Towards Content-Independent Multi-Reference Super-Resolution: Adaptive Pattern Matching and Feature Aggregation

Xu Yan[1,⋆], Weibing Zhao[1,⋆], Kun Yuan[1,2], Ruimao Zhang[3], Zhen Li[1,†], and Shuguang Cui[1]

[1] Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, [2]University of Ottawa, [3]SenseTime Research
{xuyan1@link., weibingzhao@link., lizhen@}cuhk.edu.cn

Abstract. Recovering realistic textures from a largely down-sampled low resolution (LR) image with complicated patterns is a challenging problem in image super-resolution. This work investigates a novel multi-reference based super-resolution problem by proposing a Content Independent Multi-Reference Super-Resolution (CIMR-SR) model, which is able to adaptively match the visual pattern between references and target image in the low resolution and enhance the feature representation of the target image in the higher resolution. CIMR-SR significantly improves the flexibility of the recently proposed reference-based super-resolution (RefSR), which needs to select the specific high-resolution reference (e.g., content similarity, camera view and relative scale) for each target image. In practice, a universal reference pool (RP) is built up for recovering all LR targets by searching the local matched patterns. By exploiting feature-based patch searching and attentive reference feature aggregation, the proposed CIMR-SR generates realistic images with much better perceptual quality and richer fine-details. Extensive experiments demonstrate the proposed CIMR-SR outperforms state-of-the-art methods in both qualitative and quantitative reconstructions.

Keywords: Super-Resolution, Content-Independent Multi-Reference, Universal Reference Pool, Local Feature Enhancement

1 Introduction

As one of the fundamental low-level vision problems, image super-resolution [4], which aims to reconstruct the high-resolution (HR) image from its low-resolution (LR) observation, has attracted increasing attention in both academic and industry. As shown in Fig. 1, the previous methods can be roughly divided into two categories, termed single image super-resolution (SISR) [3,31,34,27], and reference-based super-resolution (RefSR) [35,21,20,36]. For SISR, since the fine

⋆ Equal first authorship. † Corresponding author.

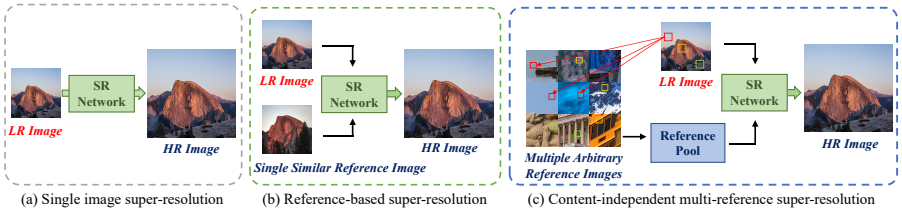

Fig. 1. Comparison between single image super-resolution (SISR), reference-based super-resolution (RefSR) and our proposed content-independent multi-reference super-resolution (CIMR-SR). Within (c), the patches marked in the same color stand for the similar patterns in feature space.

texture presented in original HR is usually lost in the LR, it becomes extremely difficult to recover fine textures when handling large up-scaling factors tasks. Therefore, RefSR [36,33] is meticulously proposed to address such issues by feature transferring between the HR reference images and the LR image. Although state-of-the-art RefSR significantly improves the quality of the reconstructed HR image, it is still suffering from the following problems to further improve the perceptual quality and generate richer fine-details.

- To get similar content or homogeneous patterns for LR images, most of the previous RefSR methods impose strict restrictions on the correlation between HR references and LR images, e.g., content similarity, camera view, and relative scale. However, such constraints are usually impractical in a lot of real applications.
- Different local regions within the LR image usually exhibit different reconstruction proprieties. The existing deep architectures lack the ability to adaptively enhance the feature representation from various patterns.

To address the above issues, in this work, we investigate a novel reference-based SR problem, that is, a universal multi-reference oriented image super-resolution, by proposing a Content-Independent Multi-Reference Super Resolution (CIMR-SR) model. As shown in Fig. 1, CIMR-SR breaks the restrictions of the previous RefSR methods on HR reference images by applying multiple arbitrary reference images. It could adaptively match local patterns from these content-independent reference images and aggregate them in the feature space to remarkably promote an ultimate representation ability of LR image.

Specifically, CIMR-SR consists of two major components, i.e., a universal reference pool (RP) and a local feature enhancement (LFE) module. The former is used to store various local patches (i.e., represented by the high-level feature representations) from the reference image to adaptively compensate for the information loss of the LR. In practice, the proposed CIMR-SR adaptively matches the visual patterns between the local patches in RP and target LR image. Then, it returns several groups of reference feature maps for further enhancement. The later is designed to aggregate the above-assembled feature maps and original fea-

ture representation of LR images. Unlike directly swapping the closest feature point on one single image [35], the LFE module achieves to aggregate similar feature points from multiple reference patches, making the network suitable for generating complex local details.

The **contributions** of this work are three folds. (1) To our best knowledge, this is the first work to address the multi-reference based image super-resolution by using deep learning with an end-to-end training mechanism. A novel CIMR-SR scheme is proposed to adaptively search the similar local visual patterns to enhance the feature representation of the target LR image. (2) A universal Reference Pool (RP) is constructed as a container for general local feature patterns and its memory burden is alleviated in conjunction with the diversity-insurance sampling strategy. In addition, an effective LFE module is firstly proposed to deal with feature aggregation for various patterns. (3) Extensive quantitative and qualitative experiments have demonstrated that the proposed CIMR-SR can generate realistic images with much better perceptual quality and richer fine-details, outperforming state-of-the-art methods.

