# Robust Reference-based Super-Resolution with Similarity-Aware Deformable Convolution

Gyumin Shim Jinsun Park In So Kweon
Robotics and Computer Vision Laboratory
Korea Advanced Institute of Science and Technology, Republic of Korea
{shimgyumin, zzangjinsun, iskweon77}@kaist.ac.kr

## Abstract

*In this paper, we propose a novel and efficient reference feature extraction module referred to as the Similarity Search and Extraction Network (SSEN) for reference-based super-resolution (RefSR) tasks. The proposed module extracts aligned relevant features from a reference image to increase the performance over single image super-resolution (SISR) methods. In contrast to conventional algorithms which utilize brute-force searches or optical flow estimations, the proposed algorithm is end-to-end trainable without any additional supervision or heavy computation, predicting the best match with a single network forward operation. Moreover, the proposed module is aware of not only the best matching position but also the relevancy of the best match. This makes our algorithm substantially robust when irrelevant reference images are given, overcoming the major cause of the performance degradation when using existing RefSR methods. Furthermore, our module can be utilized for self-similarity SR if no reference image is available. Experimental results demonstrate the superior performance of the proposed algorithm compared to previous works both quantitatively and qualitatively.*

## 1. Introduction

Single Image Super-Resolution (SISR) aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) image. Despite its notorious difficulty, SISR [34, 9] has received substantial attention due to its importance and practicality. As the Convolutional Neural Network (CNN) has demonstrated its capability in various research areas, including SISR, numerous deep learning-based SISR methods have been proposed [5, 14, 18] and have shown substantial performance improvements, especially with respect to reconstruction accuracy. To achieve a high peak signal-to-noise ratio (PSNR), the optimization process is typically defined as the minimization of the mean-squared-error (MSE) or the mean-absolute-error (MAE) between a ground truth image and a predicted high-resolution image. This type of algorithm has a critical limitation in that the generated solution is the mean or median of possible high-resolution images, with a lack of high-frequency details and a blurred visual quality level.

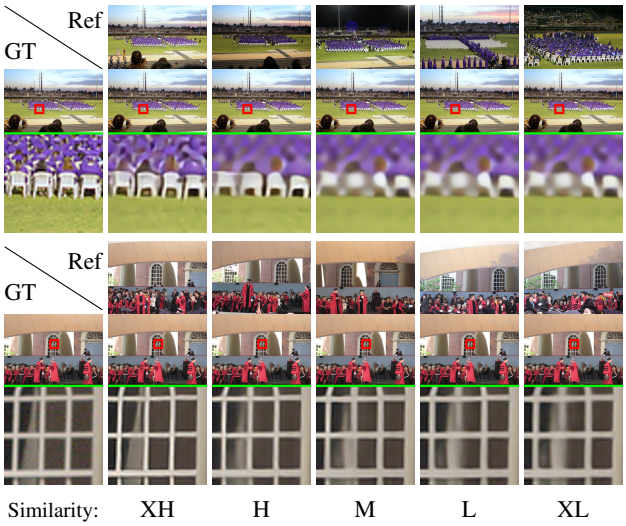


Figure 1: RefSR results with reference images with varying levels of similarity. XH, H, M, L, and XL denote very-high, high, middle, low, and very-low similarity levels, respectively.

In order to obtain high-resolution images with realistic textures, high-level feature similarity between the high-resolution and reconstructed images is enforced. Perceptual loss [12] or Generative Adversarial Network (GAN)-based algorithms [17, 24] are proposed for better output perceptual quality levels in SR. Specifically, adversarial learning helps a generator network to synthesize more realistic images while competing with a discriminator which attempts to differentiate super-resolved and original HR images. Although those algorithms provide visually pleasing outputs, they do not ensure an accurate reconstruction of the original high-resolution image, and this leads to PSNR degradation.

To mitigate this problem, some methods explicitly exploit additional information to make the SR outputs more like the ground truth and more visually pleasing [42, 40].

Because the original high-frequency information is lost due to the down-sampling process, it is highly challenging to reconstruct the precise high-frequency details of the ground truth. For such high-frequency details, providing similar content explicitly is a more reasonable approach compared to generating fake textures. Hence, the importance of reference-based SR (RefSR) is rapidly arising to overcome the limitations of SISR. RefSR aims to recover high-resolution images by utilizing an external reference (Ref) containing similar content to generate rich textures, changing the one-to-many to an one-to-one mapping problem (*i.e.*, mapping textures from the reference to the output). Many existing SR algorithms can be regarded as special cases of RefSR based on which reference image is paired with the input. For instance, reference images can be diversely acquired from video frames [19, 3], web image searches [35], or from different view points [42]. Conventional RefSR algorithms [2, 41, 42] are known to have a critical limitation in that the reference image should contain similar content to avoid any unexpected degradation in the performance. The most desired behavior of the RefSR algorithm is that it should be aware of the degree of similarity between low-resolution and reference images so as not to be affected by irrelevant reference images.

Inspired by recent works on video SR [26, 28] and RefSR methods [35, 42, 40], we propose a novel reference feature extraction module for the RefSR task. The module consists of stacked deformable convolution layers, and it can be inserted into any existing super-resolution network. The major benefit of our approach is that we aggressively search for similarity using a sophisticatedly designed offset estimator which learns the offsets of the deformable convolution. We adopt a non-local block [29] for our offset estimator, which performs pixel- or patch-wise similarity matching in a multi-scale manner. With the benchmark dataset used with RefSR, which has images paired with reference images with five different levels of similarity [40], we conduct experiments to demonstrate the superiority of the proposed algorithm. Experimental results show that our reconstruction results are more accurate and realistic with the help of the proposed module compared to the outcomes of previous algorithms.

