scheme and a learning-based patch synthesis strategy for reference-based super-resolution. For the correspondence scheme, we propose a robust Cross-scale Correspondence Network (CC-Net) to learn the cross-scale correspondence without tedious hand-engineering steps and without the down-sampling operation. Specifically, to align the LR input with the HR reference, dense correspondence learning step is utilized for patch matching. Moreover, to tackle the occlusion and mistake LR-HR matching, a feature masking is proposed in a novel unsupervised fashion. For the super-resolution synthesis strategy, we propose a Super-resolution image Synthesis Network (SS-Net) which takes advantage of the recent single image SR (SISR) method [0, 17, 18, 21] providing strong internal statistics that could be helpful to produce reasonable HR result with erroneous external HR samples. Quantitatively and qualitatively, the experiments show the superior performance of these two networks compared to the previous methods.

## 2 Related work

Image super-resolution is a typical inverse graphics problem. In literature, parametric approaches and exemplar-based approaches provide different insight into this problem. exemplar-based approaches rely on external dictionary for super-resolution. Specifically, [12] apply nearest neighbour search on exemplar dictionary for generating high-resolution patches. [9] applies manifold embedding for SR reconstruction. [0, 5, 50, 51] consequently applied varieties of manifold embedding schemes for linearly combining HR exemplar patches.

Recent works [0, 5, 24, 54, 55] further uses exemplar-base SR to handle the problem of SR with reference image, which forms a new kind of super-resolution method using an explicit reference. Especially, [0] apply the K-NN search [9] and non-local mean [5] for combining patches. [54] further applied an iterative step for enriching the exemplar database. [55] applied patch registration before nearest neighbor searching, and applied dictionary learning for reconstruction.

Parametric approaches usually formulate SR process as a function which maps a LR patch to a HR patch, and learn such mapping function by different function modeling. Specifically, simple function [36, 40], decision tree [27], random forests [27, 28], Gaussian

process regression [13] are proposed to predict HR patches given LR patches. In addition, Sparse dictionary [19, 22, 28, 29, 33] is also used in the parametric way for super-resolution.

Recent advances in deep learning further boost the performance of parametric SR. [7, 8] proposed a three layers convolutional neural network (CNN) for predicting HR image with LR input. Furthermore, [9, 29] show that using transposed/sub-pixel convolution to learn upsampling filter help to improve performance both in speed and accuracy. [18] proposed a 20-layers deep CNN for predicting the bicubic upsampling error. [10] proposed deeply-recursively convolutional network. Recently, [20] apply the residue network [24] for single image super-resolution, which achieves the current state-of-the art results. Additionally, [16, 26] also applied the residue network for image super-resolution.

With an explicit reference, the available exemplar-based approaches utilizes the explicit, task-related external dataset for super-resolution, thus they are able to achieve superior results compared to parametric SR approaches. Compared with the previous exemplar-based and parametric approaches, the proposed *Super-resolution Synthesis Network* in our RefSR can not only produce high quality HR results with the LR-HR patch pairs, but also generates reasonable HR patches even with erroneous HR patch reference.

Finding visual correspondence is crucial for many applications such as optical-flow or stereo-matching. For exemplar-based super-resolution, finding the cross scale patch correspondence (i.e., finding correspondence between LR patches and reference HR patches) is also crucial. [9] proposed a widely used gradient feature for matching exemplar with an input LR patches. Specifically, the gradient feature of LR patches are extracted for measuring  $L_2$  distance similarities between input and exemplar reference from external dictionary. This matching solution is further used in single-image SR (SISR) [1, 30, 31], and SR for specific system including hybrid imaging systems [2, 32, 35]. In contrast, our corresponding network automatically learns feature embedding which is robust to occlusion and view parallax.

### 3 Cross-scale Correspondence Network (CC-Net)


In our RefSR-Net, the image correspondence between the HR reference image and the LR target image is critical. While image correspondence that is usually formulated as feature matching has been well studied in various works [10, 32, 33], existing approaches which try to minimize average endpoint error metrics or average  $L_2$  distance error cannot serve the end of super-resolution. Because with the increase of matching error distance, the correspondence patch provides much less information (such as high-frequency details) for SR reconstruction.