2 Related Work

Deep Learning in SISR. Dong et al. [5] firstly attempted to learn an end-to-end mapping from HR-LR pairs by convolutional networks, various subsequent works tend to design the network deeper or wider to enhance the model representation capacity and computational efficiency. For example, a VGG style network VDSR [11] is proposed with a multi-scale training strategy to meet tasks at different scales. Furthermore, EDSR [12] is proposed by stacking modified residual blocks [8] to significantly improve the performance. As the depth of network plays a vital role in SISR, RCAN [33] improves the inferior performance with a very deep model by adopting attention mechanism and shared-source skip connections to bypass redundant information. SAN [3] further use the second-order attention mechanism, and implement a non-locally enhanced residual group for long-distance dependency learning.

In general, the approaches mentioned above aim at minimizing reconstructed loss with no prior included. In addition, some other works incorporate the prior in perceptual-related constraints to recover more visually plausible SR images. For instance, SRGAN [14] adds perceptual loss and generative adversarial strategy to address the issue of over-smoothing in SISR. SFTGAN [23] incorporates segmentation maps to induce categorical priors to generate offline transformation parameters for spatial-wise feature modulation. In this way, perceptual-related priors were implicitly incorporated to achieve better visual quality. However, although using adversarial strategies increases plausible visual quality, it results in the quantitative criteria (i.e., PSNR) reduction and fake textures generation.

Besides, there are some previous works focus on the LR images with a more realistic degradation (called blind SR). Specifically, blind SR introduces the complex blur kernels (e.g., motion blur in [27]) or DSLR camera's degradation process (e.g., zooming in [32], ISP pipeline in [28]) to produce the LR-HR paired

training set instead of the fixed bicubic kernel. Thus, the assumption about the degradation process limits it into specific scenarios (e.g., specific camera). On the contrary, this paper focuses on how to transfer features between LR and multiple external references with normal SISR training strategies.

Reference-based Super-Resolution. In contrast to SISR taking only a single LR image as the input, recent works use additional images from different views or scenes to assist the SR process, called RefSR. Traditional methods use sparse signal representation and linear combination [29] or example-based method [7] for reference utilization without data-driven training process. According to where the additional images come from, RefSR could be divided into two categories: internal and external RefSR. Compared with internal RefSR [21,20,6,9] utilizes the self-similarity of images to refer patches from itself, references in external RefSR could be acquired in multiple ways, such as from adjacent frames in a video [16], from web retrieval and external database [30]. As for external RefSR methods, a typical work is CrossNet [36], which learns the alignment parameters from optical flow to warp the feature between input LR and the reference image. However, this model highly depends on the assumption that the references need to be well aligned with the LR images. To address the alignment issue, SRNTT [35] refers to the work in style transfer and swap the most similar feature in the neural space during the SR process, which enables the learning of long-distance dependency and complicated feature transferring process. However, patch swap strategy limits itself to the closest feature with a narrow field of view in feature space, thus hampering the feature transferring process.

3 Methods

Given the specified down-scale factor s, the proposed CIMR-SR aims to estimate the SR image I^{SR} with the size $H \times W \times 3$ from its LR counterpart I^{LR} with the size $H/s \times W/s \times 3$ and the given M arbitrary content-independent reference images $\boldsymbol{I}^{Ref} = \{I_i^{Ref}\}_{i=1}^{M}$. To achieve the above goal, we firstly construct a reference pool (RP) in Sec. 3.1, which contains feature points converted from the features of reference patches. It provides additional information for detailed texture generation. And then we propose the local feature enhancement (LFE) module in Sec. 3.2, which selects reference features and aggregates group features by the effective feature searching and the feature aggregation mechanisms.

3.1 Reference Pool

Reference Pool (RP) is constructed for storing all reference information from multiple HR-Refs. Therefore, it needs to satisfy the following criteria:

- C1: Universal content-independent references should be collected and there is no restriction on pixel alignment with LR image;
- C2: Efficient and accurate feature searching and aggregation should be supported while exploiting such RP as external information.

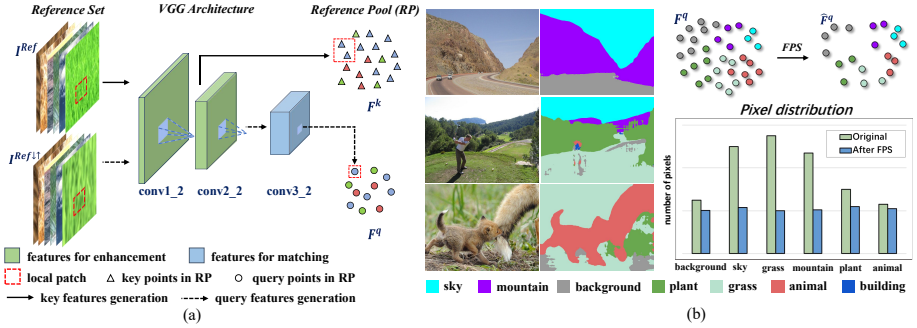

Fig. 2. (a) **RP generation.** Multiple HR-Refs (I^{Ref}) and LR-Refs ($I^{Ref\downarrow\uparrow}$) images are transferred from the RGB space to the feature space through a pre-trained CNN. Within the RP, each 3×3 local patch on feature map `conv3_2` is reshaped to a feature point and acts as a query feature, which can be mapped to larger patches on `conv2_2` or `conv1_2` stored as key features. (b) **Farthest Point Sampling (FPS).** The left feature point set indicates the original feature distribution, while the right is uniformly distributed over categories after FPS. Here different colors indicate different semantics.