Figure 1 shows the result of our method with reference images with different levels of similarity. Our method shows robustness to similarity variations. Even with a reference image with unrelated content or a much lower similarity level, our method still produces less noisy output, demonstrating the adaptiveness/robustness to various levels of content similarity in RefSR.

In summary, our contributions are as follows:

- We propose a novel end-to-end trainable reference feature extraction module termed the Similarity Search and Extraction Network, with similarity-aware deformable convolutions.
- The proposed method shows superior robustness/adaptiveness without any PSNR degradation given irrelevant references.
- The proposed method can be utilized not only for RefSR but also for exploiting self-similarity if no reference image is available.

## 2. Related Works

### 2.1. Single Image Super-Resolution

Conventional SISR algorithms aim to reconstruct HR images as accurately as possible by optimizing pixel-level reconstruction errors such as MSE and MAE. Dong *et al*. [5] propose a three-layer CNN-based SISR algorithm, referred to as SRCNN. Each layer of SRCNN is closely related to sparse representation, and it shows substantial performance improvements compared to those of conventional algorithms. Kim *et al*. [13, 14] propose a very deep CNN with input-output skip connections and a recursive architecture, offering stable and rapid convergence. Recently, the reconstruction accuracy was improved even further by adopting deeper networks with residual blocks and sub-pixel convolutions [18].

To overcome the major drawback of reconstruction-oriented SISR algorithms which produce blurred and non-realistic textures [17], perceptual loss [12] has been proposed to improve the perceptual quality of the generated images by minimizing feature-level differences extracted from a ImageNet [15] pre-trained network. Currently, GAN is known to be effective when used to generate realistic images [8], and numerous GAN-based SISR algorithms [17, 30] have been proposed. SRGAN [17] is the first GAN-based SISR algorithm which generates more realistic SR images compared to those of conventional algorithms. However, it was also found that degradation of the reconstruction accuracy is inevitable with GAN-based approaches, because generated realistic textures do not always correspond to ground truth textures.

### 2.2. Reference-based SR

Earlier works on RefSR derive from patch matching or patch synthesis schemes [2, 41]. Zheng *et al*. [41] propose a RefSR algorithm based on patch matching and synthesis with a deep network. Down-sampled patches are used for patch matching and for finding correspondences between input and reference images. However, those schemes have critical drawbacks in that they produce blur and grid artifacts and are unable to handle non-rigid image deformations

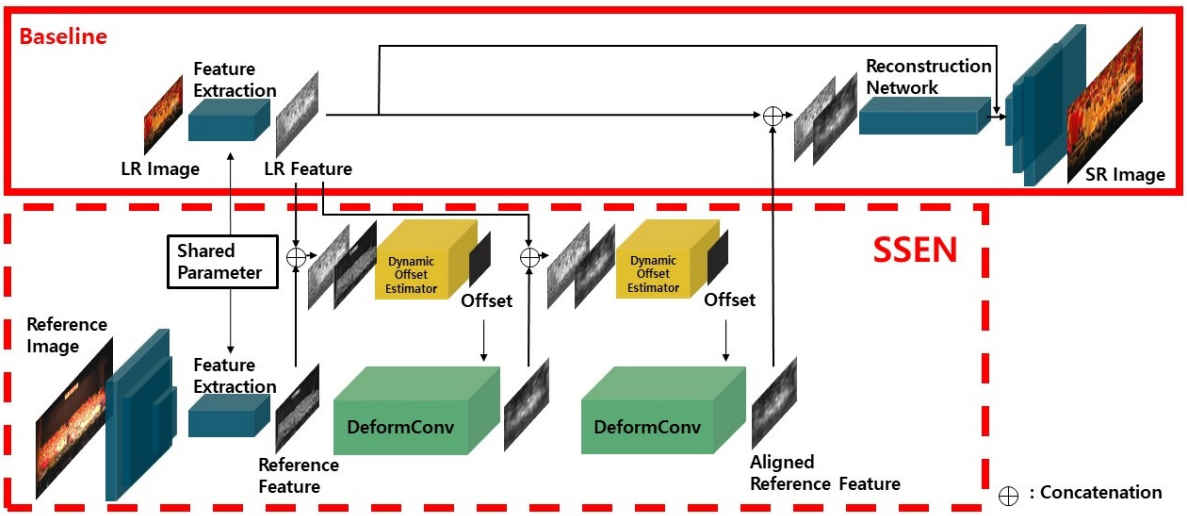


Figure 2: Illustration of our RefSR framework. A stack of two deformable convolution layers is depicted in the figure.

or inter-patch misalignments. Moreover, optimization including patch matching is inefficient due to its high computational cost. CrossNet [42] defines RefSR as a task where the reference image shares a similar viewpoint with a LR input image, and proposes an end-to-end neural network combining a warping process and image synthesis based on an optical flow [6, 10]. However, the ground truth for the optical flow is obtained at a high cost, and the flow estimation from other pre-trained networks is not accurate. In addition, although warping somewhat handles non-rigid deformation, it is highly vulnerable to large motions. SRNTT [40] points out the problem of robustness in CrossNet [42], arguing that severe performance degradation occurs when an unrelated reference image is paired with an input image. In SRNTT [40], a patch-wise matching scheme is adopted at the multi-scale feature level, which sacrifices computational efficiency for capturing long distance dependencies.

