CrossNet: An End-to-end Reference-based Super Resolution Network using Cross-scale Warping

Haitian Zheng[1] Mengqi Ji[1,2] Haoqian Wang[1] Yebin Liu[3] Lu Fang[1] *

[1] Tsinghua University, Tsinghua-Berkeley Shenzhen Institute
[2] Hong Kong University of Science and Technology
[3] Tsinghua University, Dept. of Automation

Abstract. The Reference-based Super-resolution (RefSR) super-resolves a low-resolution (LR) image given an external high-resolution (HR) reference image, where the reference image and LR image share similar viewpoint but with significant resolution gap (8×). Existing RefSR methods work in a cascaded way such as patch matching followed by synthesis pipeline with two independently defined objective functions, leading to the inter-patch misalignment, grid effect and inefficient optimization. To resolve these issues, we present CrossNet, an end-to-end and fully-convolutional deep neural network using cross-scale warping. Our network contains image encoders, cross-scale warping layers, and fusion decoder: the encoder serves to extract multi-scale features from both the LR and the reference images; the cross-scale warping layers spatially aligns the reference feature map with the LR feature map; the decoder finally aggregates feature maps from both domains to synthesize the HR output. Using cross-scale warping, our network is able to perform spatial alignment at pixel-level in an end-to-end fashion, which improves the existing schemes [1, 2] both in precision (around 2dB-4dB) and efficiency (more than 100 times faster).

Keywords: Reference-based Super resolution · Light field imaging · Image synthesis · Encoder-decoder · Optical flow

1 Introduction

Reference-based super-resolution (RefSR) methods [2] utilizes an extra high resolution (HR) reference image to help super-resolve the low resolution (LR) image that shares similar viewpoint. Benefit from the high resolution details in reference image, RefSR usually leads to competitive performance compared to single-image SR (SISR). While RefSR has been successfully applied in light-field reconstruction [1, 3, 2] and giga-pixel video synthesis [4], it remains a challenging

* The correspondence author is Lu FANG (fanglu@sz.tsinghua.edu.cn). This work is supported in part by Natural Science Foundation of China (NSFC) under contract No. 61722209, 61331015 and 61522111, in part by the National key foundation for exploring scientific instrument of China No.2013YQ140517.

and unsolved problem, due to the parallax and the huge resolution gap (8x) exist between HR reference image and LR image. Essentially, how to transfer the high-frequency details from the reference image to the LR image is the key to the success of RefSR. This leads to the two critical issues in RefSR, i.e., image correspondence between the two input images and high resolution synthesis of the LR image.

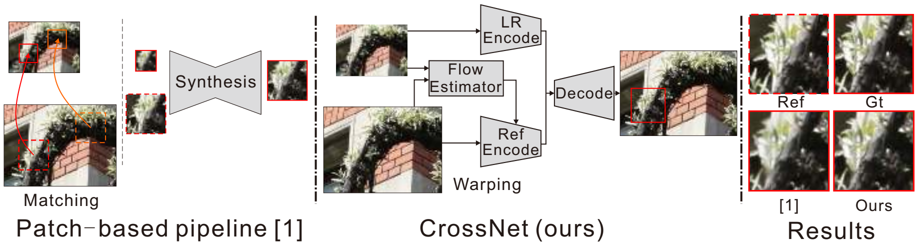


Fig. 1. Left: the 'patch maching + synthesis' pipeline of [2], middle: the proposed end-to-end CrossNet, right: results comparisons.

In the initial work of [1], to develop image correspondences between the two inputs, the gradient features on the down-sampled patches in the HR reference are used for patch-based matching, while patch averaging is designed for the image synthesis. However, the oversimplified and down-sampled correspondence estimation of [1] does not take advantage of the high frequency information for matching, while the synthesizing step does not utilize high resolution image prior for better fusion. To address the above two limitations, a recent work [2] replaces the gradient feature of [1] with the convolutional neural network (CNN) learned features to improve the matching accuracy, and then proposes an additional CNN which utilizes the state-of-the-art single image super-resolution (SISR) algorithm [5] for patch synthesis. However, the 'patch matching + patch synthesis' scheme of [1, 2] are fundamentally limited. Firstly, the adopted sliding averaging blurs the output image and causes grid artifacts. Moreover, patch-based synthesis is inherently incapable in handling the non-rigid image deformation caused by viewpoint changes. To impose the non-rigid deformation to patch-based algorithms, [3] enriches the reference images by iteratively applying non-uniform warping before the patch synthesis. However, directly warping between the low and high resolution images is inaccurate. In addition, such iterative combination of patch matching and warping introduces heavy computational burden, e.g. around 30min for synthesizing an image.

In this paper, we propose CrossNet, an end-to-end convolutional neural network based on the idea of 'warping + synthesis' for reference-based image super-resolution. We discard the idea of 'patch matching' and replace it with 'warping', which enables the design of 'Encoder-Warping-Decoder' structure, as shown in Fig. 1. Such structure contains two encoders to extract multi-scale features from LR and reference image respectively. We take advantage of the warping module originated from spatial transformer network (STN) [6], and integrate it to our

HR reference image encoder. Compared with the patch matching based methods, warping naturally supports non-rigid deformation to overcome the parallax challenge in RefSR. More over, we extract multi-scale features in the encoder, and then perform multi-scale spatial alignment using warping, as shown in Fig. 1. The introduced multi-scale features capture the complementary scale information from two images, which help to alleviate the huge resolution gap challenge in RefSR. Finally, the decoder aggregates features to synthesize the HR output. Overall, our model is fully end-to-end trainable and does not require pretraining the flow estimator.

Extensive experiments have shown the superior performance of CrossNet (around 2dB-4dB gain) compared to state-of-the-art SISR and RefSR methods, under different datasets with large/small viewpoint disparities and different scales. Our trained model that generalized to external dataset including Stanford light field maintains the ability to retain high frequency details. More importantly, CrossNet is efficient in terms that it generates a 320 × 512 image within one second, while [1], [2] and [3] take 86.3s, 105.0s and around 30 minutes to perform the same task, respectively.