C1. To satisfy condition C1, we map multiple reference images to a high-dimensional feature space offline without any pixel alignment restriction. As shown in Fig. 2 (a), for each reference image HR-Ref I_i^{Ref}, the j-th patch $I_{i,j}^{Ref}$ is transferred to a feature point through a pre-trained CNN, where i and j denote the index of reference image and patches. Concretely, there are two different kinds of features: which termed to the feature points generated from patches of HR-Refs and blurred LR-Refs as *key features* and *query features*, respectively. The *key features* are beneficial to restore high-resolution information and will be saved for feature aggregation, while *query features* will be used in feature matching with the LR image. Note that each LR-Ref is obtained by down-sampling and up-sampling corresponding HR-Ref to match the frequency band of the I^{LR} through bicubic interpolation. Therefore, the key features F_i^k and query features F_i^q for the i-th HR-Ref are obtained by,

$$F_i^k = \mathcal{F}(I_i^{Ref})_{l_k},\ F_i^q = \mathcal{F}(I_i^{Ref\downarrow\uparrow})_{l_q}, F_i^k \in \mathbb{R}^{N^k \times D^k},\ F_i^q \in \mathbb{R}^{N^q \times D^q}, \quad (1)$$

where l_k and l_q are specified layers from a feature extractor $\mathcal{F}$ (e.g., `conv3_2`), sign $\downarrow$ and $\uparrow$ denotes down-sample and up-sample operations. Besides, output F_i^k and F_i^q can also be seen as feature space point sets, which have point numbers N^k and N^q with dimensions D^k and D^q. Here D^k and D^q are the product of feature dimension of specified layers and the size of local patches on feature maps. Therefore, we break the restriction of pixel alignment and transfer image information into feature space. Furthermore, the constrains about the reference images in previous methods can be solved by building a universal reference pool (RP). Once arbitrary HR-Refs are offered, the RP would provide diverse patterns with various semantics and textures for LR reconstruction.

C2. It is non-trivial to build a universal RP not only considering the memory

limitation and time efficiency but also satisfying the constrain C2. Concretely, the distribution in RP tends to be locally dense and uneven, easily dominated by monotonous textures, which increases the computational cost and difficulty for conventional feature searching and aggregation. To solve the issues, we exploit Farthest Point Sampling (FPS) algorithm [18] on query features for RP construction.

Firstly, we collect multiple HR-Refs as $\boldsymbol{I}^{Ref} = \{I_i^{Ref}\}_{i=1}^{M}$ and transfer them into feature space point set $\boldsymbol{F}^q \in \mathbb{R}^{N \times D^q}$ by Eq. (1). Here N is the total point number of all query features, which will increase as the number of HR-Ref images increases. After that, we sample a subset of N' points ($N' < N$) as $\hat{\boldsymbol{F}}^q$ through FPS sampling. In each iteration FPS selects a new feature point from original features, it will choose the feature that has the farthest distance between all existing sampled features. It can achieve an approximate uniform sampling in feature space and has better coverage of the entire point set. Therefore, it is a good manner to reduce redundancy in a certain space. As illustrated in Fig. 2 (b), before the FPS, most of the pixels in the image belong to the sky and grass, which share similar features without complex texture. It leads the SR network dominated by monotonous textures and prone to a sub-optimal reconstruction. With the assistance of FPS, we can keep the diversity of features while reducing local redundancy, i.e., the distribution of feature points for each semantic and texture remains appropriately uniform. This resolves the efficiency and accuracy problems of RP, which benefits the subsequent LFE module.

3.2 Local Feature Enhancement Module

Be equipped with various features provided in RP, it is non-trivial to aggregate the high-dimensional features for each local region of LR effectively and efficiently. Especially, non-parametric operation, i.e., patch swapping [35], only considering the most similar features, which undoubtedly damages the diversity of features enhancement. To address the problems mentioned above, the Local Feature Enhancement (LFE) module is introduced to effectively search usable features and aggregate them for local regions of the LR image, thus producing enhanced features for reconstruction. Specifically, LFE contains two parts:

- *Feature Searching* retrieves K most similar feature points from query features $\hat{\boldsymbol{F}}^q$ for each local patch of input LR (see Fig. 3);
- *Feature Aggregation* aligns and fuses the searched key features and generates enhanced feature maps for HR reconstruction (see Fig. 4).

Feature Searching. Before matching most similar features, we firstly apply bicubic up-sampling on the LR image I^{LR} to get an up-scaled one $I^{LR\uparrow}$ with the same spatial size as I^{HR}. Then, we use the same feature extractor $\mathcal{F}$ to generate LR features $F = \mathcal{F}(I^{LR\uparrow})_{l_q}$ with size $N^l \times D^q$, where each feature point $f_i \in F$ represents a 3×3 patch on the LR feature map. Then, $\forall f_i \in F$, we find indices $\mathcal{N}(f_i)$ of K most similar features in $\hat{\boldsymbol{F}}^q$ by ranking their normalized inner product similarities [35]. After feature searching, index vectors $\{\mathcal{N}(f_i)\}_{i=1}^{N^l}$ are

