LMR: A Large-Scale Multi-Reference Dataset for Reference-based Super-Resolution

Lin Zhang* [1,3,4,5], Xin Li[2], Dongliang He[2], Fu Li[2], Errui Ding[2], Zhaoxiang Zhang[1,3,4,6]
[1] University of Chinese Academy of Sciences
[2] Department of Computer Vision Technology (VIS), Baidu Inc.
[3] Institute of Automation, Chinese Academy of Sciences
[4] State Key Laboratory of Multimodal Artificial Intelligence Systems
[5] School of Future Technology, UCAS
[6] Center for Artificial Intelligence and Robotics, HKISI_CAS

Abstract

*It is widely agreed that reference-based super-resolution (RefSR) achieves superior results by referring to similar high quality images, compared to single image super-resolution (SISR). Intuitively, the more references, the better performance. However, previous RefSR methods have all focused on single-reference image training, while multiple reference images are often available in testing or practical applications. The root cause of such training-testing mismatch is the absence of publicly available multi-reference SR training datasets, which greatly hinders research efforts on multi-reference super-resolution. To this end, we construct a large-scale, multi-reference super-resolution dataset, named **LMR**. It contains 112,142 groups of 300×300 training images, which is 10× of the existing largest RefSR dataset. The image size is also some times larger. More importantly, each group is equipped with 5 reference images with different similarity levels. Furthermore, we propose a new baseline method for multi-reference super-resolution: **MRefSR**, including a **M**ulti-Reference **A**ttention **M**odule (MAM) for feature fusion of an arbitrary number of reference images, and a **S**patial **A**ware **F**iltering **M**odule (SAFM) for the fused feature selection. The proposed MRefSR achieves significant improvements over state-of-the-art approaches on both quantitative and qualitative evaluations. Our code and data are available at: https://github.com/wdmwhh/MRefSR.*

1. Introduction

Single image super-resolution (SISR) is to restore a degraded low-resolution (LR) image to a texture-realistic high-resolution (HR) image [11]. SISR has a wide range of applications in surveillance [40], astronomy [8], medical imaging [7], film and television [24, 14], and other industries [27, 38, 34]. With the development of deep learning, SISR has made great progress over these years [4, 5, 16, 19, 13, 41, 32, 20, 3, 23]. Compared with SISR, reference-based super-resolution (RefSR) can leverage relevant textures from additional similar HR reference images, so it often achieves better performance. Similar high-definition images can be acquired from web-searching, expertly curated data repositories or websites, etc.

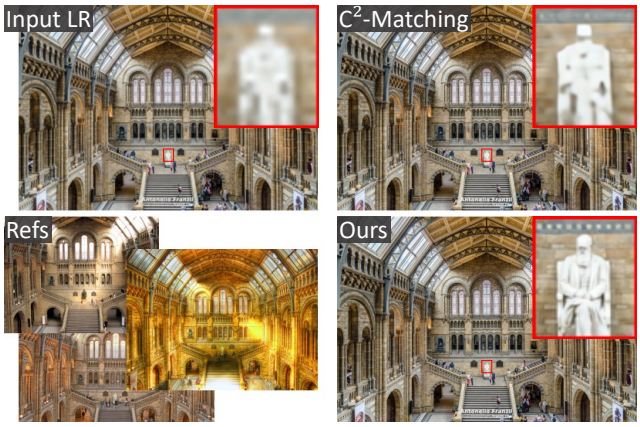

Figure 1. Visual comparison of single-reference training RefSR method C^2-Matching [12] and our multi-reference training MRefSR. Our MRefSR can more fully utilize arbitrary number of multiple reference images to achieve the best results. The full-resolution comparison is provided in supplementary material.

Because of promising results shown by recent RefSR methods [29, 37, 43, 42, 28, 30, 12, 21, 35], it attracts more and more research interest. However, all these previous RefSR methods have focused on using a single reference

*Work done during an internship at Baidu Inc.

image for training, but there are often multiple reference images available for testing or practical applications. To the best of our knowledge, the only RefSR training dataset currently available is CUFED5 [42, 33], which has only 11,871 image pairs with a small resolution of 160×160. More importantly, there is only one reference image for each LR input image. However, in practical applications, multiple reference images are often encountered. For example, testing set of CUFED5 has 126 input images and each has 5 reference images with different similarity levels. Similarly, we can also easily find multiple reference images for any real test case. Due to the limitation of the only available training dataset, previous RefSR methods do not make good use of multiple reference images in testing or practical applications. The previous RefSR methods usually stitch together several reference images to get a large resolution image as one reference image to fit the models trained with only one reference image. Nevertheless, if the resolution of the reference images is too large, this way of testing will exhaust the GPU memory. Furthermore, the relationship among multiple reference images is not modeled effectively. So this is certainly much worse than a method designed specifically for multiple reference images. Therefore, a multi-reference RefSR training dataset and a simple but effective multi-reference RefSR method are needed.