Instead of minimizing the average  $L_2$  distance error, we regard the feature learning as a classification problem. Given an input LR patch and a larger area of reference patch, our network outputs the feature (feature map) of them for computing a matching probability map, denoting the matched sub-patch in the reference patch. Concretely, we apply two dilated convolutions networks on the bi-linearly up-sampled LR image and HR image for extracting pixel-level dense feature.

Note that unlike typical dense prediction networks with skip-connection [10, 23], a dilated convolution network does not contain downsampling/up-sampling building block which incurs spatial ambiguity. Thus, it is able to achieve precise translation equivariance, which is beneficial for extracting spatial sensitive features maps. The two dilated networks share the same network structure and parameters. As shown in Table 1, our network contains six dilated convolution layers and one final  $1 \times 1$  convolutional layer. Batch normalization

Layer	Kernel	Channel	Dilation	Receptive field
conv1 + BN	$3 \times 3$	16	1	$3 \times 3$
conv2 + BN	$3 \times 3$	32	1	$5 \times 5$
conv3 + BN	$3 \times 3$	64	2	$9 \times 9$
conv4 + BN	$3 \times 3$	64	4	$17 \times 17$
conv5 + BN	$3 \times 3$	64	8	$33 \times 33$
conv6 + BN	$3 \times 3$	128	16	$65 \times 65$
fc	$1 \times 1$	256	1	$65 \times 65$

Table 1: The network architecture for feature extraction.




[,] is used for the dilated convolution layers. In Section 5.1, we also explore different strategies for upsampling the LR inputs and different weight sharing strategies.


In training stage, given a bi-linearly up-sampled LR image and the HR reference image, we first randomly sample central position  $(x, y) \in \{(8*m + 1, 8*n + 1) | m, n \in 0, 1, 2, \dots\}$  for cropping the 65 LR patch. With correspondence ground truth, a  $192 * 192$  patch from HR reference image centered at the corresponding location is selected. Afterward, LR feature  $\mathbf{f}$  and a  $129 \times 129$  feature maps  $\mathbf{g}_{i,j}$  are computed. Afterwards, inner-product is used for computing the correspondence matching score  $s_{i,j} = \langle \mathbf{f}, \mathbf{g}_{i,j} \rangle$ . Consequently, a 2-D softmax layer is applied to compute the matching probability  $p_{i,j} = e^{s_{i,j}} / \sum_{i,j} e^{s_{i,j}}$ . Finally, the loss function is defined by

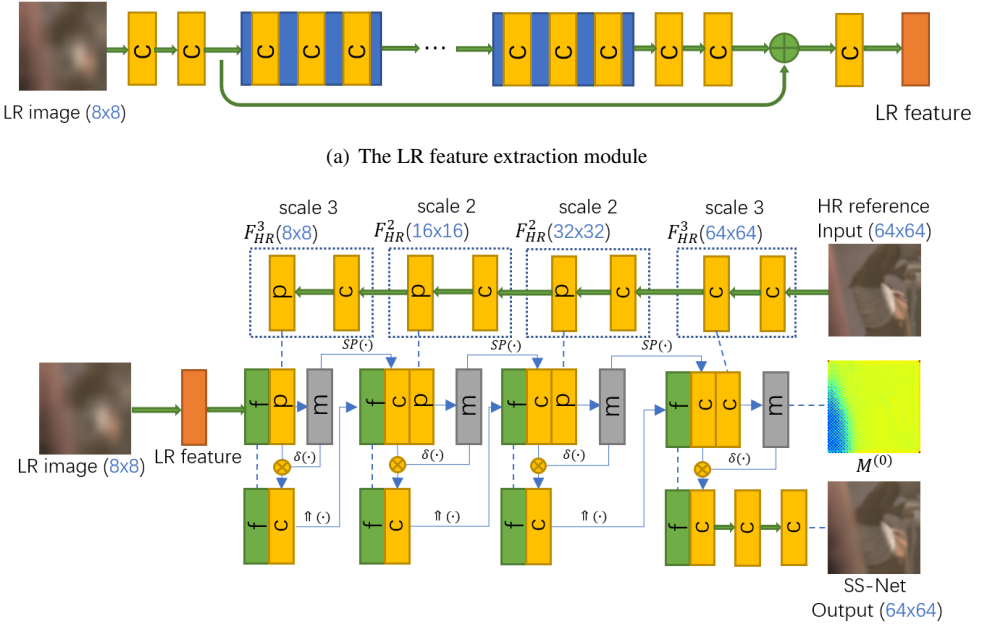
$$L = - \sum_{i,j} q_{i,j} \log p_{i,j}, \quad (1)$$