### 2.3. Self-Similarity and Non-local Block in SR

In a natural image, similar patterns tend to recur within the same image. Various methods have been studied regarding how to exploit self-similarity for image restoration [7, 33]. Those approaches attempt to utilize the internal information as a reference to reconstruct high-quality images. Huang *et al*. [9] propose a model allowing geometric transformation, which handles perspective distortions and affine transformations. However, the method of utilizing the intrinsic properties of images in deep learning-based methods remains ambiguous.

To deal with this problem, non-local block [29] based approaches [20, 38] have been proposed. The non-local operation computes pixel-wise correlations to capture long-range and global dependencies. The correlation is computed as a weighted sum of all positions in the input feature maps. This approach largely overcomes the locality of previous CNNs and is therefore suitable for various computer vision applications that require large receptive fields. The proposed method can be used to search not only for correspondences between input and reference image but also for self-similarity within a single image with the help of non-local blocks.

## 3. Similarity Search and Extraction Network

### 3.1. Network Architecture

The goal of reference-based super-resolution is to estimate a high-resolution image given a low-resolution input image and a high-resolution reference image. Inspired by the feature aligning capability of deformable convolution [26, 28], we formulate the RefSR problem as an integrative reconstruction process of matching similar contents between input and reference features and extracting the reference features in an aligned form. We propose an end-to-end unified framework that transfers HR details from reference images to restore high-frequency textures with the help of the proposed reference feature extraction module, specifically Similairity Search and Extraction Network (SSEN). The overall structure of SSEN is shown in Fig. 2. As input-reference pair of images are fed into the framework to reconstruct the high-resolution image, SSEN extracts features from the reference images in an aligned form, matching the contents in the pixel space without any flow supervision.

We design deformable convolution layers in a sequential approach, noting that the receptive field becomes larger as stacking continues in a sequential manner. Stacking multiple layers of deformable convolution, we discover that three layers of deformable convolution are the optimal structure for the best performance (*c.f*., Tab. 4). As RefSR expects to search for similar areas within the entire image, a large receptive field is the most critical issue during this task. For

this purpose, a multi-scale structure and non-local blocks are adopted to propagate offset information. Our module softly conducts pixel- or patch-level matching with an extremely large receptive field, estimating the offsets for deformable convolution kernels.

### 3.2. Stacked Deformable Convolution Layers

Deformable convolution [4] is proposed to improve the CNN's capability to model geometric transformations. It is trained with a learnable offset, which helps with the sampling of pixel points with a deformed sampling grid. Due to this characteristic, it is widely leveraged for feature alignments or implicit motion estimations without any optical-flow priors [26, 28]. In this work, we leverage the deformable convolution for the similarity search and extraction steps, adopting modulated deformable convolution [43] which additionally learns the dynamic weights of the sampling kernels with a modulation scalar.

In the modulated deformable convolution, modulation scalars are learned together with offsets to make the kernels more spatially-variant. Formally, the deformable convolution operation is defined as follows:

$$Y(p) = \sum_{k=1}^{K} w_k \cdot X(p + p_k + \triangle p_k) \cdot \triangle m_k, \qquad (1)$$

where $X$ is the input, $Y$ is the output, and $k$ and $K$ correspondingly denote the index and the number of kernel weights. $w_k$, $p$, $p_k$ and $\triangle p_k$ are the $k$-th kernel weight, indices of the center, the $k$-th fixed offset and the learnable offsets for the $k$-th location, respectively. $\triangle m_k$ is the modulation scalar, which enables relevancy-aware weight learning to robustly extract correspondences for cluttered or irrelevant input data in RefSR.

SSEN consists of several deformable convolution layers arranged in a sequential manner as shown in Fig. 2. The purpose of the stacking of deformable convolution layers sequentially is to sample more locations from reference images for aligning features with a larger receptive field. SSEN gradually aligns reference features to input features in each layer according to the offset provided from the dynamic offset estimator, which will be covered in detail in the following section.

### 3.3. Dynamic Offset Estimator

To capture the similarity located from near to far distances, the offset should be learned dynamically (*i.e.*, the offset should be able to cover a wide range of area, actively reaching various and distant positions). We design an offset estimator for learning the dynamic offsets, called *dynamic offset estimator*. Because the offset for deformable convolution should be learned based on the similarity between the reference image and the input low-resolution image, a reference feature and an input feature are concatenated as an

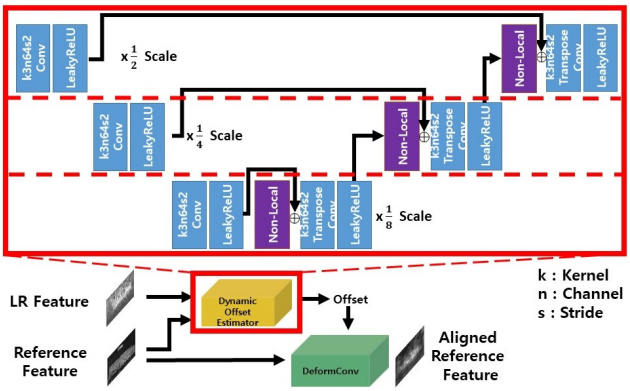


Figure 3: The proposed dynamic offset estimator. Features are down-sampled and fed into non-local blocks. The estimator is designed to learn residuals with skip connections.

input for the dynamic offset estimator, as shown in Fig. 3. We follow the multi-scale philosophy commonly adopted in optical flow estimations [6, 10]. The concatenated input is down-sampled three times such that multiple levels of scales can be considered when predicting offsets.