2 Related work

2.1 Single-image Super-resolution

The single-image super-resolution (SISR) problem aims to super-resolve an LR image without additional references. Despite that, the SISR problems are closely related to the Reference-based Super-resolution (RefSR) problem. In the early days, approaches based on adaptive sampling [7, 8] has been applied to SISR. However, such approaches did not utilize the statistics of nature images. In contrast, model-based approaches try to design image prior which helps to super-resolve the image-specific patterns. Such works usually utilize edge prior [9], total variation model [10], hyper-Laplacian prior [11], sparsity priors [12–15], or exemplar patches [16, 17].

More recently, the SISR problem was casted into a supervised regression problem, which try to learn a mapping function from LR patches to HR patches. Those works relies on varieties of learning techniques including nearest-neighbor search [18, 19], decision tree [20], random forests [21, 22], simple function [23, 24], Gaussian process regression [25], and deep neural networks.

With the increasing model capacity of the deep neural networks, the SISR performance has been rapidly improved. Since the appearance of the first deep learning-based SR method [26], a large number of works have been proposed to further improve the SISR performance. For example, Dong et al. [27] and Shi et al. [28] accelerate the efficiency of SISR by computing features on low-resolution domains. Kim et al. [29] proposed a 20-layers deep network for predicting the bicubic upsampling residue. Ledig et al. [5] proposed a deep residue network with adversarial training for SISR. Lai et al. [30] reconstructed the sub-band residuals using a multi-stage Laplacian network. Lim et al. [31] improved [5] by introducing a multi-scale feature extraction residue block for better performance. Because of the impressive performance of the MDSR network from [31], we employ MDSR as a sub-module for LR images feature extraction and RefSR synthesis.

2.2 Reference-based Super-resolution

Recent works such as [2, 1, 3, 32–34] uses additional reference images from different viewpoints to help super-resolving the LR input, which forms a new kind of SR method called RefSR. Specifically, Boominathan et al. [1] used an DSLR captured high-definition image as reference, and applies a patch-based synthesizing algorithm using non-local mean [19] for super-resolving the low-definition light-field images. Wu et al. [34] improved such algorithm by employing patch registration before the nearest neighbor searching, then applies dictionary learning for reconstruction. Wang et al. [3] iterate the patch synthesizing step of [1] for enriching the exemplar database. Zheng et al. [35] decompose images into subbands by frequencies and apply patch matching for high-frequency subband reconstruction. Recently, Zheng et al. [2] proposed a deep learning-based approach for the cross-resolution patch matching and synthesizing, which significantly boosts the accuracy of RefSR. However, the patch-based synthesizing algorithms are inherently incapable in handling the non-rigid image deformation that is often caused by the irregular foreground shapes. Under such cases, patch-based synthesize causes issues such as blocky artifact and blurring effect. Despite that sliding windows [1, 2] or iterative refinement [3] mitigate such difficulties to some extends, these strategies usually introduce heavy computational cost. On the contrary, our fully convolutional network makes it possible to achieve more than 100 times speedup compared to existing RefSR approaches, allowing the model to be applicable for real-time applications.

2.3 Image/video Synthesis using Warping

Our task is also related to image/video synthesis tasks that use additional images from other viewpoints or frames. Such tasks include view synthesis [36, 37], video denoising [38], super-resolution [38–40], interpolation or extrapolation [41, 42]. To solve this type of problems, deep neural networks based of the design of "warping and synthesis" has been recently proposed. Specifically, the additional images are backward warped to the target image using the estimated flow. Afterward, the warped image is used for image/frame synthesis using an additional synthesis module. We follow such "warping and synthesis" pipeline. However, our approach is different from existing works in the following ways: 1) in stead of the common practice where warping was performed on image-domain at pixel-scale [41, 42, 36, 37], our approach performs multi-scale warping on feature domain, which accelerates the model convergence by allowing flow to be globally updated at higher scales. 2) after the warping operations, a novel fusion scheme is proposed for image synthesis. Our fusion scheme is different from the existing synthesizing practices that include image-domain early fusion (concatenation) [41, 37] and linearly combining images [42, 36].

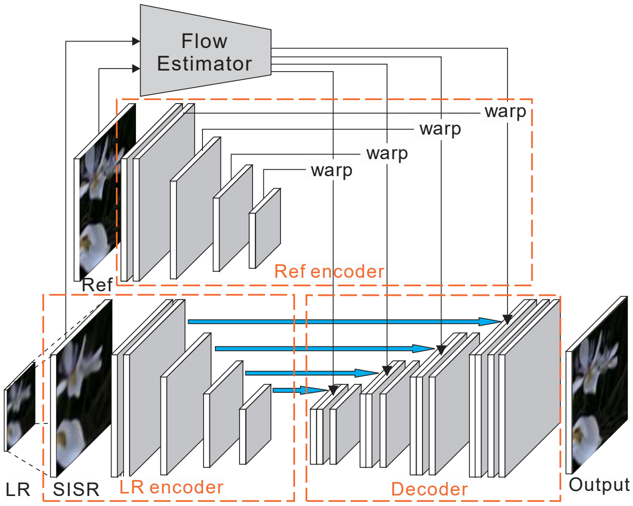


Fig. 2. Network structure of our proposed CrossNet.

3 Approach

Our reference-based super resolution scheme, named CrossNet, is based on a fully convolutional cross-scale alignment module that spatially aligns the reference image information to the LR image domain. Along with the cross-scale alignment module, an encoder-decoder structure is proposed to directly synthesize the RefSR output in an end-to-end, and fully convolutional fashion. The entire network is plotted in Fig. 2. In Section 3.1, we introduce the designs and properties of the fully convolutional cross scale alignment module. In Section 3.2, the end-to-end network structure is described, followed by the image synthesis loss function depicted in Section 3.3.

3.1 Fully Convolutional Cross-scale Alignment Module

Since the reference image is captured at different view points from LR image, it is necessary to perform spatial alignment. In [1–3], such correspondence is estimated by matching every LR patches with its surrounding reference patches. However, such sparsely-sampled and non-rigidly upsampled correspondence can easily fail around the region with varying depth or disparity.

Cross-scale warping. We propose cross-scale warping to perform non-rigid image transformation. Comparing to patch matching, we do not assume the depth plane to be locally constant. Our proposed cross-scale warping operation considers a pixel-wise shift vector V:

$$I_o = warp(y_{Ref}, V), \tag{1}$$

which assigns a specific shift vector for each pixel location, so that it avoids the blocky and blurry artifacts.