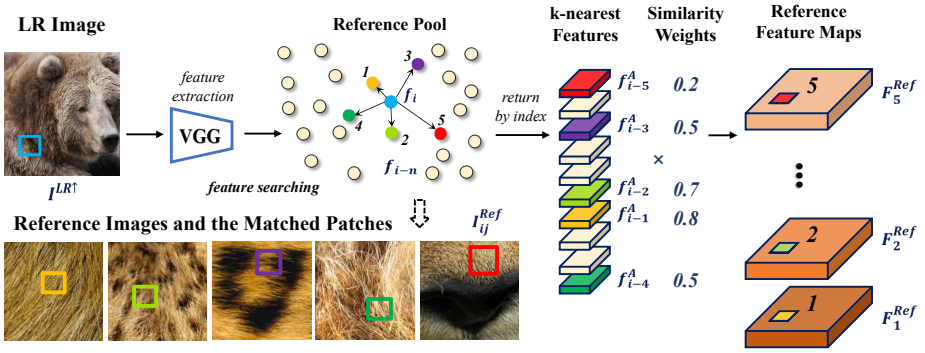

Fig. 3. Feature searching and realignment with $K = 5$. It shows features of i-th LR local patch (blue) corresponds to 5 points in query features (e.g., none-white points), which will generate 5 selected reference feature maps by assembling the key features. Besides, i-th position on reference feature maps can be obtained by multiplying the selected key features with their corresponding similarity weights.

stacked by rows into a global index matrix $\mathcal{N}(F)$.

Feature Aggregation. After feature searching, we use global index matrix $\mathcal{N}(F)$ to choose key features according to their corresponding query features (see the mapping in Fig. 2 (a)). The overlaps of adjacent key features on the rearranged feature map will be divided by their overlap number. Then, we obtain K aligned reference feature maps $\boldsymbol{F}^A = \{F_k^A\}_{k=1}^K$. Besides, the normalized similarity score of each query feature between LR features is also rearranged to K aligned similarity matrix $\boldsymbol{S}^A = \{S_k^A\}_{k=1}^K$. The i-th element of the k-th matrix S_k^A records the inner product between the k-th query feature patch and the i-th LR feature patch, which is corresponding to the i-th patch in $\boldsymbol{F}_k^A$. Finally, we generate reference feature maps by multiplying aligned reference feature maps with corresponding element-wise similarity. Concretely, the i-th patch of the k-th reference feature map is obtained by

$$f_{i-k}^{Ref} = f_{i-k}^A \cdot S_{i-k}^A, \ k = 1, ..., K \tag{2}$$

Therefore, the LFE module can be constructed as a general component for any SR backbone (collectively called SRNet). Fig. 4 shows that conducting a LFE module after the last pixel shuffle layer of SRNet (the feature map with original scale). For input LR image I^{LR}, it obtains K reference feature maps $\boldsymbol{F}^{Ref}$ through feature searching. After that, to obtain enhanced features, it aggregate features by using hidden output of the SRNet (F^H) and K reference feature maps,

$$F^P = \mathcal{P}_{k=1}^K \{\mathcal{R}(F^H || F_k^{Ref})\}, \tag{3}$$

where $(\cdot||\cdot)$ denotes concatenate operation and $\mathcal{P}, \mathcal{R}$ denotes element-wise addition and passing through share-weighted residual blocks.

As shown in Fig. 4 and Eq. (3), the feature aggregation firstly concatenates the hidden features of SRNet with each reference feature map respectively. Then

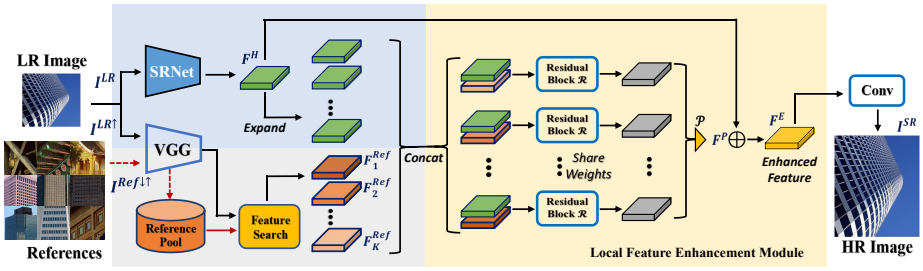

Fig. 4. Inner structure of LFE module. Blue and grey part illustrate the SISR network and the procedure of feature searching from RP detailed in Fig. 2 and 3. Yellow part depicts the Local Feature Enhancement (LFE) module. LFE module can be a general component embedded in any SRNet, which aggregates hidden output of SRNet with reference feature maps to an enhanced feature map.

a set of share-weighted residual blocks is used for each combined feature map. Finally, a fusion function aggregates information of all reference features to a single fused feature map, which retains the most informative texture feature in all reference feature maps and greatly facilitates the reconstruction of HR. Besides, a skip-connection will be used between hidden features and fused features, generating an enhanced feature map F^E by conducting element-wise addition between F^P and F^H. After a convolution layer changes the channel dimension, the output I^{SR} can be obtained. It should be noted that here we just give a simple case of the LFE module, which is a plug-and-play component and can be used in other positions of SRNet (i.e., all of three scales in 4× upscaling).

3.3 Loss Function

To recover the mapping from LR to HR and leverage the natural texture from references, l_1-norm content loss $\mathcal{L}_c$ is adopted to optimize weights of our model,

$$\mathcal{L}_c = \| I^{SR} - I^{HR} \|_1, \tag{4}$$

where I^{SR} and I^{HR} denotes output and ground truth of single sample.