To localize relevant features which can be located at far distances effectively, we exploit non-local blocks in the dynamic offset estimator. The non-local operations capture the global correlation of intra- or inter-features, which helps with the prediction of dynamic offsets with an extremely large receptive field to handle both small and large displacements. We utilize three non-local blocks in the dynamic offset estimator so that the features are amplified with attention in each level of scale. Note that the processing of non-local operations with regard to down-sampled features can be considered as measuring the patch-wise similarity rather than the pixel-wise similarity.

Given an input $\mathbf{x}$ and an output $\mathbf{y}$, the non-local block operation is defined as follows:

$$\mathbf{y}_i = \mathbf{x}_i + W_{\mathbf{y}} \frac{1}{C(\mathbf{x})} \sum_j f(\mathbf{x}_i, \mathbf{x}_j) g(\mathbf{x}_j), \qquad (2)$$

where *i* is the index of the output position and *j* is the index of all possible positions. $W_{\mathbf{y}}$ denotes the weight matrix and $C(\mathbf{x})$ is the normalization factor. $f(\cdot)$ and $g(\cdot)$ represent the pair-wise operation and the linear embedding function, respectively.

Here, we can consider $\mathbf{y}$ as an attention guided feature, which highlights the global correlation between the input feature and the reference feature at the pixel- or patch-level. The function $g(\mathbf{x}_j)$ can be expressed as $\mathbf{W}_g \mathbf{x}_j$, which computes the linear embedding of the input signal $\mathbf{x}$ at position $j$. $f(\mathbf{x}_i, \mathbf{x}_j)$ calculates the pairwise similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. In this operation, we expect the similarity to be calculated in a similar manner to an inner product which is commonly used in patch matching. We adopt an embedded Gaussian function [29] for this pairwise operation defined

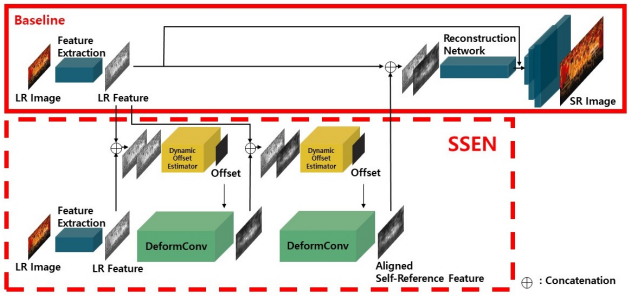


Figure 4: Illustration of the self-similarity SR framework

as follows:

$$f(\mathbf{x}_i, \mathbf{x}_j) = \exp(\theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j)), \quad (3)$$

where $\theta(\cdot)$ and $\phi(\cdot)$ are two linear embedding functions. For parameter-efficiency, we halve the dimension of the embeddings within the dynamic offset estimator.

### 3.4. RefSR and Self-Similarity SR Framework

**RefSR** The baseline is implemented with stacked residual blocks [18], which consist of residual blocks without batch normalization. Following previous work [18], we remove batch normalization from the network for better super-resolution performance. Note that SSEN can easily be attached to any existing super-resolution architecture.

The reference features extracted from the SSEN are fused with the input feature at the mid-level, after which the fused features are processed further at reconstruction network before they are up-sampled (*c.f.*, Fig. 2). In addition, a global skip connection between the input and output layers is adopted to ensure that our network focuses on residual feature learning. All feature manipulations are conducted at one-quarter of the size of the input spatial dimensions for efficient computation.

**Self-Similarity SR** If no appropriate reference image is available, the SSEN can be utilized in a self-reference manner. As shown in Fig. 4, the feature to be referenced is extracted from the same input image. In this situation, the proposed module is expected to exploit cues from the input image (*i.e.*, cues from itself), which can be helpful to minimize the reconstruction loss.

For the reconstruction network, we adopt RCAN [37] as our baseline network, which shows the best performance in SISR without a self-ensemble scheme. RCAN is designed with a residual-in-residual structure, which consists of several residual blocks with short skip connections. In addition, it adopts a channel attention mechanism to consider inter-dependencies among channels. The performance of SISR can be improved by seamlessly inserting the SSEN into a baseline network.

### 3.5. Training Objective

For the objective function, we adopt the Charbonnier penalty function [16] as the final training objective, which is known to help with the handling of outliers to improve the performance. The Charbonnier penalty is defined as follows:

$$\mathcal{L}_{rec} = \sqrt{\| I^{HR} - I^{SR} \|^2 + \varepsilon^2}, \quad (4)$$

where $I^{HR}$ and $I^{SR}$ denote the ground truth and the output from the proposed algorithm, respectively. $\varepsilon$ is set to $1e^{-6}$.

Moreover, perception-oriented objectives [12, 11] can be incorporated together to generate rich and realistic textures. For instance, PatchGAN judges realism at the scale of the image patches and shows performance superior to those of other GAN classifiers [11].