In this paper, we propose a large-scale, multi-reference RefSR dataset, named LMR. The training set of LMR consists of 112,142 groups of 300×300 training images, each group containing 5 reference images of different similarity levels. LMR training dataset has 10 times images more than CUFED5, and the image size is also some times larger. Such a sufficiently large training dataset will be beneficial for improving the generalization ability of models. We believe this training dataset will greatly facilitate the RefSR research as it is the first RefSR training dataset with multiple reference images. Meanwhile, the testing set of LMR has 142 groups of images and each group with 2∼6 reference images. The side length of the testing images ranges from 800 to 1600.

With the help of LMR, we propose a new RefSR baseline method for multiple reference RefSR, named MRefSR. First, we develop a **M**ulti-Reference **A**ttention **M**odule (MAM) for feature fusion from an arbitrary number of reference images. We treat the LR input feature as query, and candidate keys and values are generated from the aligned reference features corresponding to different reference images. Then, attention across different aligned reference features is conducted to fuse features from different reference images. Second, since not all LR feature points can well match the reference features, we use **S**patial **A**ware **F**iltering **M**odule (SAFM) for fused feature selection. As shown in Figure 1, our MRefSR effectively utilizes information from multiple reference images to produce visually pleasing details. In summary, our contributions are threefold:

- We contribute the first multi-reference RefSR dataset, named LMR, which contains 112,142 groups of 300×300 training images and each group has 5 reference images for the input image. This dataset will enable RefSR research from single-reference to multi-reference images and largely promote the development of the RefSR research field.

- We propose a novel multi-reference baseline RefSR method MRefSR, using a multi-reference attention module for feature fusion of an arbitrary number of reference images, and a spatial aware filtering module for the fused feature selection. Our method effectively learns the relationship among multiple references and makes the best use of them, this is also thanks to the multi-reference dataset LRM.

- We conduct extensive experiments which demonstrate the superiority of the proposed LMR and the potential of multi-reference RefSR methods. Our method achieves significant improvements over state-of-the-art approaches on both quantitative and qualitative evaluations.

2. Related Work

2.1. Reference-based Image Super-Resolution

RefSR is gradually becoming an emerging research field. Compared with SISR, RefSR is more advantageous because it can utilize the information of additional HR reference images with similar contents. SRNTT [42] proposed an end-to-end network structure that performs multi-scale adaptive texture transfer from the reference image to recover the SR image. Subsequently, TTSR [36] applied a cross-scale feature integration method to merge multi-scale reference features. MASA [21] designed a coarse-to-fine patch matching scheme to reduce the computational complexity. Consequently, C^2-Matching [12] got more accurate pre-offsets of reference features to LR features by a teacher-student correlation distillation and a dynamic DCN [2, 44] aggregation module. AMSA [35] made an incremental extension of C^2-Matching by introducing multi-scale aggregation and coarse-to-fine patch matching. Huang *et al*. [10] also used the C^2-Matching model, but added an additional SISR network to decouple the texture transfer and the super-resolution, which made the network parameters much larger and the inference much slower. Recently, RRSR [39] and DATSR [1] also introduce reciprocal learning and transformers to boost the performance. Although previous methods have made great progress, all of the above methods focus on research exploration using only one single reference