Furthermore, we modify texture matching loss [19] to multi-reference scenario and take into account the texture difference between I^{SR} and F^{Ref}. We require the texture of neural feature space $\mathcal{F}(I^{SR})$ to be close to each aligned reference feature maps, where $\mathcal{F}(\cdot)_{l_k}$ denotes the same feature extractor in key features generation. Specifically, we define a texture loss $\mathcal{L}_{tex}$ as

$$\mathcal{L}_{tex} = \frac{1}{K} \sum_{k=1}^{K} \| Gr(\mathcal{F}(I^{SR})_{l_k} \cdot S_k^A) - Gr(F_k^A \cdot S_k^A) \|_F, \tag{5}$$

where $Gr(\cdot)$ computes the Gram matrix, and K normalized similarity matrix defined in Eq. (2) are used to determine the effect of each position. Intuitively, textures dissimilar to I^{SR} will have lower weight, and thus receiving lower penalty in texture learning.

Apart from content loss and texture loss, generative adversarial loss [10] $\mathcal{L}_{adv}$ will also be utilized upon the WGAN [1] with gradient penalty. This helps stabilize the training process and improve the visual quality of synthesised images. Furthermore, we adopt the perceptual loss [2] to motivate our model inclined to the solutions in the manifold of natural images. The perceptual loss $\mathcal{L}_p$ measures the distance in the feature space and enforce the feature alignment. Combining all together, the overall loss function of our model is finally designed as

$$\mathcal{L} = \mathcal{L}_c + \gamma_{tex}\mathcal{L}_{tex} + \gamma_{adv}\mathcal{L}_{adv} + \gamma_p\mathcal{L}_p, \tag{6}$$

where weights γ_p , γ_{adv} and γ_{tex} are $1e-4$, $1e-6$ and $1e-4$, respectively.

3.4 Network Architecture

During the experiment, we use MDSR [15] as our SRNet. We set the residual block number equal to 80 and hidden layer channel number equal to 64. Considering the efficiency and effeteness, we use $K = 3$ for feature searching. The feature extractor VGG-19 is pre-trained on ImageNet, whose `conv3_2` is used for query feature while all of `conv1_2`, `conv2_2` and `conv3_2` are responsible for key feature representations. Therefore, for the $4\times$ upscaling SR, we conduct three LFE modules on different scales. On each scale feature map (i.e., $4\times$ down-scale, $2\times$ down-scale or original scale), we generate K reference feature maps according to search results of query features. The architecture of share-weighted residue blocks is the same as the neural texture transfer in [35].

For RP generation, we directly select 300 HR-Refs from Outdoor Scene (OST) [23] as our reference, which is a dataset for scene images reconstruction. It contains seven categories, sky, mountain, plant, grass, water, animal, and building. For each category, there are 1k to 2k images that only cover that category. The total amount of the training set is 10,324. Here, we adopt a training set with explicit semantic prior for our RP construction. We firstly randomly crop 1,000 HR candidates with size 128×128 in each category, then we use the FPS algorithm to sample final 300 of them as our initial reference pool.

4 Experiment Results

4.1 Dataset

Following the setting in [35,36], we trained our model on CUFED5 dataset and test with down-scale factor $4\times$ on three standard benchmark datasets: Urban100 [9], Sun80 [22] and CUFED5 test set. Noted that CUFED5 only contains 13,761 160×160 input-reference pairs as the training set and 126 images with different sizes as the test set. In CUFED5, most images have relatively low resolution, and there are many moving people and complicated objects in each image, which makes the training and testing on CUFED5 extremely challenging.

4.2 Implementation Details

To further verify the rationality of our model, we test our model on both content-independent and content-similar references. For the former, we use our RP as extra information, which will be used in both training and evaluation processes. As for content-similar references, we conduct a similar training strategy in [35] and use reference images with different similarity levels during evaluation.

Content-independent references. Before the training, we use the pre-trained VGG to collect key features and query features offline from the initial RP. Then we use Farthest Point Sampling with the sample ratio factor $r = 16$ on both query and key features to generate final RP and save key features for further usage. After that, we conduct feature searching and save the global indices $\mathcal{N}(F)$ for each LR image. We adopt strategy mentioned in [35] to match K query features with the largest inner product for each LR patch, which can be implemented by a convolution operator using the LR patch as the kernel. During the training, we feed input images and corresponding global indices into our model, and the model selects features from preserved key features and synthesizes the aligned reference feature maps. Unlike [35] uses offline storage for feature maps of the entire training set, we only save entire key features and the mapping index of each local region. This strategy greatly reduces memory consumption and achieves efficiency in data augmentation at the same time. Besides, calculating indices offline will notably accelerate training speed and require almost no extra time in data feeding. More detailed training protocol will be described in the supplementary material.

Content-similar references. The overall strategy is consistent with the process mentioned above, but here we use given reference images to generate and save aligned reference feature maps offline rather than global indices. At the same time, specific data augmentation (the "warp" in the caption of Tab. 2) is conducted on each reference image.

4.3 Quantitative Evaluation

Content-independent references. Following standard protocols, we obtain all LR images by bicubic down-scaling (4×) from the HR images. For fair comparison on PSNR/SSIM with those methods mainly minimizing MSE, e.g., MDSR [15] and SRNTT-ℓ_2 [35], we first train a PSNR-oriented model emphasizing on the ℓ_2 minimization, called CIMR-ℓ_2. Next, we train a GAN-based model named CIMR, focusing on the aspect of visual quality compared with other methods with GAN fine-tuning. Since [35] uses content-similar references in their paper, we implement SRNTT*-ℓ_2 by using reference patches in initial RP to compare their model in content-independent references scenario. Specifically, we use FPS to sample 80 patches from initial RP as their references (i.e., the maximum amount of references SRNTT could hold). At the same time, we compare the result reported in their paper.