The objective of the PatchGAN is defined as follows:

$$\mathcal{L}_{adv} = \min_G \max_D \mathbb{E}_{I^{HR} \sim \mathbb{P}_r}[\log D(I^{HR})] + \mathbb{E}_{I^{SR} \sim \mathbb{P}_g}[\log(1 - D(I^{SR}))], \quad (5)$$

where $G$ denotes the generator which generates $I^{SR}$, and $D$ stands for the discriminator. $\mathbb{P}_r$ and $\mathbb{P}_g$ are the real data distribution and the model distribution, respectively.

## 4. Experimental Results

### 4.1. Dataset

We use the CUFED dataset [31] which contains 1883 albums that describe daily life events, to train the network for RefSR. For the RefSR task, there is the assumption that reference images contain contents similar to those of the input low-resolution images. To guarantee this assumption, Zhang *et al*. [40] reorganize the CUFED dataset into 13,761 image pairs scored based on the number of SIFT [22] correspondence matches. Furthermore, to evaluate the RefSR methods, the authors propose the CUFED5 dataset with 126 groups. Each high-resolution image is paired with five different reference images with five different levels of similarity. We adopt a random 90° rotation for augmentation during the training process.

For self-similarity SR, we utilize the DIV2K dataset [27], which has been widely used as a benchmark dataset in SR tasks. We augment the training data with random cropping with a patch size of $192 \times 192$ and a random 90° rotation. For the evaluation, the Urban100 [9], Set5 [1], Set14 [36], and B100 [23] datasets are utilized.

### 4.2. Training Details

All experiments are conducted with a scaling factor of $\times 4$ between the LR and HR images. The network is trained with an initial learning rate of $1e^{-4}$ using the ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use PyTorch to implement the model on an NVIDIA 1080Ti GPU. For the training of the network with $L_{adv}$, we initially pre-train the network only with the reconstruction loss, after which we fine-tune the network with the GAN loss attached to

| **SISR** | PSNR | SSIM |
|---|---|---|
| Bicubic | 24.18 | 0.684 |
| SRCNN [5] | 25.33 | 0.745 |
| LapSRN [16] | 24.92 | 0.730 |
| MDSR [18] | 25.93 | 0.777 |
| SRGAN [17] | 24.40 | 0.702 |
| ENet [24] | 24.24 | 0.695 |
| **Baseline** | **26.36** | **0.779** |

(a) **SISR**

| **RefSR** | PSNR | SSIM |
|---|---|---|
| Landmark [35] | 24.91 | 0.718 |
| CrossNet [42] | 25.48 | 0.764 |
| SRNTT [40] | 25.61 | 0.764 |
| SRNTT-$\ell_2$ [40] | 26.24 | 0.764 |
| **Ours** | **26.78** | **0.791** |

(b) **RefSR**

Table 1: PSNR / SSIM comparisons with other SR methods on the CUFED5 dataset. We group methods by (a) SISR and (b) RefSR. The best performances are written in bold.

the training objective. We train the network for 100K iterations with a batch size of 32 and adopt a cosine learning rate schedule scheme [21] with $\gamma = 0.9$.

### 4.3. Quantitative and Qualitative Evaluations

**RefSR** We compare the quantitative and qualitative results to those of other GAN-based SISR and RefSR methods to demonstrate the superior performance of the proposed method. Figure 5 shows comparisons of our method with other SISR and RefSR methods in terms of the visual quality. Overall, our method shows better visual quality with less noise, clear contents, and greater relevancy with the ground truth.

To evaluate our approach quantitatively, we measure the PSNR and SSIM (structural similarity) [32] as a means of distortion-oriented measurements. As shown in Tab. 1, our method is compared with various methods, including SISR and RefSR methods. The SISR methods are SRCNN [5], LapSRN [16], MDSR [18], SRGAN [17], and ENet [24], and the RefSR methods are Landmark [35], CrossNet [42], SRNTT [40], and SRNTT-$\ell_2$. The difference between SRNTT and SRNTT-$\ell_2$ is the presence of perceptual and adversarial losses during the training. SRNTT-$\ell_2$ is trained only with the reconstruction loss to acheive a higher PSNR. All methods are trained with the CUFED dataset and are tested on the CUFED5 dataset. Our method achieves the highest PSNR and outperforms all previous methods by a large margin (see Tab. 1). We emphasize that the effectiveness of the proposed algorithm is caused by less noisy outputs, the clearer boundaries, and the relevant textures softly transferred from the reference (see Fig. 5).

**Self-Similarity SR** For a fair comparison with the SISR methods, we utilize SSEN in a self-reference manner. As depicted in Fig. 4, the overall network is modified to enable the self-reference capability. In this variant, a LR input image is fed into the baseline and the SSEN simultaneously to extract self-reference features from the input image. Our method is mainly tested on the Urban100 [9] dataset, which contains structured scenes and rich textures to be utilized for self-similarity SR.