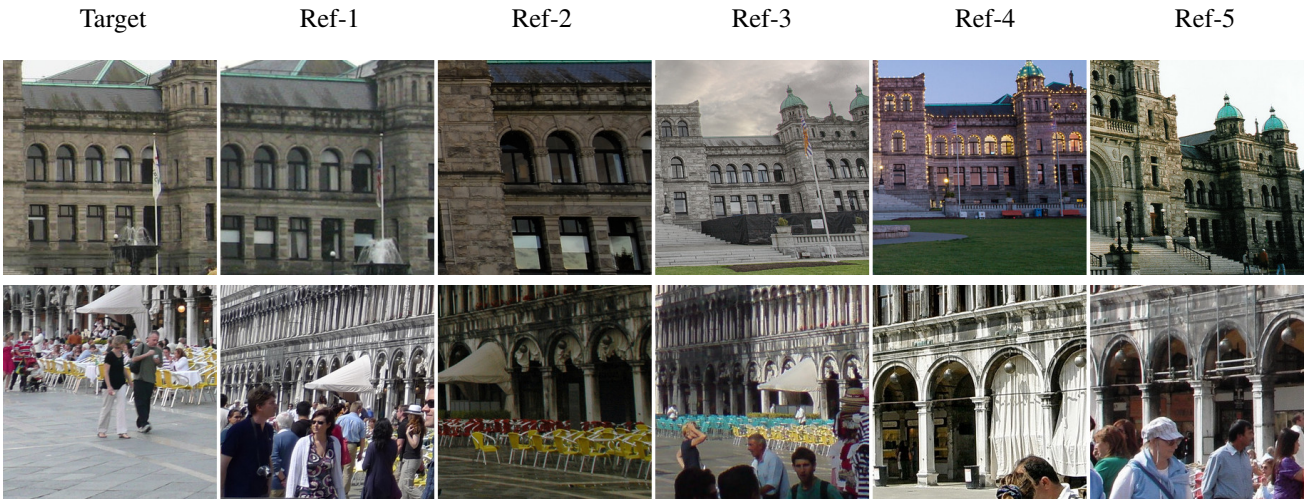

Figure 2. Two groups of sample images from our LMR training dataset. From left to right, there is one target image, one high-similarity (H) reference image, two medium-similarity (M) reference images, and two low-similarity (L) reference images.

image due to the limitation of the only available training dataset, CUFED5.

2.2. RefSR Datasets

To the best of our knowledge, there are five datasets commonly used in RefSR research: Sun80 [29], Urban100 [9], Manga109 [22], WR-SR [12] and CUFED5 [42, 33]. However, the first four are all testing sets. The Sun80 dataset contains 80 natural images, each with 20 web-search reference images, but these reference images are not very similar to the corresponding LR input, so it is not suitable as a testing set for RefSR. The Urban100 dataset contains 100 building images, lacking references. Because of self-similarity in the building image, the corresponding LR image is usually treated as the reference image. The Manga109 dataset contains 109 manga images without references. Since all the images in Manga109 are the same category (manga cover), the previous methods randomly use one HR image in the dataset as a reference image. The WR-SR dataset with more diverse categories, contains 80 image pairs, each target image accompanied by a web-searching reference image. CUFED5 [42, 33] is the only dataset with a training set, which has 11,871 image pairs with a small resolution of 160×160 and only one reference image for the LR input in each image pair. CUFED5 testing set has 126 input images and each has 5 reference images with different similarity levels. Recently, Wang *et al*. [31] proposed a new dataset named CameraFusion for dual-camera super-resolution with 131 training image pairs and 15 testing image pairs. Similarly, the RefVSR work [17] focuses on the multiple-camera setting with fixed camera relative positions. However, the images captured by specific cameras is too ideal for the RefSR task. In this paper, to better meet the demands of RefSR research, we propose LMR, a large-scale multi-reference RefSR dataset.

3. Approach

In this section, we first introduce the proposed **L**arge-scale **M**ulti-reference **R**efSR dataset **LMR** in Sec. 3.1. Subsequently, we detail a new baseline RefSR method **MRefSR** using multiple references in Sec. 3.2.

3.1. Construction of LMR

The MegaDepth [18] dataset was originally proposed for single-view depth prediction. They used a large number of Internet images from overlapping viewpoints to obtain the dense depth by COLMAP, a state-of-the-art SfM system [25] (for reconstructing camera poses and sparse point clouds) and MVS system [26] (for generating dense depth maps). The generated dense depth maps of the COLMAP are used as the supervised targets for single-view depth prediction model training. MegaDepth contains 1,070,468 internet photos of landmarks around the world and reconstructs 196 3D landmark models from these photos. Each photo of the same landmark varies widely in viewpoint, scene extent, and focused buildings. The scene of finding Internet images from overlapping viewpoints for 3D reconstruction is very similar to finding reference images for target images to do reference-based super-resolution. Inspired by this, the image groups in the off-the-shelf MegaDepth dataset are very suitable for making a RefSR dataset. Consequently, we propose a new large-scale multi-reference RefSR dataset, dubbed LMR.