which denotes the cross entropy between  $p$  and  $q$ , and  $q_{i,j}$  is the one-hot matrix that represents the ground-truth flow. Regarding the testing stage, the correspondence is calculated by finding the maximum inner product response  $s$  between the feature of a LR patch and its corresponding HR features. However, we use the search range of 100 instead of 64 in testing stage, which helps to deal with large displacement.

## 4 Super-resolution Synthesis Network (SS-Net)

Recent works on single image SR using deep learning, such as [, , ,] have pushed the performance of SR a great leap. However, RefSR with deep learning still remains unexplored. In this section, we propose a fusion network, denoted as Super-resolution Synthesis Network (SS-Net), which takes advantages of both the state-of-the-art single image super-resolution techniques and available information from a corresponding reference patch extracted using the trained CC-Net. Specifically, as shown in Fig. 1, SS-Net consists of a *LR Feature Extraction Module* which is pre-trained on ImageNet, a *HR Feature Extraction Module* for extracting multiple scales feature from the HR reference patches, and a *prediction module* which fuses the LR image features with reference image features for SR prediction.

*LR Feature Extraction Module* encodes the representation of LR patch. As shown in Figure 1(a), our LR feature extraction module is derived from SRResNet [,]. Concisely, the SRResNet relies on a sequential of 16 convolutional residue blocks following skip connection for extracting low-resolution feature map, and  $2 \times 2$  upsampling block for outputting SR image. Different from SRResNet, we replace the sub-pixel upsampling layer with the nearest neighbour upsampling layer (denoted by SRResNet + NN). The network is pretrained on ImageNet for  $\times 4$  super-resolution following the same setting of *SRResNet*. Afterward the



(b) The HR feature encoding module (upper part) and the prediction module (lower part).

Figure 1: Illustration of Super-resolution Synthesis Network (SS-Net).

weights are frozen, and our LR feature extraction module utilizes the pretrained SRResNet + NN (without upsampling blocks) for extracting the feature representation of LR patch. Note that in principle the LR feature extraction module in SS-Net can be directly trained without pretraining. However, since the model capacity of LR feature extraction network (16 residue blocks) is too large for the MPI Sintel dataset and LF Video Dataset, we instead use the large ImageNet to train the LR feature extraction network.

*HR Feature Extraction Module* consists a convolutional layer following three feature extraction blocks, in which a convolution layer is followed by a stride 2 max pooling layer. Given an input high-resolution patch  $x_{HR}$ , the output feature map of the first convolutional layer  $F_{HR}^{(0)}$ , and the feature map of each conv + pooling block  $F_{HR}^{(1)}, F_{HR}^{(2)}, F_{HR}^{(3)}$  extract HR image feature at resolution scale  $\{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$  respectively. The HR Feature Extraction Module is illustrated in the upper part of Figure 1(b).

After the extraction of the LR feature  $F_{LR}$  and the multiple scale HR features  $F_{HR}^{(0)}, F_{HR}^{(1)}, F_{HR}^{(2)}, F_{HR}^{(3)}$ , the *Prediction module* subsequently performs feature fusion from scale 3 to scale 0, and output the final SR image prediction. Specifically, the fused feature map at scale  $i$  (for  $i = \{0, 1, 2, 3\}$ ) is defined by