Figure 6 shows a qualitative comparison between models with and without the SSEN on the Urban100 dataset. We confirm that our method successfully recovers structured and recurring details by transferring finer textures from distant pixels. Relevant reference features extracted from distant pixels are well aligned by the SSEN, and this leads to more accurate and visually pleasing reconstruction results.

| Methods | Urban100 | Set5 | Set14 | B100 |
|---|---|---|---|---|
| Bicubic | 23.14/0.657 | 28.42/0.810 | 26.00/0.702 | 25.96/0.667 |
| SRCNN [5] | 24.52/0.722 | 30.48/0.862 | 27.50/0.751 | 26.90/0.710 |
| VDSR [14] | 25.18/0.754 | 31.35/0.883 | 28.02/0.768 | 27.29/0.072 |
| LapSRN [16] | 25.21/0.756 | 31.54/0.885 | 28.19/0.772 | 27.32/0.727 |
| MemNet [25] | 25.50/0.763 | 31.74/0.889 | 28.26/0.772 | 27.40/0.728 |
| EDSR [18] | 26.64/0.803 | 32.46/0.896 | 28.80/0.787 | 27.71/0.742 |
| RDN [39] | 26.61/0.802 | 32.47/0.899 | 28.81/0.787 | 27.72/0.741 |
| RCAN [37] | 26.82/0.808 | 32.63/0.900 | 28.87/0.788 | 27.77/0.743 |
| **Baseline** | 26.61/0.802 | 32.46/0.898 | 28.79/0.787 | 27.69/0.740 |
| **Ours** | 26.71/0.808 | 32.48/0.899 | 28.84/0.788 | 27.72/0.742 |

Table 2: PSNR / SSIM comparisons with other SISR methods on benchmark datasets.

| Methods | XH | H | M | L | XL |
|---|---|---|---|---|---|
| SRNTT | 25.17/0.734 | 25.13/0.729 | 25.06/0.728 | 25.07/0.720 | 25.14/0.729 |
| SRNTT-$\ell_2$ | 26.06/0.765 | 25.97/0.760 | 25.90/0.758 | 25.88/0.758 | 25.87/0.757 |
| Baseline | 26.36/0.779 | | | | |
| Ours | **26.78/0.791** | **26.52/0.783** | **26.48/0.782** | **26.42/0.781** | **26.41/0.780** |
| Ours (GAN+) | 25.35/0.742 | 25.05/0.732 | 24.99/0.730 | 24.95/0.729 | 24.98/0.730 |

Table 3: PSNR / SSIM comparisons with five different levels of similarity. The best numbers for each level of similarity are written in bold.

Table 2 shows quantitative comparisons with other SISR methods. SRCNN [5], VDSR [14], LapSRN [16], MemNet [25], EDSR [18], RDN [39], and RCAN [37] are compared with our self-similarity SR method. We mainly demonstrate the effectiveness of our module by verifying the performance improvement when the SSEN is attached during the training process. Note that while we reproduce our baseline with the RCAN [37] network unit, we achieve a slightly lower PSNR than the result reported in an earlier paper [37]. However, a PSNR improvement of 0.1dB is observed with our module on the Urban100 dataset. This confirms the effectiveness of the SSEN in self-similarity SR.

### 4.4. Ablation Studies

**Robustness to Irrelevancy** In this section, we compare the robustness to irrelevancy of the proposed method and SRNTT [40] at five different similarity levels. In Tab. 3, XH, H, M, L, and XL denote very-high, high, middle, low, and very-low similarities, respectively. In a perceptual quality-oriented training condition (*i.e.*, SRNTT and ours (GAN+)), our method outperforms at the very-high similarity level. Note that the performance of the proposed method with GAN for the other similarity levels can be improved by adjusting the weights for $L_{rec}$ and $L_{adv}$. In a reconstruction-oriented training condition (*i.e.*, SRNTT-$\ell_2$ and ours), our method shows superior robustness at every