To construct the LMR training image patch groups, we first perform the following preprocessing steps on the original MegaDepth dataset to obtain similar image pairs.

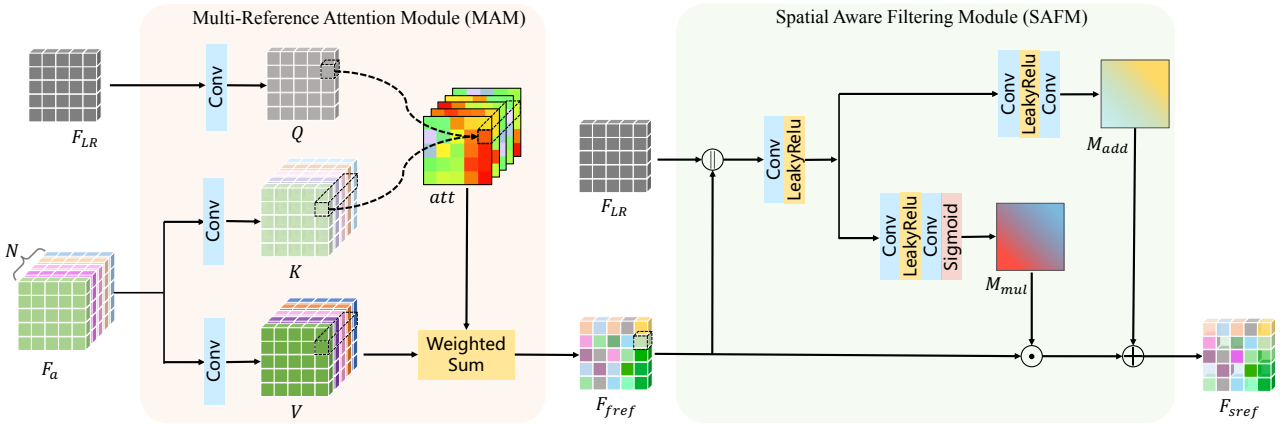

Figure 3. The proposed Multi-Reference Attention Module (left) for the multi-reference feature fusion and the Spatial Aware Filtering Module (right) for the fused feature selection. Both modules perform pixel-wise functions.

- Firstly, the PSNR of the target image and the candidate reference images should be lower than 30dB to filter duplicate images.
- Secondly, to ensure that the candidate reference images and the target image should have some similar contents, we filter them by controlling the overlap ratio R_{olp} of matched keypoints in the sparse 3D point clouds.
- Thirdly, the size ratio R_s of the same object in the reference image and the target image cannot be too small, otherwise the reference image cannot provide enough detailed texture information.

We calculate R_s and R_{olp} following the existing code in D2-Net [6], a method for image matching and 3D reconstruction.

Further, we define three similarity levels for these image pairs, that is high similarity (H), medium similarity (M) and low similarity (L). A image pair is categorized as H if the overlap ratio R_{olp} is greater than 30% and the size ratio R_s is larger than 0.9, M if R_{olp} is greater than 10% and the R_s is larger than 0.66, otherwise L. [1]

Through the above operations, we can obtain a large number of image groups, each containing one target image and multiple reference images. However, due to GPU memory limitation, it is often not possible to use the entire large image to train the network. For SISR, it is common to randomly crop a patch from the image for training. While in the case of RefSR, it is better to crop corresponding patches with similar contents in the reference images and the target image, e.g. CUFED5 cropped 11,871 paired 160×160 patches as the training set. For the multi-reference dataset LMR, we first randomly crop a patch from the target image. Then, we map the center point of the cropped patch into 3D sparse point cloud and pick up 5 keypoints near the mapped point, which are from 5 reference images with different similarities (one H, two M, two L). Next, we take the selected keypoints as centers and crop the corresponding patches. In this way, we collect a total of 112,142 groups of 300×300 patches as the training set. The number of collected training groups is ten times larger than that of training pairs of CUFED5, and the image size of collected training patches is nearly four times as large as that of training patches of CUFED5. More importantly, each group has 5 reference image patches of different similarities. Some representative samples are presented in Figure 2. As shown in Sec. 4, the model trained on the LMR dataset shows good generalization performance on other RefSR datasets, demonstrating the effectiveness of the LMR.

In addition to the LMR training set, we also prepare a testing set for multi-reference RefSR testing. We remove the images containing target or reference patches that appeared in the training set. From the remaining image pairs, we construct a testing set consisting of 142 groups, each containing a target image and $2\sim6$ reference images with image side lengths between 800 and 1600.