$$\begin{cases} M_L^{(3)} = \text{relu}((F_{LR}, F_{HR}^{(3)}) * W_{mask}^{(3)}), & M^{(3)} = \text{sig}(M_L^{(3)}), \\ F_{fuse}^{(3)} = \delta((F_{LR}, F_{HR}^{(3)}) \odot M^{(3)}) * W_{fuse}^{(3)}, \\ M_L^{(i)} = \text{relu}((SP(F^{(i+1)}), F_{HR}^{(i)}, \uparrow(M_L^{(i+1)})) * W_{mask}^{(i)}), & M^{(i)} = \text{sig}(M_L^{(i)}), \\ F_{fuse}^{(i)} = \delta((\uparrow(F^{(i+1)}), F_{HR}^{(i)}) \odot M^{(i)}) * W_{fuse}^{(i)}), & i = \{0, 1, 2\} \end{cases} \quad (2)$$

where at scale  $k$ ,  $M_L^{(k)}$  is the 2-D mask confidence for masking HR feature  $F_{HR}^{(k)}$ , and  $F_{fuse}^{(k)}$

represents the fused feature map. The operation  $*$  stands for convolution and  $\odot$  stands for element wise matrix multiplication.  $\delta(\cdot)$  and  $\text{relu}(\cdot)$  denote the sigmoid function and the rectified linear function.  $SP(\cdot)$  is  $\times 2$  sub-pixel upsampling functions [24] that are parameterized by trainable convolutional kernels, and the  $\uparrow(\cdot)$  stands for  $\times 2$  nearest neighbour upsampling function. The lower part of Figure 1(b) gives a detailed illustration of the prediction module.

As the SS-Net employs a patch-wise prediction scheme, indicating that for every overlapping  $8 \times 8$  LR patch the prediction network outputs a  $64 \times 64$  HR prediction, the *Sliding Windows Fusion* is necessary. In [2], an simple averaging is performed among the overlapping patches. Following the same setting, we test this fusion strategy for our prediction network. However, the weighted averaging is utilized for our fusion scheme. Specifically, the intermediate mask at scale 0,  $M^{(0)}$  is designed for masking features in occlusion regions, where the values near 0 indicate occlusion (As illustrated in Figure 1(b)). Therefore, we additionally employ a weighted averaging scheme which utilizes  $M^{(0)}$  for indicating the occlusion regions. More specifically,  $M_i^{(0)}$  represents the mask produced by LR patch  $i$ . We first smooth the masks with Gaussian kernel  $\mathcal{G}(\sigma)$  to produce the confidence weights  $O_i = M_i^{(0)} * \mathcal{G}(3)$ . For every pixel in the output image, weighted averaging is applied using the weights of the smoothed masks  $O_i$ . The radius of the smoothing kernel is set to  $\sigma = 3$  empirically as 1 pixel in  $F_{fuse}^{(0)}$  correspond a receptive field of size  $5 * 5$  in final predictions.

## 5 Experiment

**Dataset** To learn robust feature representation, it is crucial that the training correspondence contain rich variations, such as lighting, geometric changes and occlusions. In this paper, we use MPI Sintel Flow Dataset for training the CC-Net and the SS-Net. The MPI Sintel training set contains 1041 image frames from 23 different clip videos with optical flow ground truth. The MPI Sintel testing set contains a total of 564 frames from 12 clips. We further divide the training set into a subset of 956 images for model training and a subset of 85 images for validation. Data augmentation by randomly rotating the image by  $\{0, 90, 180, 270\}$  degree is applied. Furthermore, by randomly offsetting the HR image by  $(i, j)$  pixel ( $0 \leq i, j < 8$ ) and then perform the down-sampling, the HR image with its corresponding LR image is further augmented by a factor of 64. The HR frame is taken as reference image, while the LR version of the next frame is taken as input image, for finding correspondence and super-resolution.

### 5.1 Cross-scale Correspondence Network (CC-Net)

**Training** For training the correspondence network, the LR images is bi-linearly upsampled to the same scale as the HR reference image. During training,  $65 \times 65$  patches are random sampled from the upsampled LR images. Using the optical-flow ground truth, the larger  $193 \times 193$  patches from reference images are sampled accordingly. The 2-D softmax layer output  $129 \times 129$  matrices which represent the matching probability.