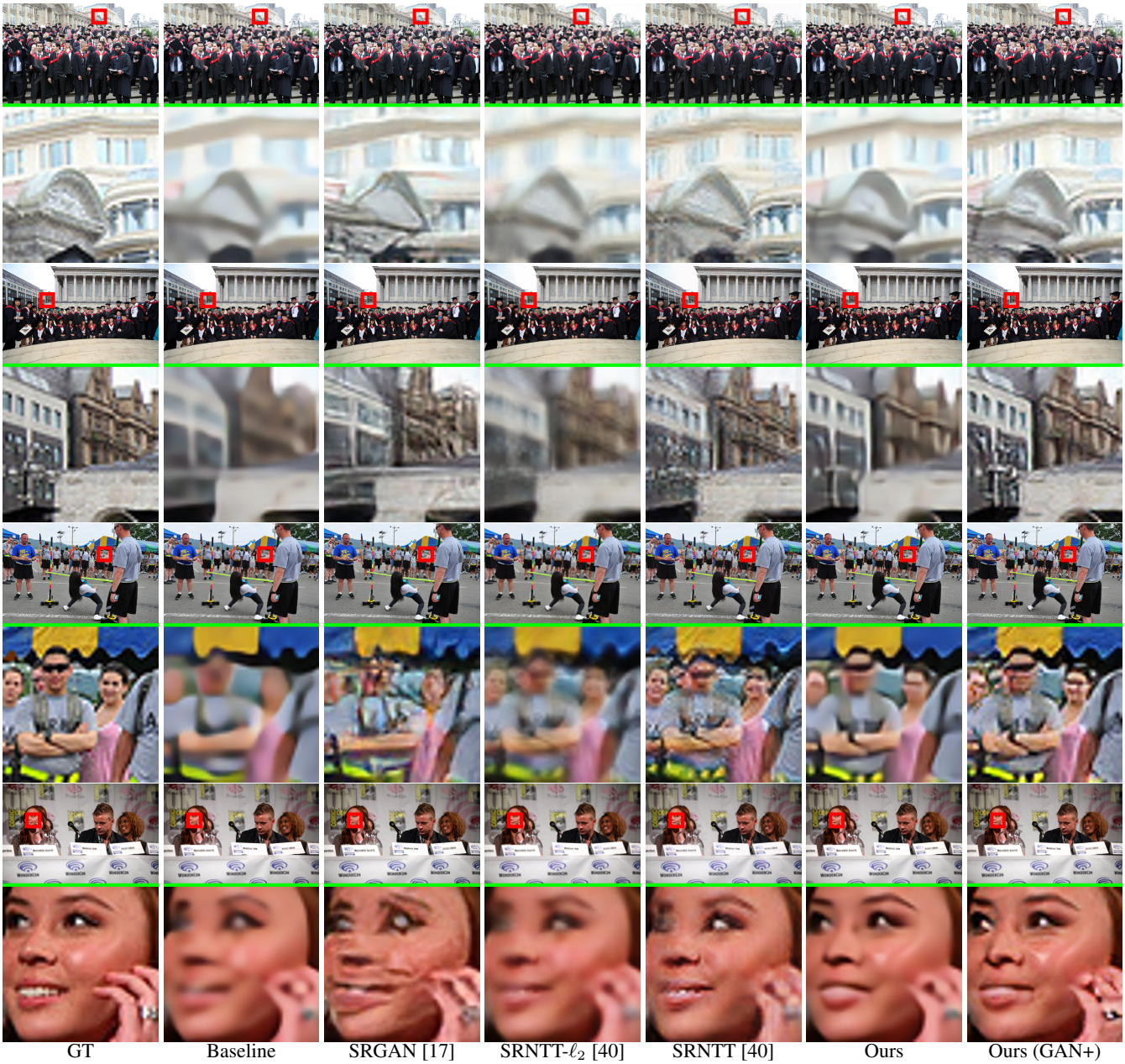


Figure 5: Qualitative comparisons with other SR methods on the CUFED5 dataset.

level of similarity compared to SRNTT-$\ell_2$. Robustness can be strongly proven given that our method consistently outperforms the baseline, which means that even a reference image with a low level of similarity is still quite effectively utilized by our method.

**Number of Deformable Convolution Layers** As shown in Tab. 4, we conduct an ablation study to investigate the optimal number of deformable convolution layers in Sec. 3.2. We confirm that three layers of deformable convolution are the optimal number for the best performance. Compared to the performance of the baseline, adding only one deformable convolution layer shows a great improvement, with a PSNR of 0.23dB at the XH level, whereas the performance rarely improves without any deformable convolution layers. Because the output features in each layer have different levels of alignment, a skip connection would be inappropriate if added to each deformable convolution layer. This implies that the deeper the network, the more difficult it is to train. Hence, stacking more layers (*e.g.*, more than four layers) does not guarantee better performance.

**Effect of Non-Local Block** To validate the importance of non-local blocks in our dynamic offset estimator in Sec. 3.3, we compare the performance of the proposed algorithm with and without non-local blocks, as shown in

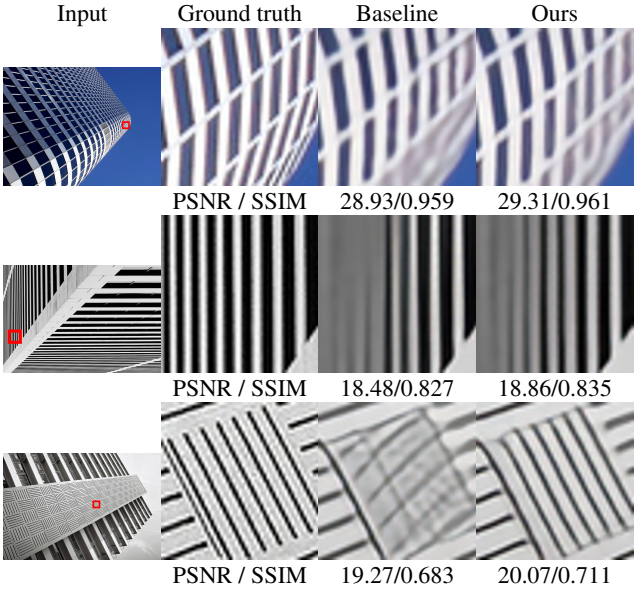


Figure 6: Qualitative results of self-similarity SR. PSNR and SSIM scores are shown together.

| # of layers | XH | H | M | L | XL |
|---|---|---|---|---|---|
| Baseline | 26.36/0.779 | | | | |
| No layers | 26.42/0.779 | 26.40/0.779 | 26.39/0.779 | 26.39/0.778 | 26.39/0.778 |
| 1 layer | 26.59/0.785 | 26.47/0.781 | 26.44/0.780 | **26.43**/0.780 | **26.41**/0.779 |
| 2 layers | 26.64/0.786 | 26.47/0.781 | 26.44/0.780 | 26.41/0.780 | 26.40/0.779 |
| 3 layers | **26.78/0.791** | **26.52/0.783** | **26.48/0.782** | 26.42/**0.781** | **26.41/0.780** |
| 4 layers | 26.70/0.789 | 26.45/0.781 | 26.43/0.781 | 26.40/0.780 | 26.38/0.779 |
| w/o *NB* | 26.68/0.788 | 26.47/0.782 | 26.44/0.781 | 26.41/0.780 | 26.39/**0.780** |
| w *NB* | **26.78/0.791** | **26.52/0.783** | **26.48/0.782** | **26.42/0.781** | **26.41/0.780** |

Table 4: PSNR / SSIM comparisons of (top) the number of layers and (bottom) the presence of non-local blocks (NB). The best numbers are written in bold.