Cross-scale flow estimator. As shown on the top of Fig. 2, given an upsampled LR image and its corresponding reference image, we adopt the widely used FlowNetS [43] as our flow estimator to generate the cross-scale correspondence at multiple scale. To further improve the FlowNetS, we replace the final ×4 bilinear upsampling layer of FlownetS with two ×2 upsampling module, whereas each ×2 upsampling module contains a skip connection structure following a deconvolution layer. Such additional upsampling procedure allow the modified model to predict the flow-field with much finer definition. The modified flow estimator works to generate the multi-scale flow-fields as follows:

$$\{V^{(3)}, V^{(2)}, V^{(1)}, V^{(0)}\} = Flow(I_{LR\uparrow}, I_{REF}), \tag{2}$$

where the I_{REF} denotes the reference image, and $I_{LR\uparrow}$ denotes an representative Single-Image SR (SISR) approach [31] upsampled the LR image (I_{LR}):

$$I_{LR\uparrow} = SISR(I_{LR}). \tag{3}$$

More discussions on the choice of flow estimator are presented in discussion in Section 4.3.

3.2 End-to-end Network Structure

The patch matching calculates pixel-wise flow using a sliding window scheme. Such matching is computationally expensive, compared with the proposed fully convolutional network for cross-scale flow field estimation.

Resorting the cross-scale warping as a key component for spatial alignment, we propose an end-to-end network for RefSR synthesis. Our network, contains a **LR image encoder** which extracts multi-scale feature maps from the LR image I_L, a **reference image encoder** which extracts and aligns the reference image feature maps at multiple scales , and a **decoder** which perform multi-scale feature fusion and synthesis using the U-Net[44] structure. Fig. 2 summarizes the structure of our proposed CrossNet. The major modules, i.e., encoder, estimator and decoder, are elaborated as follows.

LR image encoder. Given the LR image I_L, we design a LR image encoder to extract reference feature maps at 4 scales. Specifically, we utilize SISR approach in Equation 3 to upsample the LR image. After that, we convolve the upsampled images with 64 filters (of size 5×5) with stride 1 to extract feature map at scale 0. We repeatedly convolve the feature map at the scale $i - 1$ (for $0 < i \leq 3$) with 64 filters (of size 5×5) with stride 2 to extract feature map at scale i. Such operations can be represented as

$$\begin{aligned} F_{LR}^{(0)} &= \sigma(\boldsymbol{W}_{LR}^{(0)} * I_{LR\uparrow} + \boldsymbol{b}_{LR}^{(0)}), \\ F_{LR}^{(i)} &= \sigma(\boldsymbol{W}_{LR}^{(i)} * F_{LR}^{(i-1)} + \boldsymbol{b}_{LR}^{(i)})\Downarrow_2,\ i = 1, 2, 3, \end{aligned} \tag{4}$$

where $F_{LR}^{(i)}$ is the LR feature map at scale i, σ stands for the activation function of rectified linear unit (ReLU) [45], $*$ denotes convolution, and $\Downarrow_2$ denotes 2D sampling with stride 2.

Note that resorting independent SISR approaches (such as [31]) to encode LR image owns two advantages. First, the SISR approaches that are validated on large-scale external datasets help the LR image encoder to generalize better on unseen scenes. Second, new state-of-the-art SISR approaches can be easily integrated into our system to improve the performance without changing our network structures.

Reference image encoder. Given the raw reference image I_R, a 4 scale feature extraction network with the exact structure from Equation 4 are used to sequentially extract reference image features $\{F_{REF}^{(0)}, F_{REF}^{(1)}, F_{REF}^{(2)}, F_{REF}^{(3)}\}$ from multiple scales. We allow the reference feature extractor and the LR feature extractor to learn different weights, which helps the two sets of features to complement each other.

After that, we perform backward warping operation on the reference image features $F_R^{(i)}$ using the cross-scale flow $V^{(i)}$ in equation 2, to generate the spatially aligned feature $\hat{F}_R^{(i)}$.

$$\hat{F}_{REF}^{(i)} = warp(F_{REF}^{(i)}, V^{(i)}),\ i = 0, 1, 2, 3. \tag{5}$$

More discussions on the multi-scale warping are presented in Section 4.3.

Decoder. After extracting the LR image feature and the warped reference image feature at different scales, a U-Net like decoder is proposed to perform fusion and SR synthesis. Specifically, the warped features and the LR image features at scale i (for $0 \leq i \leq 3$), as well as the decoder feature from scale $i - 1$ (if any) are concatenated following a deconvolution layer with 64 filters (of size 4×4) and stride 2 to generate decoder features at scale i,

$$\begin{aligned} F_D^{(3)} &= \sigma(\boldsymbol{W}_D^{(3)} \star (F_{LR}^{(3)}, \hat{F}_{REF}^{(3)}) + \boldsymbol{b}_D^{(3)}), \\ F_D^{(i)} &= \sigma(\boldsymbol{W}_D^{(i)} \star (F_{LR}^{(i+1)}, \hat{F}_{REF}^{(i+1)}, F_D^{(i+1)}) + \boldsymbol{b}_D^{(i)}),\ i = 2, 1, 0, \end{aligned} \tag{6}$$

where $\star$ denotes the deconvolution operation.

After generating the decoder feature at scale 0, three additional convolution layers with filter sizes 5×5 and filter number $\{64, 64, 3\}$ are added to perform post-fusion and to generate the SR output,

$$\begin{aligned} F_1 &= \sigma(\boldsymbol{W}_1 * F_D^{(0)} + \boldsymbol{b}_1), \\ F_2 &= \sigma(\boldsymbol{W}_2 * F_1 + \boldsymbol{b}_2), \\ I_p &= \sigma(\boldsymbol{W}_p * F_2 + \boldsymbol{b}_p). \end{aligned} \tag{7}$$

3.3 Loss Function

Our network can be directly trained to synthesize the SR output. Given the network prediction I_p, and the ground truth high-resolution image I_{HR}, the loss function can be written as

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \sum_{s} \rho(I_{HR}^{(i)}(s) - I_p^{(i)}(s)), \tag{8}$$

where $\rho(x) = \sqrt{x^2 + 0.001^2}$ is the Charbonnier penalty function [46], N is the number of samples, i and s iterate over training samples and spatial locations, respectively.