All the correspondence network are trained by Adam [20] with mini-batch size 16 and learning rate of 0.001 for the first 30K iterations. The learning rate is optionally set to 0.00005 to further train the network for additional 20K iterations.

**Different training schemes** Figure 2 compares the impact of different training schemes. In the figure, the **Siamese** represents when weight sharing is applied on two streams of net-

works, where as the **Pseudo-siamese** represents when two streams of networks do not share weights. Different interpolation schemes for interpolating the LR input, such as **bilinear** and **SRResNet + NN + sliding** (as discussed in Section 5.2) are also tested. Figure 2 shows siamese network outperforms the pseudo-siamese network. We also found that despite SR-ResNet + NN + sliding output interpolation reconstruction with much higher precision in terms of PSNR measurement, such output is not suitable for the cross-scale correspondence task. That is because some spatial clue for learning robust correspondence is lost.

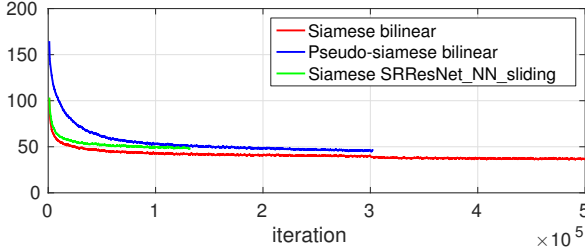


Figure 2: Training loss comparison for different weight sharing schemes and input schemes. The horizontal and vertical axis represent training iteration and training error respectively.

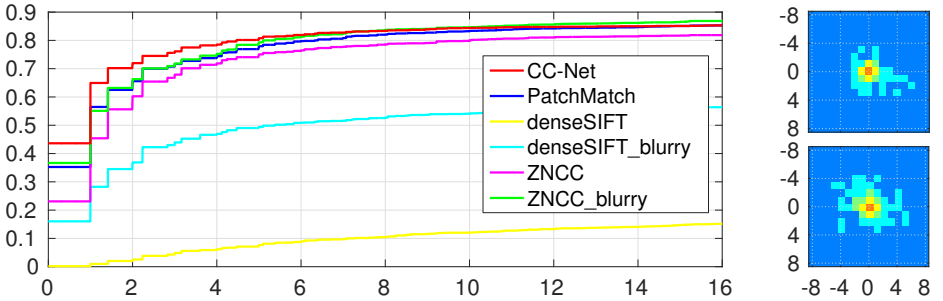


Figure 3: Left: Cross-scale matching accuracy under different threshold error distance on MPI dataset. The horizontal and vertical axis represent threshold error distance and matching accuracy respectively. Right: Matching visualization depicts the offset (in pixels) of matched position from ground-truth position, upper: CC-Net, lower: PatchMatch [4]

**Evaluation** We validate the performance of our cross correspondence network (denoted by **CC-Net**) by computing the matching accuracy with different threshold error distance. In comparisons, a widely used feature [4] (denoted by **PatchMatch**) for exemplar-based SR [2], the **denseSIFT** feature [22], the **ZNCC** feature [11] is also evaluated. Additionally, we explored a the variant of dense SIFT feature and ZNCC feature, which take the bicubic downsampled and upsampled reference image as input, instead of the HR reference image (denoted by **denseSIFT\_blurry** and **ZNCC\_blurry**).

Figure 3 compares different schemes by showing accuracy changing with the different threshold error distance. As showed in Figure 3, CC-Net substantially outperforms the rest approaches when threshold error distance are in range of  $[0, 6]$  pixels. We also the observe denseSIFT\_blurry and ZNCC\_blurry improves the results of denseSIFT and ZNCC by large margin; ZNCC\_blurry slightly outperform PatchMatch feature. The matching accuracy of



all approaches under threshold error distance in range  $[0, 7]$  are shown also in Table 2.<sup>1</sup> For finding exact matching (threshold error distance = 0), CC-Net outperforms the second best approaches by 6.9.