3.2. Multi-Reference RefSR network

Equipped with the LMR dataset, we propose a multi-reference RefSR network to make good use of multiple reference images, dubbed MRefSR. Our MRefSR is based on C^2-Matching [12] as it is currently the open source method with the best performance, and is easy to be started. Note that other RefSR frameworks such as TTSR [36] are also applicable since we aim to exploit multi-reference fea-

[1] These hyper-parameters of R_{olp} and R_s are set empirically. By grid search on R_{olp} and R_s, we balance the number of image pairs with different similarities.

tures instead of single-reference feature transfer. As C^2-Matching did, a *Content Extractor* (CE) is used to extract features F_{LR} from LR image. Multi-scale ($1\times$, $2\times$ and $4\times$) reference features $F^s_{Ref_i}$ are extracted by a *VGG* extractor, where $s = 1, 2, 4$ and $i \in \{1, 2, ..., N\}$, N is the number of reference images. For the sake of brevity, the s in the following is omitted, and F_{Ref_i} is used instead of $F^s_{Ref_i}$. A pretrained *Contrastive Correspondence Network* (CCN) is used to obtain the relative target offsets O_i of the LR input and the corresponding multiple reference images. Afterwards, as shown in Figure 3, we develop a **M**ulti-Reference **A**ttention **M**odule (MAM) for the multi-reference feature fusion and a **S**patial **A**ware **F**iltering **M**odule (SAFM) for the fused feature selection.

Dynamic Aggregation Module in C^2-Matching is used to get the aligned features F_{a_i} from the reference features F_{Ref_i} by the corresponding pre-offsets O_i. After that, we introduce MAM to fuse the aligned features from different reference images. In detail, at each feature scale, we first generate corresponding N attention maps for the aligned features of N reference images:

$$\begin{aligned} att_i(x, y) &= softmax(\langle Q(x, y), K_i(x, y)\rangle) \\ &= \frac{exp(\langle Q(x, y), K_i(x, y)\rangle)}{\sum_{j=1}^{N} exp(\langle Q(x, y), K_j(x, y)\rangle)}. \end{aligned} \tag{1}$$

We use inner product to measure the similarity between the features $Q(x, y)$ and $K_i(x, y)$ at the point (x, y), where query Q is obtained from the LR input feature F_{LR}, key K_i and value V_i are obtained from the i-th reference image aligned feature F_{a_i}:

$$\begin{aligned} Q &= conv_q(F_{LR}), \\ K_i &= conv_k(F_{a_i}), \\ V_i &= conv_v(F_{a_i}), \end{aligned} \tag{2}$$

where $conv_q$, $conv_k$ and $conv_v$ are convolutions with kernel size 3×3 and stride 1. Then, we get fused reference feature F_{fref} from all reference images:

$$F_{fref}(x, y) = \sum_{i=1}^{N} (att_i(x, y) \cdot V_i(x, y)). \tag{3}$$

The proposed MAM enables MRefSR to handle an arbitrary number of reference images during training and testing phases, making the MRefSR more flexible for practical applications.

Since not all LR feature pixels can be well matched with reference features, we use the proposed SAFM for the selection of fused reference features F_{fref}. As shown in Figure 3, we get two masks M_{mul} and M_{add} from the concatenated feature of F_{LR} and F_{fref} and a $sigmoid$ function is used to limit the range of the M_{mul}.

$$\begin{aligned} M_{mul} &= sigmoid(f_1(F_{LR} \| F_{fref})) \cdot 2, \\ M_{add} &= f_2(F_{LR} \| F_{fref}), \end{aligned} \tag{4}$$

where f_1 and f_2 are nonlinear mapping functions consisting of convolution and leaky ReLU layers. At last, the M_{mul} and M_{add} are used for the final selected reference features F_{sref}:

$$F_{sref} = F_{fref} \odot M_{mul} + M_{add}, \tag{5}$$

where $\odot$ denote element-wise multiplication.