As shown in Tab. 1, it is obvious that our model gains higher scores on all the benchmark datasets. We achieve better performance than SRNTT-ℓ_2

Table 1. PSNR/SSIM comparison among different SR methods on four datasets: methods are grouped by SISR (top) and RefSR (bottom), and the best result is in bold. All the SR results are evaluated by PSNR and SSIM metrics on the Y channel of transformed YCbCr space.

Algorithm	CUFED5	Urban100	Sun80
Bicubic	24.18 / 0.684	23.14 / 0.674	27.24 / 0.739
SRCNN [4]	25.33 / 0.745	24.41 / 0.738	28.26 / 0.781
SCN [26]	25.45 / 0.743	24.52 / 0.741	27.93 / 0.786
DRCN [12]	25.26 / 0.734	25.14 / 0.760	27.84 / 0.785
LapSRN [13]	24.92 / 0.730	24.26 / 0.735	27.70 / 0.783
MDSR [15]	25.93 / 0.777	25.51 / 0.783	28.52 / 0.792
EDSR [15]	25.90 / 0.776	25.50 / 0.783	28.49 / 0.789
SRGAN [14]	24.40 / 0.702	24.07 / 0.729	26.76 / 0.725
RCAN [33]	**26.32 / 0.789**	**25.65 / 0.785**	**28.67 / 0.795**
SAN [3]	26.29 / 0.789	25.63 / 0.783	28.66 / 0.795
LandMark [30]	24.91 / 0.718	-	27.68 / 0.776
CrossNet [36]	25.48 / 0.764	25.11 / 0.764	28.52 / 0.793
SRNTT [35]	25.61 / 0.764	25.09 / 0.774	27.59 / 0.756
CIMR	26.16 / 0.781	25.24 / 0.778	29.67 / 0.806
SRNTT*-ℓ_2 [35]	25.98 / 0.776	25.54 / 0.784	28.49 / 0.791
SRNTT-ℓ_2 [35]	26.24 / 0.784	25.50 / 0.783	28.54 / 0.793
CIMR-ℓ_2	**26.35 / 0.789**	**25.77 / 0.792**	**30.07 / 0.813**

on CUFED5 even if they use references with high similarity of input, which is usually invalid and impractical in the real world. Despite the nonexistence of references with high similarity, it is surprising to notice our performance is much higher than previous methods on Sun80. This success may owe to that many outdoor-related textures like the wave, sky and vegetation are adaptively covered in our constructed RP, strengthening the applicability for outdoor scenarios. Furthermore, our method achieves a large improvement compared with the baseline method MDSR [15].

Content-similar references. Although we achieve satisfactory performance in content-independent references evaluation, to further investigate our performance when meeting content-similar references, we designed the experiment using reference images with different similarity levels. Specifically, there are four similar level HR-Refs in the CUFED5 test set, ranked by SIFT [17] feature matching, which decline from L1 to L4. We also compare results using augmented HR and all the four HR-Refs in Tab. 2.

As shown in Tab. 2, we compare CIMR, CrossNet, and SRNTT in both PSNR-oriented and GAN implementation, where our results are much higher. For the HR (warp) column in Tab. 2, the ability of our model for directly transferring information on HR itself is slightly worse than SRNTT. This is because our model finds multiple reference patches for each LR local region, but it is worthy note that there is no practical value in using HR itself for SR.