4 Experiment

4.1 Dataset

The representative Flower dataset [47] and Light Field Video (LFVideo) dataset [42] are used here. The Flower dataset [47] contains 3343 flowers and plants light-field images captured by Lytro ILLUM camera, whereas each light field image has 376×541 spatial samples, and 14×14 angular samples. Following [47], we extract the central 8×8 grid of angular sample to avoid invalid images, and randomly divide the dataset into 3243 images for training and 100 images for testing. The LFVideo dataset [42] contains real-scene light-field image captured by Lytro ILLUM camera. Similar to the Flower dataset, each light field image has 376×541 spatial samples and 8×8 angular samples. There are in total 1080 light-field samples for training and 270 light-field samples for testing.

For model training using these two dataset, the LR and reference images are randomly selected from the 8×8 angular grid. While for testing, the LR images at angular position $(i, i), 0 < i \leq 7$ and reference images at position $(0, 0)$ are selected for evaluating RefSR algorithms. As our model requires the input size being a factor of 32, the images from the two dataset are cropped to 320×512 for training and validation.

To validate the generalization ability of CrossNet, we also test it on the images from Stanford Light Field dataset [48] and Scene Light Field dataset [49], where we apply our trained model using sliding windows approach, with windows size being 512×512 and stride being 256 to output the SR result of the entire image. More details are presented in the generalization analysis in 4.2.

4.2 Evaluation

We train the CrossNet for 200K iterations on the Flower and LFVideo datasets for ×4 and ×8 SR respectively. The learning rates are initially set to 1e-4 and 7e-5 for the two dataset respectively, and decay to 1e-5 and 7e-6 after 150k iterations. As optimizer, the Adam [50] is used with $\beta_1 = 0.9$, and $\beta_1 = 0.999$. In comparison to CrossNet, we also test the latest RefSR algorithms SS-Net [2] and PatchMatch [1], and the representative SISR approaches including SRCNN [26], VDSR [29] and MDSR [31].

We evaluate the results using three image quality metrics: PSNR, SSIM [51], and IFC [52]. Table 1 shows quantitative comparisons for ×4 and ×8 RefSR under the two parallax settings, where the reference images are sampled at position $(0, 0)$ while LR images are sampled at position $(1, 1)$ and $(7, 7)$. Examining Table 1, the proposed CrossNet outperforms the previous approaches considerably under various settings including small/large parallax, different upsampling scales and different datasets, achieving 2dB-4dB gain in general.

Algorithm	Scale	Flower (1,1)	Flower (7,7)	LFVideo (1,1)	LFVideo (7,7)
		PSNR/SSIM/IFC	PSNR/SSIM/IFC	PSNR/SSIM/IFC	PSNR/SSIM/IFC
SRCNN [26]	×4	32.76 / 0.89 / 2.46	32.96 / 0.90 / 2.49	32.98 / 0.86 / 2.07	33.27 / 0.86 / 2.08
VDSR [29]	×4	33.34 / 0.90 / 2.73	33.58 / 0.91 / 2.76	33.58 / 0.87 / 2.29	33.87 / 0.88 / 2.30
MDSR [31]	×4	34.40 / 0.92 / 3.04	34.65 / 0.92 / 3.07	34.62 / 0.89 / 2.62	34.91 / 0.90 / 2.63
PatchMatch [1]	×4	38.03 / 0.97 / 5.11	35.23 / 0.94 / 3.85	38.22 / 0.95 / 4.60	37.08 / 0.94 / 3.99
CrossNet (ours)	×4	**42.09 / 0.98 / 6.70**	**38.49 / 0.97 / 5.02**	**42.21 / 0.98 / 5.96**	**39.03 / 0.96 / 4.61**
SRCNN [26]	×8	28.17 / 0.77 / 0.98	28.25 / 0.77 / 1.00	29.43 / 0.75 / 0.82	29.63 / 0.76 / 0.82
VDSR [29]	×8	28.58 / 0.78 / 1.04	28.68 / 0.78 / 1.06	29.83 / 0.77 / 0.89	30.04 / 0.77 / 0.89
MDSR [31]	×8	29.15 / 0.79 / 1.17	29.26 / 0.80 / 1.19	30.43 / 0.78 / 1.04	30.65 / 0.79 / 1.05
PatchMatch [1]	×8	35.26 / 0.95 / 4.00	30.41 / 0.85 / 2.07	36.72 / 0.94 / 3.81	34.48 / 0.91 / 2.84
SS-Net [2]	×8	37.46 / 0.97 / 4.72	32.42 / 0.91 / 2.95	37.93 / 0.95 / 4.06	35.81 / 0.93 / 3.30
CrossNet (ours)	×8	**40.31 / 0.98 / 5.74**	**34.37 / 0.93 / 3.45**	**41.26 / 0.97 / 5.22**	**36.48 / 0.93 / 3.43**

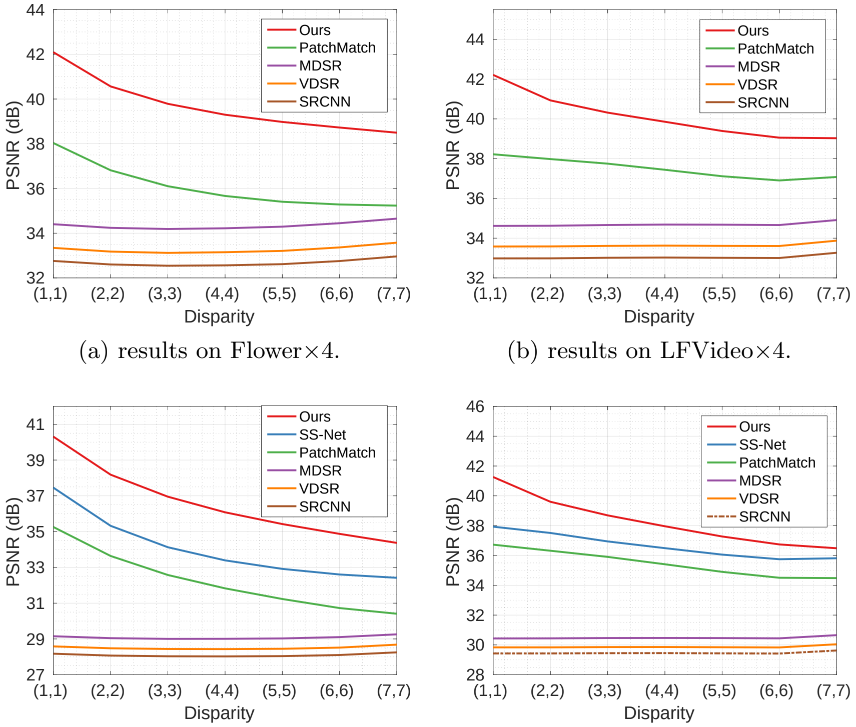
Table 1. Quantitative evaluation of the state-of-the-art SISR and RefSR algorithms, in terms of PSNR/SSIM/IFC for scale factors ×4 and ×8 respectively.