Methods	$\leq 0$	$\leq 1$	$\leq 2$	$\leq 4$	$\leq 5$	$\leq 6$
denseSIFT [22]	0.2%	1.0%	2.6%	6.1%	7.65%	8.92%
ZNCC [14]	23.1%	45.4%	60.3%	71.9%	75.1%	76.5%
denseSIFT_blurry	16.0%	28.2%	36.9%	47.0%	49.4%	50.9%
ZNCC_blurry	36.7%	55.1%	66.3%	75.3%	80.3%	81.3%
PatchMatch [4]	35.3%	56.5%	65.6%	74.4%	77.9%	79.8%
<b>CC-Net (ours)</b>	<b>43.6%</b>	<b>65.0%</b>	<b>72.0%</b>	<b>78.5%</b>	<b>81.3%</b>	<b>81.9%</b>

Table 2: Cross-scale correspondence matching accuracy on MPI Sintel validation set under different threshold error distance.

The mismatching pattern of CC-Net and PatchMatch is also shown in the right part of Figure 3, where the horizontal and vertical axis represent offset from ground-truth matching position respectively, and color represent matching point density (blue to red means small to large density). The CC-Net has more condensed matching pattern.

## 5.2 Super-resolution Synthesis Network (SS-Net)

**SISR network** We use the **SRResNet** [24] trained on ILSVRC2012 [25] training set as the baseline network for extracting LR feature. Apart from it, we replace the sub-pixel convolution layers in SRResNet with x2 nearest neighbor upsampling layers for feature upsampling, which result in an additional SISR network (denoted as **SRResNet + NN**). In additional, we also test the sliding windows version of SRResNet\_NN (**SRResNet + NN + sliding**) with stride size = 2, and patch size = 24.

Single image based Methods	Sintel	LF video
bicubic	30.59	30.32
SRCNN [2]	32.93	31.24
VDSR [18]	33.59	31.65
SRResNet [24]	33.15	31.40
SRResNet [24] + NN	33.68	31.72
SRResNet [24] + NN + sliding	<b>33.75</b>	<b>31.80</b>
Exemplar based Methods	Sintel	LF video
PatchMatch [4]	35.72	37.96
<b>RefSR (CC-net+SS-Net, ours)</b>	<b>38.03</b>	<b>38.88</b>

Table 3: Average PSNR comparison of 552 images in MPI Sintel testing set and 268 images in LF video testing set.

The SR reconstruction result on MPI Sintel testing set is shown in the lower part of Table 3. In comparisons, SR reconstruction result of **bicubic**, **SRCNN** [2], **VDSR** [18] is also listed. The table clearly shows SRResNet + NN based approach (33.68 and 33.75) outperforms the rest ones (VDSR 33.59). This is because as SRResNet based network utilize residue network for learning deeper representation.

On MPI Sintel dataset, the SRResNet + NN outperforms SRResNet. This is probably due to the images in MPI Sintel dataset are less sharper compared to ImageNet images, and

<sup>1</sup>Note when error distance becomes large, reference patch provides much less information for SR reconstruction. The matching accuracy under large threshold error distance is therefore not shown in Table 2.



the SRResNet + NN trend to output less sharper feature map and output image compared to SRResNet. The sliding window version of SRResNet + NN slightly outperform the full convolutional SRResNet + NN. We conjecture that for SRResNet + NN, statics discrepancy between training and testing exists due to the zero padding is applied on different size of inputs, where as the SRResNet + NN + sliding remedy such statics discrepancy.

Besides, we also test our result on LF video dataset [53]. The LF video dataset contains 1080 light field images for training and 270 light field images for testing. Unlike MPI Sintel, the LF video dataset does not provide the ground truth correspondence. Therefore, we directly apply the CC-Net that is trained on MPI Sintel dataset for extracting HR reference patches, and then use the LF video dataset to train the SS-Net. For training the SS-Net, we randomly select LR and reference image from the  $8 \times 8$  LF grid at different positions. While for testing, we select LR image at position (3,3) and reference image at (0,0) for RefSR. As shown in Table 3, we also observe improvement of SRResNet + NN + sliding over SRResNet + NN and SRResNet on the LF video dataset. Furthermore, our RefSR achieves 28.72 average PSNR, outperforming PatchMatch by a margin of 0.92 on this real image dataset.