In the end, a restoration module $\mathcal{G}$ takes the LR features F_{LR} and the selected reference features F_{sref} to reconstruct the target image:

$$X_{SR} = \mathcal{G}(F_{LR}, F_{sref}). \tag{6}$$

3.3. Implementation Details

We train and evaluate our MRefSR in a scale factor $4\times$. In detail, we train the network for 255K iterations using Adam optimizer [15] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and constant learning rate of 1e-4. Each mini-batch includes 48 groups of image patches, each consisting of an LR input patch with size 40×40 and five reference HR patches with size 160×160. We use three commonly used loss functions to train our model, including reconstruction loss L_{rec}, perceptual loss L_{per} , and adversarial loss L_{adv}, referring to supplementary material for the network training loss details. The weight coefficients for L_{rec}, L_{per} and L_{adv} are set to 1, 1e-4 and 1e-6. The network is first trained with L_{rec} only and then finetuned with all losses. During training, we augment the training data by randomly horizontally flipping and vertically flipping, and random 90° rotation. Following the standard protocol, we generate all LR images by bicubically downsampling the HR images with a scale factor of $4\times$. All experiments run in parallel on 4 NVIDIA V100 GPUs. For the quantitative comparison, we train MRefSR without GAN loss and perceptual loss as other methods did. Benefiting from the large-resolution training images of LMR, we get an LPF (Large Patch Finetuning) version of the model, which is finetuned using the large-patch training images.

4. Experiments

4.1. Datasets and Metrics

We train our network on the proposed LMR training set and evaluate it on the testing set of LMR, CUFED5 [42, 33], Sun80 [29] and WR-SR [12]. As mentioned earlier, LMR and CUFED5 are two real multi-reference testing sets. Although Sun80 has multiple reference images, these reference images are not very similar to the corresponding target images. WR-SR is a single-reference testing set. When testing on CUFED5, previous RefSR methods can stitch

Table 1. We report PSNR/SSIM on Y channel of YCbCR space to compare among different SR methods on the testing set of LMR, CUFED5 [42, 33], Sun80 [29], and WR-SR [12]. Methods are grouped by SISR methods (top) and reference-based methods (bottom). The best results are marked **in bold and underlined**. The second best and the third best results are marked in **bold** and underlined, respectively. C^2-Matching-LMR means C^2-Matching-rec is trained on the LMR dataset and the Ours-rec-LPF indicates that the model was finetuned using large patch size (300×300) training images.

Method	Training Dataset	LMR	CUFED5 [42, 33]	Sun80 [29]	WR-SR [28]
		PSNR↑ / SSIM↑	PSNR↑ / SSIM↑	PSNR↑ / SSIM↑	PSNR↑ / SSIM↑
SRCNN [4]	CUFED5	-	25.33 / 0.745	28.26 / 0.781	27.27 / 0.767
EDSR [19]	CUFED5	-	25.93 / 0.777	28.52 / 0.792	28.07 / 0.793
RCAN [41]	CUFED5	-	26.33 / 0.781	29.97 / 0.814	27.91 / 0.793
RRDB [32]	CUFED5	-	26.41 / 0.783	29.99 / 0.814	27.96 / 0.793
RCAN* [41]	LMR	29.63 / 0.841	26.58 / 0.785	**30.36 / 0.821**	28.24 / 0.798
RRDB* [32]	LMR	29.68 / 0.842	26.61 / 0.786	**30.37 / 0.821**	28.25 / 0.798
Landmark [37]	CUFED5	-	24.91 / 0.718	27.68 / 0.776	-
CrossNet [43]	CUFED5	-	25.48 / 0.764	28.52 / 0.793	-
SRNTT-rec [42]	CUFED5	-	26.24 / 0.784	28.54 / 0.793	27.59 / 0.780
TTSR-rec [36]	CUFED5	29.13 / 0.832	27.09 / 0.804	30.02 / 0.814	27.97 / 0.792
MASA-rec [21]	CUFED5	29.42 / 0.837	27.54 / 0.814	30.15 / 0.815	28.19 / 0.796
C^2-Matching-rec [12]	CUFED5	30.01 / 0.856	28.40 / 0.846	30.18 / 0.817	28.32 / 0.801
AMSA-rec [35]	CUFED5	-	28.50 / 0.849	30.29 / 0.819	-
TDF-rec [10]	CUFED5	-	28.64 / 0.850	30.31 / 0.820	28.52 / **0.807**
C^2-Matching-LMR	LMR	30.64 / 0.869	28.65 / 0.853	30.31 / 0.819	**28.53 / 0.807**
Ours-rec	LMR	**31.81 / 0.895**	**28.94 / 0.860**	30.28 / 0.819	28.52 / 0.806
Ours-rec-LPF	LMR	**31.98 / 0.898**	**29.05 / 0.862**	30.32 / 0.819	**28.59 / 0.807**