For better comparison, we also visualize the PSNR performance under different parallax setting in Fig. 3. As expected, the RefSR approaches such as CrossNet, PatchMatch, SS-Net outperform SISR approaches owe to the high-frequency details provided by reference images. However, RefSR results deteriorate as the parallax enlarges, due to the fact that the correspondence searching is more difficult for large parallax. In contrast, the performance of SISR approaches appears as 'U-shape' for different views, i.e., at the corners of LF image for disparity being (1, 1) and (7, 7), the SISR performs slightly better. This is probably due to the occurrence of easily super-resolved invalid region becomes larger at corners. Finally, it can be seen that the proposed CrossNet consistently outperforms the resting approaches under different disparities, datasets and scales.

Fig. 4 presents the visual comparisons of CrossNet with SISR approaches including SRCNN, VDSR, MDSR and RefSR approaches including PatchMatch and SS-Net under the challenging ×8 scale setting. Benefiting from the reference image, RefSR approaches show competitive results compared to the SISR methods, where the high frequency details are explicitly retained. Among them, the proposed CrossNet can further provide finer details, resembling the details in ground truth image. More visual comparison are shown in the supplementary material and supplementary video [4].

Generalization: to further estimate the cross-dataset generalization capacity of our model, we report the results on Stanford light field dataset (Lego Gantry) [48] and the Scene Light Field dataset [49], where the former one contains light field images captured by a Canon Digital Rebel XTi that set on a movable Mindstorms motors on the Lego gantry, and images from the latter one are also captured on a motorized stage with a standard DSLR camera. Under such equipment settings, the captured light-field images of the two datasets have much large parallax comparing to the ones captured by Lytro ILLUM cameras. The parallax discrepancy between datasets yields difficulty to our trained model, as our model is not particularly trained with large parallax.

[4] https://youtu.be/7htEaaNkxG8



(a) results on Flower×4.

(b) results on LFVideo×4.

(c) results on Flower×8.

(d) results on LFVideo×8.

Fig. 3. The PSNR measurement under different parallax settings: the reference images are select at $(0, 0)$ LF grid, while the LR image are selected at (i, i) LR grid $((i, i), 0 < i \leq 8)$.

To handle these two datasets, we employ a parallax augmentation procedure during training, which randomly offsets the reference input by $[-15, 15]$ pixels both horizontally and vertically. We take the pre-trained model parameters using LFVideo dataset (in Section 4.2) as the initialization, and re-train the CrossNet on the Flower dataset for 200K iterations in order to achieve better generalization. We use 7e-5 as the initial learning rate, and decay the learning rate using factors 0.5, 0.2, 0.1 at 50K, 100K, 150K iterations.

Table 2 and Table 3 compare in PSNR measurement our re-trained model with PatchMatch [1], SS-Net [2] for ×8 RefSR on the Stanford light-field dataset and the Scene Light Field dataset respectively. It can be seen that our approach outperforms the resting approaches with different parallax settings on the Stanford dataset. On average, our approach outperforms the competitive SS-Net by 1.79-2.50dB on the Stanford light-field dataset and 2.84dB on the Stanford light-field dataset.

Efficiency: It is worth mentioning that the proposed CrossNet generates an 320×512 image for $\times 8$ RefSR within 1 second, i.e., 0.75 second to perform

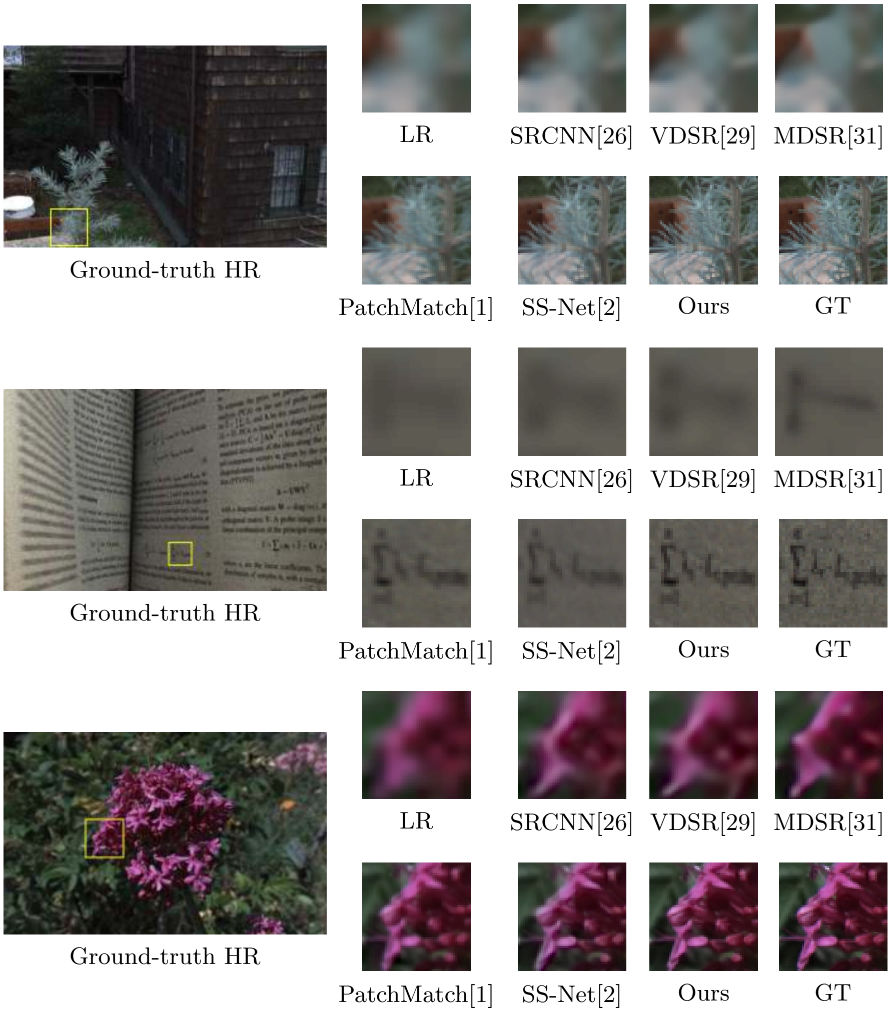


Fig. 4. Visual comparison for ×8 RefSR on LFVideo$(1, 1)$, LFVideo$(3, 3)$, Flower$(1, 1)$. In the experiment, our approach is compared against SRCNN[26], VDSR[29], MDSR[31], PatchMatch[1], and SS-Net[2].