**Training** Our network is trained with  $64 \times 8 \times 8$  LR feature of SRResNet + NN + sliding, a  $64 \times 64$  correspondence HR patch and a  $64 \times 64$  HR ground truth as input. In practice, randomly generating samples is inefficient, as it requires computing correspondence feature map for every  $65 \times 65$  upsampled LR patch and  $265 \times 265$  reference patch (we use 100 as search range). In our experiment, we compute 4 correspondence feature maps for every 100 iterations, where as for each iteration, we randomly sample LR patches and compute corresponding HR patches using the pre-computed correspondence feature maps. The network is trained with Adam [20] with mini-batch size 8, learning rate 0.0001 for 30K iterations.

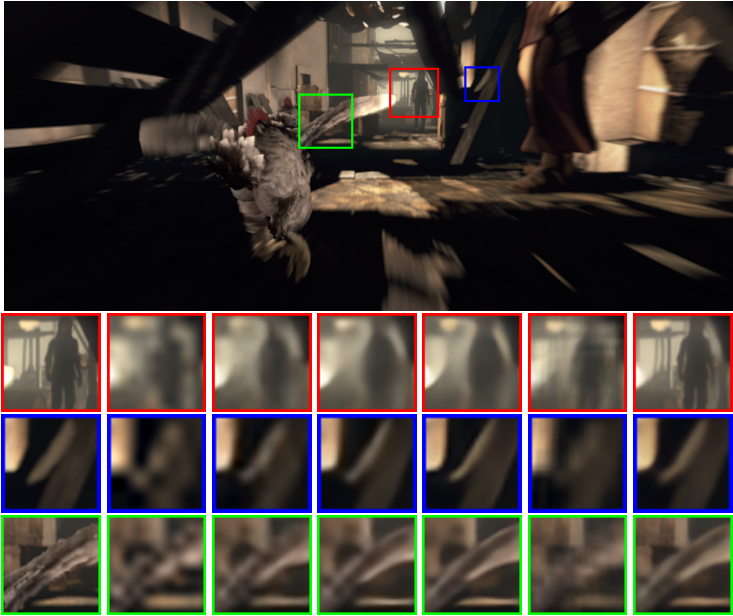


Figure 4: From left to right is ground-truth, bicubic interpolation, SRCNN [4], VDSR [18], our SRResNet + NN + sliding, PatchMatch [2], our CC-Net + SS-Net.

**Evaluation** We evaluate the results of RefSR-Net on the MPI Sintel testing set, and compare

with representative single image SR approaches such as bicubic interpolation, SRCNN [10], VDSR [18] and SRResNet based models, as well as exemplar-based SR approach PatchMatch [9] in Table 2. Figure 4 and 5 further depict the visual comparisons in testing clips among different methods.



Figure 5: From left to right is ground-truth, bicubic interpolation, SRCNN [10], VDSR [18], our SRResNet + NN + sliding, PatchMatch [9], our CC-Net + SS-Net.

## 6 Conclusion

In this paper, we preposed the learning-based RefSR scheme for cross-scale image super-resolution. Our method is divided into two components, i.e., a learning-based cross-scale correspondence and a learning-based patch synthesis. Experimental results on super-resolution of image using the temporal precedent frame as reference demonstrate the state-of-the-art performance of the proposed method.

## 7 Acknowledgement

This work is supported in part by Natural Science Foundation of China (NSFC) under contract No. 61331015, 6152211, 61571259 and 61531014, in part by the National key foundation for exploring scientific instrument No.2013YQ140517, in part by Shenzhen Fundamental Research fund (JCYJ20170307153051701).