multiple reference images into one large reference image according to the convention. However, the image resolution of reference images in the testing sets of LMR and Sun80 is larger than that in CUFED5. So when other methods are tested on LMR and Sun80, each LR image is tested with only one reference image due to GPU memory limitation. With the multi-reference attention module (MAM), our MRefSR can utilize multiple reference images for prediction on the LMR, CUFED5 and Sun80 testing sets. we adopt two quantitative metrics, PSNR and SSIM, both calculated on Y channel in the transformed YCbCr color space. To evaluate the results qualitatively, we show the visual results of different methods and conduct a user study for subjective visual quality comparison.

4.2. Comparison with State-of-the-Art Methods

We compare the proposed MRefSR with previous state-of-the-art SISR methods and single-reference RefSR methods. SISR methods include SRCNN [4], EDSR [19], RCAN [41], RRDB [32] and ESRGAN [32]. As for single-reference RefSR methods, Landmark [37], CrossNet [43], SRNTT [42], TTSR [36], MASA [21], C^2-Matching [12], AMSA [35] and TDF [10] are included. For fair comparison, we retrain three high-performance SISR methods RCAN, RRDB and ESRGAN, and one open-sourced top-performing single-reference RefSR method C^2-Matching on the training set of LMR.

Quantitative evaluation. As shown in Table 1, our MRefSR outperforms other methods by a large margin on two real multiple reference datasets, CUFED5 and LMR. On the most commonly used CUFED5 benchmark, MRefSR outperforms the retrained C^2-Matching-LMR by 0.29dB. Models trained on LMR can achieve better performance on CUFED5, which also demonstrates the generalization ability and effectiveness of LMR. What's more, MRefSR shows a significant improvement of 1.15 dB over the second best method on the LMR testing set. The above two results demonstrate the superiority of learning the interaction among multiple references, further manifesting the necessity of the LMR dataset that enables multi-reference RefSR training. On Sun80, SISR methods RRDB* and RCAN* get the best two results. The results gap of the top RefSR methods AMSA-rec, TDF-rec, C^2-Matching-LMR and MRefSR are less than 0.04 dB, which further proves the reference image and its target image in Sun80 are not very similar. On the WR-SR benchmark, since there is only one reference image per LR, our results are very close to C^2-matching-LMR.