SISR preprocessing using the MDSR [31] model, and 0.12 seconds to synthesize the final output. In contrast, the PatchMatch [1] takes 86.3 seconds to run on Matlab2016 using GPU parallelization while the SS-Net [2] takes on average 105.6 seconds running on GPU. The above running times are profiled using a machine with 8 Intel Xeon CPU (3.4 GHz) and a GeForce GTX 1080 GPU, while the model inferences of our CrossNet and SS-Net [2] are implemented on Python with Theano deep learning package [53].

Image, parallax=(1/3/5,0)	MDSR [31]	PatchMatch [1]	SS-Net [2]	Ours
Amethyst	29.64 / 29.66 / 29.74	36.44 / 34.71 / 33.20	36.91 / 34.97 / 33.35	**39.52** / **36.92** / **35.13**
Bracelet	24.66 / 24.68 / 24.64	35.66 / 33.71 / 25.47	36.33 / 34.19 / **32.53**	**38.19** / **34.27** / 27.20
Chess	30.39 / 30.42 / 30.39	38.68 / 36.68 / 34.99	39.85 / 38.64 / 37.12	**41.85** / **40.68** / **39.34**
Flowers	30.01 / 30.00 / 29.98	33.74 / 33.24 / 32.58	37.46 / 35.44 / 34.09	**39.50** / **36.53** / **34.56**
JellyBeans	41.09 / 41.00 / 41.15	39.48 / 38.68 / 37.19	37.98 / 36.60 / 35.14	**43.81** / **42.29** / **40.11**
LegoBulldozer	29.58 / 29.56 / 29.58	35.60 / 31.39 / 28.87	35.99 / 33.26 / 31.86	**38.79** / **35.00** / **32.61**
LegoGantry	26.58 / 26.52 / 26.58	31.73 / 29.86 / 27.15	32.68 / **30.97** / **30.06**	**33.42** / 30.83 / 29.96
LegoKnights	29.49 / 29.48 / 29.46	33.57 / 30.73 / 27.57	33.48 / 31.45 / 30.20	**37.60** / **34.40** / **32.11**
LegoTruck	30.82 / 30.80 / 30.67	34.96 / 34.22 / 33.30	37.80 / 36.44 / 34.87	**39.87** / **38.51** / **37.17**
TarotCardsLarge	22.91 / 22.89 / 22.86	27.98 / 20.90 / 20.40	29.69 / **26.71** / **24.63**	**31.27** / 23.69 / 22.16
TarotCardsSmall	23.98 / 23.98 / 23.97	30.08 / 29.30 / 27.60	32.92 / **31.44** / **30.70**	**35.48** / 31.42 / 28.22
StanfordBunny	36.82 / 36.90 / 36.96	37.39 / 37.15 / 36.75	40.36 / 39.77 / 39.09	**41.99** / **41.48** / **40.88**
Average	29.66 / 29.66 / 29.67	34.61 / 32.55 / 30.42	35.96 / 34.16 / 32.81	**38.44** / **35.50** / **33.29**

Table 2. ×8 super-resolution experiment on the Stanford light field dataset [48].

Image, parallax=(1,0)	MDSR [31]	PatchMatch [1]	SS-Net [2]	Ours
Bikes	28.38	36.70	36.36	**37.91**
Church	36.04	41.89	40.10	**43.68**
Couch	32.52	33.93	35.86	**39.83**
Mansion	28.03	32.83	33.39	**36.38**
Statue	29.72	35.96	35.21	**37.30**
Average	30.94	36.26	36.18	**39.02**

Table 3. ×8 super-resolution experiment on the Scene light field dataset [49].

4.3 Discussions

One may concern that our loss is designed for image synthesis, and does not explicitly define terms for flow estimation. However, since the correctly aligned features are extremely informative for decoder to reconstruct high-frequency details, our model actually learns to predict optical flow by aligning features maps in an unsupervised fashion. To validate the effectiveness of the learned flow by aligning feature, we visualize the intermediate flow field at all scales in Fig. 5(d), where flow predictions at scales $0, 1, 2, 3 (\times 1, \times 2, \times 4, \times 8)$ are reasonably coherent, yet noisy flow predictions are observed at scales $4, 5 (\times 16, \times 32)$, because the flow at scale $4, 5$ are not used for the feature-domain warping.

In addition to the multi-scale feature warping module proposed in this paper, we investigate a single-scale image warping counterpart which performs reference image warping **before** the following image encoder for feature extraction. This counterpart is inspired by the common practice in [42, 36] that performs image warping before synthesis. More concretely, our single-scale image warping counterpart performs image warping using the flow from scale 0: $\hat{I}_{REF} = warp(I_{REF}, V^{(0)})$. After that, reference image encoder with the same structure is used to extract features from the warped reference image. Without changing the structure of encoder and decoder, such image warping counterpart CrossNet-iw has the same model size as CrossNet.