## References

- [1] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi Morel. Low-complexity single-image super-resolution based on non-negative neighbor embedding.
- [2] Vivek Boominathan, Kaushik Mitra, and Ashok Veeraraghavan. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In *Computational Photography (ICCP), 2014 IEEE International Conference on*, pages 1–10. IEEE, 2014.
- [3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *CVPR*, volume 2, pages 60–65. IEEE, 2005.
- [4] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *CVPR*, volume 1, pages I–I. IEEE, 2004.
- [5] Donghyeon Cho, Minhaeng Lee, Sunyeong Kim, and Yu-Wing Tai. Modeling the calibration pipeline of the lytro camera for high quality light-field image reconstruction. In *ICCV*, pages 3280–3287, 2013.
- [6] Dengxin Dai, Radu Timofte, and Luc Van Gool. Jointly optimized regressors for image super-resolution. In *Computer Graphics Forum*, volume 34, pages 95–104. Wiley Online Library, 2015.
- [7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199. Springer, 2014.
- [8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 38(2):295–307, 2016.
- [9] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *ECCV*, pages 391–407. Springer, 2016.
- [10] Marius Drulea and Sergiu Nedevschi. Motion estimation using the correlation transform. *IEEE Transactions on Image Processing*, 22(8):3260–3270, 2013.
- [11] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. *arXiv preprint arXiv:1504.06852*, 2015.
- [12] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002.
- [13] He He and Wan-Chi Siu. Single image super-resolution using gaussian process regression. In *CVPR*, pages 449–456. IEEE, 2011.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

- [15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.
- [17] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, pages 1637–1645, 2016.
- [18] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016.
- [19] Kwang In Kim and Younghee Kwon. Single-image super-resolution using sparse regression and natural image prior. *TPAMI*, 32(6):1127–1133, 2010.
- [20] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [21] 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. *arXiv preprint arXiv:1609.04802*, 2016.
- [22] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William Freeman. Sift flow: Dense correspondence across different scenes. *ECCV*, pages 28–42, 2008.
- [23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.
- [24] Kaushik Mitra and Ashok Veeraraghavan. Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 22–28. IEEE, 2012.
- [25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.
- [26] Mehdi SM Sajjadi, Bernhard Schölkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. *arXiv preprint arXiv:1612.07919*, 2016.
- [27] Jordi Salvador and Eduardo Pérez-Pellitero. Naive bayes super-resolution forest. In *ICCV*, pages 325–333, 2015.
- [28] Samuel Schulter, Christian Leistner, and Horst Bischof. Fast and accurate image up-scaling with super-resolution forests. In *CVPR*, pages 3791–3799, 2015.
- [29] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, pages 1874–1883, 2016.

- [30] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *ICCV*, pages 1920–1927, 2013.
- [31] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *ACCV*, pages 111–126. Springer, 2014.
- [32] Shenlong Wang, Linjie Luo, Ning Zhang, and Jia Li. Autoscaler: Scale-attention networks for visual correspondence. *arXiv preprint arXiv:1611.05837*, 2016.
- [33] Ting-Chun Wang, Jun-Yan Zhu, Nima Khademi Kalantari, Alexei A Efros, and Ravi Ramamoorthi. Light field video capture using a learning-based hybrid imaging system. *arXiv preprint arXiv:1705.02997*, 2017.
- [34] 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*, 2016.
- [35] Judong Wu, Haoqian Wang, Xingzheng Wang, and Yongbing Zhang. A novel light field super-resolution framework based on hybrid imaging system. In *Visual Communications and Image Processing (VCIP)*, 2015, pages 1–4. IEEE, 2015.
- [36] Chih-Yuan Yang and Ming-Hsuan Yang. Fast direct super-resolution by simple functions. In *ICCV*, pages 561–568, 2013.
- [37] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *CVPR*, pages 1–8. IEEE, 2008.
- [38] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.
- [39] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang. Coupled dictionary training for image super-resolution. *IEEE Transactions on Image Processing*, 21(8):3467–3478, 2012.
- [40] Jianchao Yang, Zhe Lin, and Scott Cohen. Fast image super-resolution based on in-place example regression. In *CVPR*, pages 1059–1066, 2013.
- [41] Xiaoyun Yuan, Fang Lu, Qionghai Dai, David Brady, and Liu Yebin. Multiscale gigapixel video: A cross resolution image matching and warping approach. In *IEEE International Conference on Computational Photography*, 2017.
- [42] Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32): 2, 2016.
- [43] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on Curves and Surfaces*, pages 711–730. Springer, 2010.