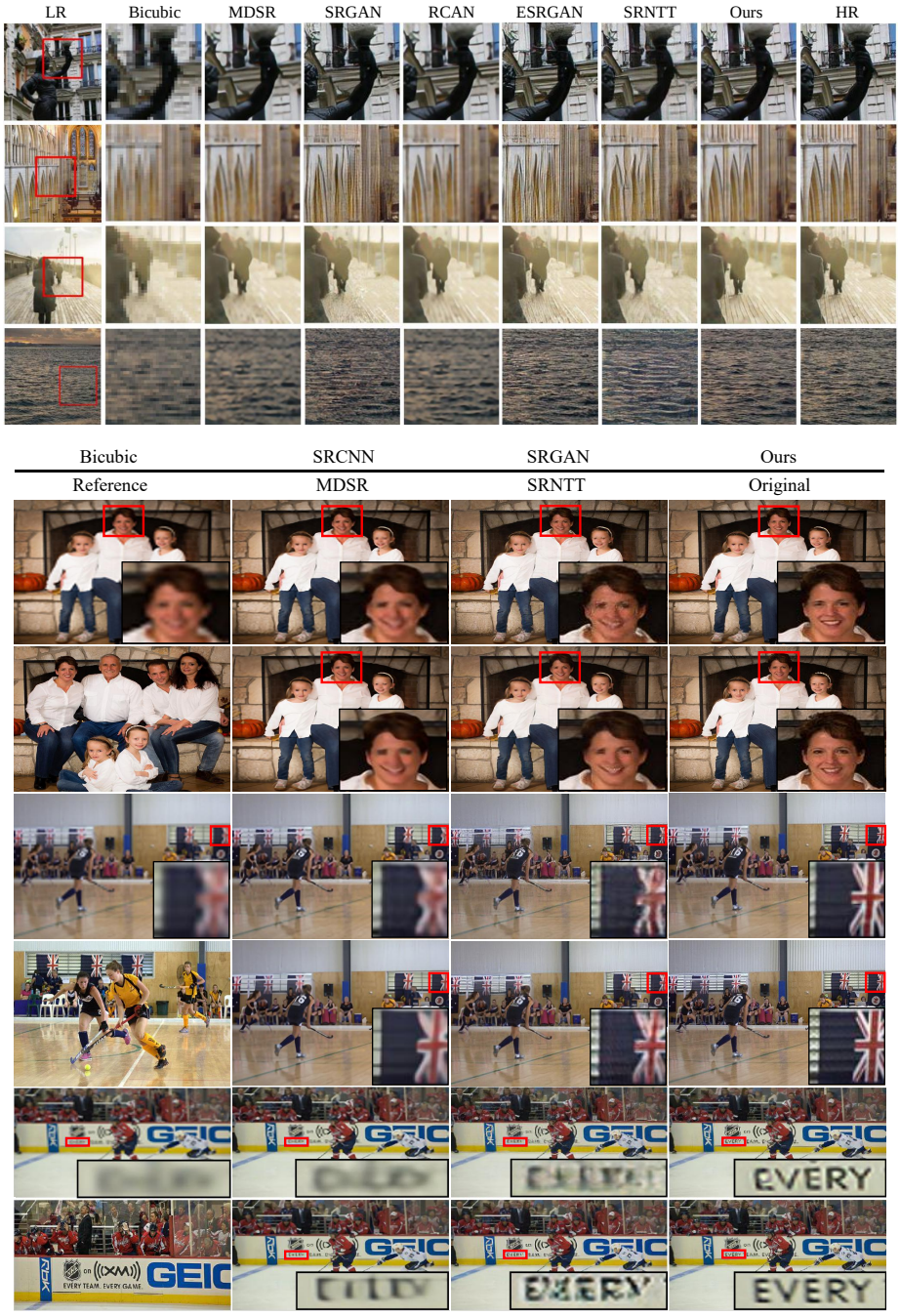

Fig. 5. Upper part (content-independent references): visualization of 003, 032 and 065 in Urban100 and N00_17_0_34matches29 in CUFED [25]. **Lower part** (content-similar references): visualization of 047_0, 044_0, 002_0 in CUFED5, where our CIMR-SR and SRNTT both use content-similar references (i.e., 047_3, 044_3, 002_1).

Table 2. PSNR/SSIM at different reference levels on CUFED5 dataset. The "warp" denotes the data augmentation with random translation (quarter to half width/height), rotation (10∼30 degree), and scaling (1.2∼2.0× upscaling) from the original HR image.

Algorithm	HR (warp)	L1	L2	L3	L4	All
CrossNet	25.49/.764	25.48/.764	25.48/.764	25.47/.763	25.46/.763	-
SRNTT-ℓ_2	29.29/.889	26.15/.781	26.04/.776	25.98/.775	25.95/.774	26.24/.784
CIMR-ℓ_2	29.82/.903	**27.32/.805**	**27.05/.799**	**26.92/.796**	**26.86/.794**	**27.44/.810**
SRNTT	**33.87/.959**	25.42/.758	25.32/.752	25.24/.751	25.23/.750	25.61/.764
CIMR	30.73/.918	26.50/.786	26.47/.784	26.45/.784	26.44/.784	26.63/.790

4.4 Qualitative Evaluation

Content-independent references. For the fair comparison, the state-of-the-art SISR and RefSR methods we choose are MDSR [15], SRGAN [14], ESRGAN [24] and RCAN [33], among which RCAN and MDSR are the most powerful SISR methods with the highest PSNR/SSIM scores. The SRGAN and ESRGAN could achieve a satisfactory performance on visual quality because of adversarial learning. SRNTT* with the same setting in the quantitative experiment is included as the representative of state-of-the-art RefSR.

As shown in the upper part of Fig. 5, although PSNR-oriented methods like RCAN and MDSR could present higher criteria, they tend to produce blurry textures while preserving sharp edges. SRGAN and ESRGAN could largely improve the high-frequency details since the generative adversarial learning strategy. However, they tend to generate unnatural textures, like the noise around the sculpture. SRNTT [35] is a powerful technique that could produce an extremely visually pleasing result and high criteria when the HR-Ref images and the LR image share the same scene. However, due to the fact that our sampled patches for SRNTT include some water wave patterns from the swimming pool, the generated water waves in SRNTT are a little bit brighter and tend to present monotonous blue, which indicates the inferior feature transferring capability of patch swapping in SRNTT. In contrast, our method employed with RP and LFE module could lead to more natural and realistic textures when reconstructing the head of the sculpture and boardwalk.

Content-similar references. We further compare qualitative evaluation with similar references. We selected three samples from CUFED5 and compared SRCNN [4], SRGAN [14], SRNTT [35], and our baseline MDSR [15] in the lower part of Fig. 5. We achieve better results in texture details and a huge improvement over baseline. More visualization results will be provided in the supplementary material.

4.5 Ablation Study

In this section, we investigate the effectiveness of utilizing multiple scales key features as compared to using a single scale. We use the feature maps extracted

Table 3. Comparison of adopting different `conv` layers for CIMR. Left part shows PSNR at different reference levels on CUFED5 test set. Right part describes running time, including offline feature searching and forward inference time.