We train both CrossNet-iw and CrossNet according to the same procedure in Section 4.2. We also adopt a pretraining strategy to train CrossNet-iw. We pretrain the flow estimator of WS-SRNet with image warping task for 100K iterations, and then apply the joint training for another 100K iterations, resulting

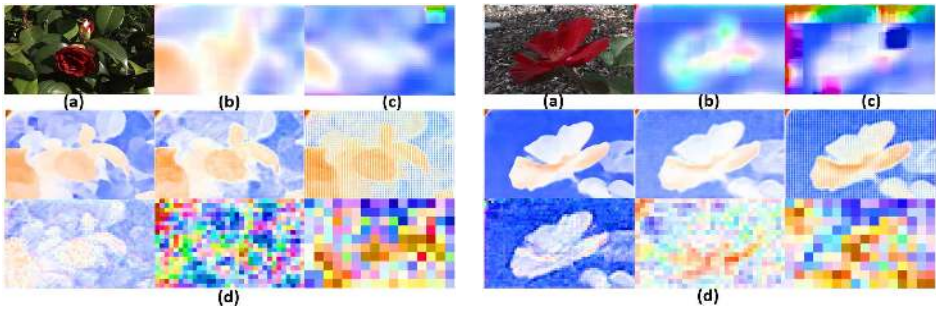
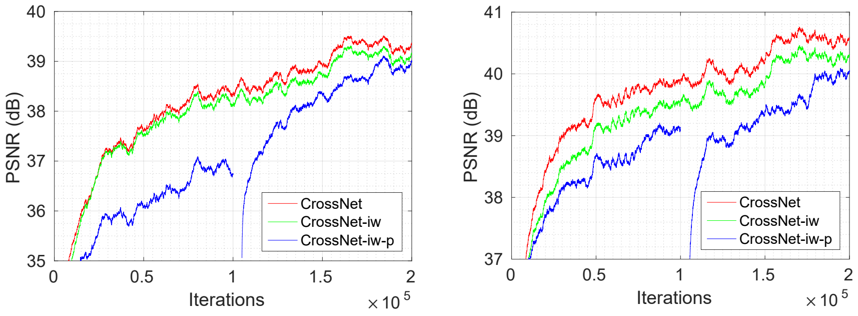


Fig. 5. Flow visualization and comparison for sample #1, #99 in the Flower ×8 testing set. (a) the HR image, (b)(c)(d) flow visualization of PatchMatch [1], SS-Net [2], and our approach respectively. In (d), the flow is visualized at scales ×1, ×2, ×4 (row 1), and ×8, ×32, ×64 (row 2).

the CrossNet-iw-p model. Fig. 6 shows the PSNR convergence curves on training set for ×8RefSR on the Flower and LFVideo dataset. It can be noticed that our CrossNet converges faster than the CrossNet-iw counterparts. At the end of the training, CrossNet outperforms CrossNet-iw 0.20dB and 0.27dB on training set. Table 4 shows the RefSR precision on the test sets with three representative point views. CrossNet outperforms CrossNet-iw, especially on small parallax setting. It is reasonable because the training uses random sampled pairs from the LF grid, which are mostly took up by small parallax training pairs.



(a) Training convergence curve on Flower×8.

(b) Training convergence curve on LFVideo×8.

Fig. 6. The convergence analysis on our feature domain warping scheme (CrossNet) versus image warping schemes. Our model (CrossNet, red) converges faster than the image-domain warping counterpart (CrossNet-iW) with or without pre-training.

As our method relies on the cross-scale flow estimators, it is also important to study the flow predicting capacities of different flow estimator. For such purpose, we train the FlowNetS and our modified model (FlowNetS+) on the Flower and the LFVideo dataset for warping the reference images to the ground truth images given the reference and LR image as input. As shown In Table 5,

Model	parameter size	Flower×8 (1,1)	Flower×8 (3,3)	Flower×8 (7,7)
		PSNR/SSIM/IFC	PSNR/SSIM/IFC	PSNR/SSIM/IFC
CrossNet	41M	**40.31 / 0.98 / 5.74**	**36.95 / 0.96 / 4.61**	34.37 / 0.93 / 3.45
CrossNet-iw	41M	40.14 / 0.98 / 5.75	36.94 / 0.96 / 4.59	**34.49 / 0.93 / 3.52**
CrossNet-iw-p	41M	40.01 / 0.98 / 5.75	36.85 / 0.96 / 4.58	34.35 / 0.93 / 3.49
		LFVideo×8 (1,1)	LFVideo×8 (3,3)	LFVideo×8 (7,7)
		PSNR/SSIM/IFC	PSNR/SSIM/IFC	PSNR/SSIM/IFC
CrossNet	41M	**41.26 / 0.97 / 5.22**	**38.69 / 0.96 / 4.32**	**36.48 / 0.93 / 3.43**
CrossNet-iw	41M	41.12 / 0.97 / 5.20	38.61 / 0.96 / 4.32	36.32 / 0.93 / 3.43
CrossNet-iw-p	41M	40.96 / 0.97 / 5.16	38.47 / 0.96 / 4.29	36.11 / 0.93 / 3.43

Table 4. Ablation study to evaluate the effectiveness of multi-scale feature warping.

while the FlowNetS+ contains 2% more parameters in comparison to FlowNetS, the additional upscaling layers of FlowNetS+ reasonably improves the warping precision in both the Flower dataset [47] and the LFVideo dataset [42], as they help to generate finer flow field. In addition, we also observe that by plain warping, the FlowNetS+ achieves notably better compatible performance compared to SS-Net [2], as depicted by the SS-Net (×8) row in Table 1.

Model	# of parameters	Flower×8(1,1)	Flower×8(7,7)	LFVideo×8(1,1)	LFVideo×8(7,7)
		PSNR/SSIM/IFC	PSNR/SSIM/IFC	PSNR/SSIM/IFC	PSNR/SSIM/IFC
FlownetS	31.9 million	37.78 / 0.97 / 5.41	31.23 / 0.90 / 3.02	**39.39 / 0.97 / 4.97**	34.94 / 0.92 / 3.30
FlownetS+	32.6 million	**38.04 / 0.97 / 5.46**	**31.66 / 0.90 / 3.11**	39.37 / 0.97 / 4.88	**35.85 / 0.93 / 3.54**

Table 5. Quantitative evaluation and the parameter sizes comparison, using different flow estimators to warp the reference image. The LR images are located at angular position $(3,3)$.

5 Conclusion

Aiming for the challenge large-scale (8×) super-resolution problem, we propose an end-to-end reference-based super resolution network named as CrossNet, where the input is a low-resolution (LR) image and a high-resolution (HR) reference image that shares similar view-point, the output is the super-resolved (4x or 8x) result of LR image. The pipeline of CrossNet is full-convolutional, containing encoder, cross-scale warping, and decoder respectively. Extensive experiment on several large-scale datasets demonstrate the superior performance of CrossNet (around 2dB-4dB) compared to previous methods. More importantly, CrossNet achieves a speedup of more than 100 times compared to existing RefSR approaches, allowing the model to be applicable for real-time applications.