Tab. 4. The network with non-local blocks consistently outperforms that without non-local blocks. This implies that the non-local blocks are helpful to capture the global context and the correlations of each feature, which are necessary to estimate the long-range dependencies. Without non-local blocks, we observe performance degradation of 0.1dB at the XH level, as shown in Tab. 4.

**Visualization of Offset** To validate the similarity- and relevancy-awareness of the proposed method, we visualize offsets which are the sampling locations of deformable convolution. We visualize all sampling locations of a pixel of the reference image. As the SSEN consists of three layers of deformable convolution with a $3 \times 3$ kernel, there are $9^3$ sampling points in total per pixel.

As shown in Fig. 7, the sampling points in the originally aligned regions tend to cluster near the output pixel location. On the other hand, in the misaligned area, the sampling locations tend to be spread apart by enlarging their receptive field. By enlarging its receptive field, the SSEN attempts


Figure 7: Offset visualization for (a) RefSR and (b) Self-Similarity SR. Reference points are shown on the left images and their corresponding sampling points are shown on the right images. Sampling points are drawn in colors identical to those of the reference points.

to find the best matching pairs of pixels between the input image and the reference image. In the self-similarity SR case, the input and reference images are previously aligned. Therefore, the sampling points are spread out along similar regions or structures.

**Computation Time** We validate the efficiency of our algorithm further by comparing its computation time with those of other algorithms on a $\times 4$ SR task. PatchMatch [2] takes 86.3s and SS-Net [41] takes 105.6s on average on a GPU. SRNTT [40] takes 9.053s for the patch matching process and 2.909s for reconstruction, *i.e.*, 11.962s in total. The patch matching process is the bottleneck in these methods. Our method generates a $322 \times 550$ image within 0.95s on average due to the effective feature alignment capability without any patch matching process. This is highly beneficial for many real-time applications. The inference time is measured on a machine equipped with an Intel Xeon CPU (2.10 GHz) and an NVIDIA GTX 1080 Ti GPU.

## 5. Conclusion

In this paper, we introduced a reference feature extraction module termed the Similarity Search and Extraction Network (SSEN), which extracts features from reference images in an aligned form relative to the low-resolution features. The proposed method is the first end-to-end trainable RefSR method that does not require heavy computation or explicit flow estimations. Our algorithm outperforms other RefSR methods with more robustness. Moreover, the proposed module can be utilized for self-similarity SR if no reference image is available. To deal with the long distance similarity issue, we adopt a multi-scale structure and non-local blocks for the dynamic offset estimator to predict a wide range of offsets. Experimental results demonstrate that our method achieves state-of-the-art performance quantitatively and qualitatively.

**Acknowledgements** This work was supported by the Institute for Information & Communications Technology Promotion (2017-0-01772) grant funded by the Korea government.

## References


[1] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proc. of British Machine Vision Conf. (BMVC)*, 2012.

[2] Vivek Boominathan, Kaushik Mitra, and Ashok Veeraraghavan. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In *IEEE Int'l Conf. on Computational Photography (ICCP)*, 2014.

[3] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2017.

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2014.

[6] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2015.

[7] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Trans. on Graph. (ToG)*, 2011.

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

[9] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2015.

[10] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2016.

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. of Advances in Neural Information Processing Systems*, 2012.

[16] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] 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 *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proc. of Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[19] Ce Liu and Deqing Sun. A bayesian approach to adaptive video super resolution. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2011.

[20] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *Proc. of Advances in Neural Information Processing Systems*, 2018.

[21] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *Proc. of International Conference on Learning Representations (ICLR)*, 2017.

[22] David G Lowe et al. Object recognition from local scale-invariant features. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 1999.

[23] David Martin, Charless Fowlkes, Doron Tal, Jitendra Malik, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2001.

[24] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2017.

[25] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2017.

[26] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally deformable alignment network for video super-resolution. *arXiv preprint arXiv:1812.02898*, 2018.

[27] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proc. of Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[28] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proc. of Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[29] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2018.

[30] 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 *Proc. of European Conf. on Computer Vision (ECCV)*, 2018.

[31] Yufei Wang, Zhe Lin, Xiaohui Shen, Radomir Mech, Gavin Miller, and Garrison W Cottrell. Event-specific image importance. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Processing (TIP)*, 2004.

[33] Chih-Yuan Yang, Jia-Bin Huang, and Ming-Hsuan Yang. Exploiting self-similarities for single frame super-resolution. In *Proc. of Asian Conf. on Computer Vision (ACCV)*, pages 497–510. Springer, 2010.

[34] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Trans. on Image Processing (TIP)*, 2010.

[35] Huanjing Yue, Xiaoyan Sun, Jingyu Yang, and Feng Wu. Landmark image super-resolution by retrieving web images. *IEEE Trans. on Image Processing (TIP)*, 2013.

[36] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Proc. of Int'l Conf. on Curves and Surfaces*, 2010.

[37] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2018.

[38] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019.

[39] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2018.

[40] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Reference-conditioned super-resolution by neural texture transfer. *arXiv preprint arXiv:1804.03360*, 2018.

[41] Haitian Zheng, Mengqi Ji, Lei Han, Ziwei Xu, Haoqian Wang, Yebin Liu, and Lu Fang. Learning cross-scale correspondence and patch-based synthesis for reference-based super-resolution. In *Proc. of British Machine Vision Conf. (BMVC)*, 2017.

[42] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2018.

[43] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2019.