Method-ℓ_2	HR(warp)	L1	L2	L3	L4	FS	Forward
SRNTT	29.29	26.15	26.04	25.98	25.95	2.726	**0.351**
`conv1`	28.24	26.97	26.89	26.84	26.79	2.045	0.783
`conv2`	28.77	27.06	26.93	26.92	26.85	1.551	0.535
`conv3`	27.31	26.26	26.06	26.02	26.03	**1.020**	0.361
`conv2/3`	27.87	27.27	27.01	26.89	26.82	2.571	1.077
`conv1/2/3`	**29.82**	**27.32**	**27.05**	**26.92**	**26.86**	4.616	1.551

from `conv1_2`, `conv2_2` and `conv3_2` to generate key features and feed them into corresponding LFE modules. In addition to quantitative metrics, comparison in the speed of feature searching (FS, also represents patch swap in SRNTT) (s/sample), model forwarding (s/batch) during the training process with batch size 8 are applied to measure the efficiency of our model.

In Tab. 3, it shows that adding key features in $2\times$ scale hidden layer of SRNet obtains the best improvement while using key features from `conv3_2` perform the worst. Furthermore, in our proposed CIMR-SR model, we also provide users with more flexible options to achieve further improvement according to specific requirements, at the cost of importing more LFE modules and more computations. It should be noted that even if we only use the key features from `conv3_2`, we can achieve comparable results with SRNTT while increase the speed by twice. These results fully demonstrate the effectiveness of our model to enhance LR from multiple similar patches in feature space.

5 Conclusion

To our best knowledge, this is the first work to deal with arbitrary multiple references oriented image super-resolution problem with deep learning. To achieve this goal, we proposed a Content-Independent Multi-Reference Super-Resolution (CIMR-SR) model. It can adaptively match local patterns from a universal reference pool and aggregate them in the feature space by the LFE module to strengthen the discriminative learning ability on representing the LR image. Extensive experiments demonstrate that our proposed CIMR-SR model can achieve better quantitative results and generate realistic images with more details as well, outperforming the state-of-the-art methods.

Acknowledgements: The work was supported in part by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the National Key R&D Program of China with grant No. 2018YFB1800800, by Natural Science Foundation of China with grant NSFC-61629101, by Guangdong Zhujiang Project No. 2017ZT07X152, by Shenzhen Key Lab Fund No. ZDSYS201707 251409055, by NSFC-Youth 61902335, Guangdong Province Basic and Applied Basic Research Fund Project Regional Joint Fund-Key Project No. 2019B151512 0039 and CCF-Tencent Open Fund.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017)
2. Bruna, J., Sprechmann, P., LeCun, Y.: Super-resolution with deep convolutional sufficient statistics. arXiv preprint arXiv:1511.05666 (2015)
3. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11065–11074 (2019)
4. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European conference on computer vision. pp. 184–199. Springer (2014)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015)
6. Freedman, G., Fattal, R.: Image and video upscaling from local self-examples. ACM Transactions on Graphics (TOG) **30**(2), 12 (2011)
7. Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-based super-resolution. IEEE Computer graphics and Applications **22**(2), 56–65 (2002)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5197–5206 (2015)
10. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
11. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016)
12. Kim, J., Kwon Lee, J., Mu Lee, K.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1637–1645 (2016)
13. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 624–632 (2017)
14. 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. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
15. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
16. Liu, C., Sun, D.: A bayesian approach to adaptive video super resolution. In: CVPR 2011. pp. 209–216. IEEE (2011)
17. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the seventh IEEE international conference on computer vision. vol. 2, pp. 1150–1157. Ieee (1999)

18. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660 (2017)
19. Sajjadi, M.S., Scholkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4491–4500 (2017)
20. Shaham, T.R., Dekel, T., Michaeli, T.: Singan: Learning a generative model from a single natural image. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4570–4580 (2019)
21. Shocher, A., Cohen, N., Irani, M.: "zero-shot" super-resolution using deep internal learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3118–3126 (2018)
22. Sun, L., Hays, J.: Super-resolution from internet-scale scene matching. In: 2012 IEEE International Conference on Computational Photography (ICCP). pp. 1–12. IEEE (2012)
23. Wang, X., Yu, K., Dong, C., Change Loy, C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 606–615 (2018)
24. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
25. Wang, Y., Lin, Z., Shen, X., Mech, R., Miller, G., Cottrell, G.W.: Event-specific image importance. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
26. Wang, Z., Liu, D., Yang, J., Han, W., Huang, T.: Deep networks for image super-resolution with sparse prior. In: Proceedings of the IEEE international conference on computer vision. pp. 370–378 (2015)
27. Xu, X., Ma, Y., Sun, W.: Towards real scene super-resolution with raw images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1723–1731 (2019)
28. Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., Yang, M.H.: Learning to super-resolve blurry face and text images. In: Proceedings of the IEEE international conference on computer vision. pp. 251–260 (2017)
29. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE transactions on image processing **19**(11), 2861–2873 (2010)
30. Yue, H., Sun, X., Yang, J., Wu, F.: Landmark image super-resolution by retrieving web images. IEEE Transactions on Image Processing **22**(12), 4865–4878 (2013)
31. Zhang, K., Zuo, W., Zhang, L.: Learning a single convolutional super-resolution network for multiple degradations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3262–3271 (2018)
32. Zhang, X., Chen, Q., Ng, R., Koltun, V.: Zoom to learn, learn to zoom. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3762–3770 (2019)
33. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 286–301 (2018)
34. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2472–2481 (2018)

35. Zhang, Z., Wang, Z., Lin, Z., Qi, H.: Image super-resolution by neural texture transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7982–7991 (2019)
36. Zheng, H., Ji, M., Wang, H., Liu, Y., Fang, L.: Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 88–104 (2018)