References

1. Boominathan, V., Mitra, K., Veeraraghavan, A.: Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In: ICCP, IEEE (2014) 1–10
2. Zheng, H., Ji, M., Wang, H., Liu, Y., Fang, L.: Learning cross-scale correspondence and patch-based synthesis for reference-based super-resolution. In: BMVC. (2017)
3. Wang, Y., Liu, Y., Heidrich, W., Dai, Q.: 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)
4. Yuan, X., Lu, F., Dai, Q., Brady, D., Yebin, L.: Multiscale gigapixel video: A cross resolution image matching and warping approach. In: IEEE International Conference on Computational Photography. (2017)
5. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint arXiv:1609.04802 (2016)
6. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. (2015) 2017–2025
7. Li, X., Orchard, M.T.: New edge-directed interpolation. IEEE transactions on image processing **10**(10) (2001) 1521–1527
8. Zhang, L., Wu, X.: An edge-guided image interpolation algorithm via directional filtering and data fusion. IEEE transactions on Image Processing **15**(8) (2006) 2226–2238
9. Tai, Y.W., Liu, S., Brown, M.S., Lin, S.: Super resolution using edge prior and single image detail synthesis. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 2400–2407
10. Babacan, S.D., Molina, R., Katsaggelos, A.K.: Total variation super resolution using a variational approach. In: Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on, IEEE (2008) 641–644
11. Krishnan, D., Fergus, R.: Fast image deconvolution using hyper-laplacian priors. In: Advances in Neural Information Processing Systems. (2009) 1033–1041
12. Yang, J., Wright, J., Huang, T., Ma, Y.: Image super-resolution as sparse representation of raw image patches. In: CVPR, IEEE (2008) 1–8
13. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE transactions on image processing **19**(11) (2010) 2861–2873
14. Kim, K.I., Kwon, Y.: Single-image super-resolution using sparse regression and natural image prior. TPAMI **32**(6) (2010) 1127–1133
15. Yang, J., Wang, Z., Lin, Z., Cohen, S., Huang, T.: Coupled dictionary training for image super-resolution. IEEE Transactions on Image Processing **21**(8) (2012) 3467–3478
16. Glasner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: ICCV, IEEE (2009) 349–356
17. Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-based super-resolution. IEEE Computer graphics and Applications **22**(2) (2002) 56–65
18. Chang, H., Yeung, D.Y., Xiong, Y.: Super-resolution through neighbor embedding. In: CVPR. Volume 1., IEEE (2004) I–I
19. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 2., IEEE (2005) 60–65

20. Salvador, J., Pérez-Pellitero, E.: Naive bayes super-resolution forest. In: ICCV. (2015) 325–333
21. Schulter, S., Leistner, C., Bischof, H.: Fast and accurate image upscaling with super-resolution forests. In: CVPR. (2015) 3791–3799
22. Schulter, S., Leistner, C., Bischof, H.: Fast and accurate image upscaling with super-resolution forests. In: CVPR. (2015) 3791–3799
23. Yang, C.Y., Yang, M.H.: Fast direct super-resolution by simple functions. In: ICCV. (2013) 561–568
24. Yang, J., Lin, Z., Cohen, S.: Fast image super-resolution based on in-place example regression. In: CVPR. (2013) 1059–1066
25. He, H., Siu, W.C.: Single image super-resolution using gaussian process regression. In: CVPR, IEEE (2011) 449–456
26. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: ECCV, Springer (2014) 184–199
27. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: ECCV, Springer (2016) 391–407
28. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: CVPR. (2016) 1874–1883
29. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: CVPR. (2016) 1646–1654
30. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: CVPR. (2017) 624–632
31. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: CVPRW. Volume 1. (2017) 3
32. Wanner, S., Goldluecke, B.: Spatial and angular variational super-resolution of 4d light fields. In: European Conference on Computer Vision, Springer (2012) 608–621
33. Mitra, K., Veeraraghavan, A.: 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, IEEE (2012) 22–28
34. Wu, J., Wang, H., Wang, X., Zhang, Y.: A novel light field super-resolution framework based on hybrid imaging system. In: Visual Communications and Image Processing (VCIP), 2015, IEEE (2015) 1–4
35. Zheng, H., Guo, M., Wang, H., Liu, Y., Fang, L.: Combining exemplar-based approach and learning-based approach for light field super-resolution using a hybrid imaging system. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 2481–2486
36. Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. ACM Transactions on Graphics (TOG) **35**(6) (2016) 193
37. Ji, D., Kwon, J., McFarland, M., Savarese, S.: Deep view morphing. Technical report, Technical report (2017)
38. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. arXiv preprint arXiv:1711.09078 (2017)
39. Sajjadi, M.S., Vemulapalli, R., Brown, M.: Frame-recurrent video super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6626–6634
40. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution
41. Liu, Z., Yeh, R., Tang, X., Liu, Y., Agarwala, A.: Video frame synthesis using deep voxel flow. In: ICCV. Volume 2. (2017)

42. Wang, T.C., Zhu, J.Y., Kalantari, N.K., Efros, A.A., Ramamoorthi, R.: Light field video capture using a learning-based hybrid imaging system. ACM Transactions on Graphics (TOG) **36**(4) (2017) 133
43. Fischer, P., Dosovitskiy, A., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. arXiv preprint arXiv:1504.06852 (2015)
44. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241
45. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10). (2010) 807–814
46. Bruhn, A., Weickert, J., Schnörr, C.: Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. IJCV **61**(3) (2005) 211–231
47. Srinivasan, P.P., Wang, T., Sreelal, A., Ramamoorthi, R., Ng, R.: Learning to synthesize a 4d rgbd light field from a single image. In: ICCV. Volume 2. (2017) 6
48. : The (new) stanford light field archive. http://lightfield.stanford.edu/lfs.html
49. Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A., Gross, M.H.: Scene reconstruction from high spatio-angular resolution light fields. ACM TOG (2013)
50. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
51. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4) (2004) 600–612
52. Sheikh, H.R., Bovik, A.C., De Veciana, G.: An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Transactions on image processing **14**(12) (2005) 2117–2128
53. Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y.: Theano: A cpu and gpu math compiler in python