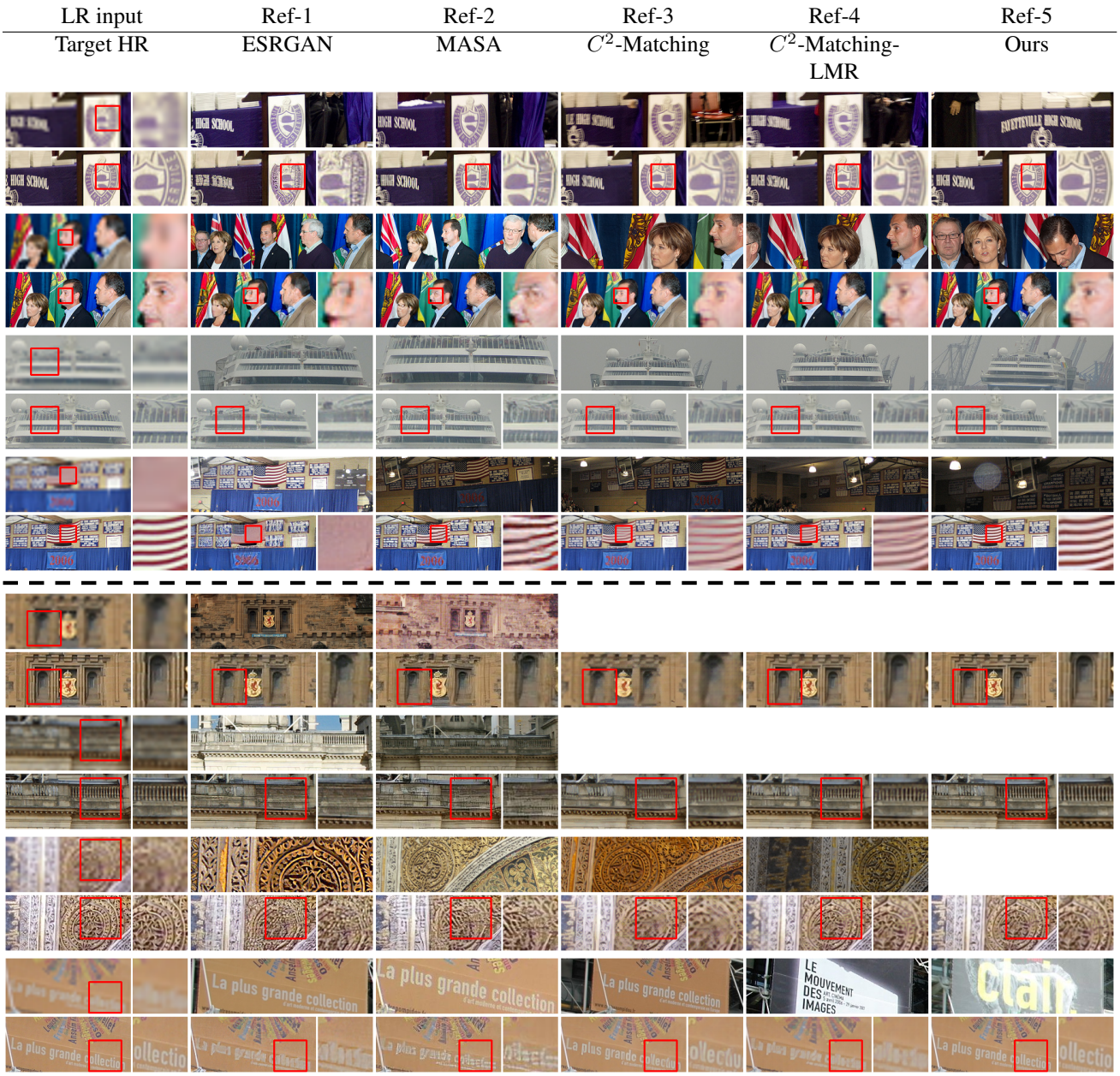

Figure 4. Qualitative comparisons on the testing set of CUFED5 (the top four examples) and LMR (the bottom four examples). We compare our results with ESRGAN, MASA, C^2-Matching, C^2-Matching-LMR. All these methods are trained with GAN loss. Our method reconstructs sharper details than other methods.

Qualitative evaluation. As shown in Figure 4, we compare the results of ESRGAN, MASA, C^2-Matching, C^2-Matching-LMR and our MRefSR. The top four examples are from CUFED5, and the models trained on LMR generalize well on the CUFED5 testing set, demonstrating the effectiveness of the proposed LMR. What's more, the results of our MRefSR trained with multiple references are much better than those trained with a single reference image. We also show four examples from the LMR testing set, and MRefSR can recover more texture details than other methods.

Besides, we perform a user study to compare with some typical methods including ESRGAN, MASA and C^2-Matching. Specifically, in each test, we present paired super-resolution results, one of which is generated by our MRefSR, and ask the users to choose the one with higher

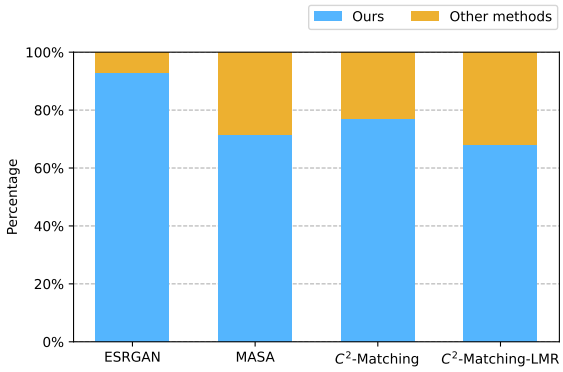

Figure 5. User study results. Values on Y-axis denote the voting percentage of users favoring our method.

Table 2. Ablation study on the influence of MAM, SAFM and LPF. All the methods are trained on the LMR dataset. The data results with $^\diamond$ means that they are tested with only one reference due to the limitation of GPU memory.

Method	LMR	CUFED5
	PSNR↑ / SSIM↑	PSNR↑ / SSIM↑
Baseline(C^2-Matching-LMR)	30.64 / 0.869 $^\diamond$	28.65 / 0.853
Baseline+MAM	31.70 / 0.894	28.85 / 0.859
Baseline+MAM+SAFM	31.81 / 0.895	28.94 / 0.860
Baseline+MAM+SAFM+LPF	31.98 / 0.898	29.05 / 0.862

visual quality. As shown in Figure 5, the users prefer our results over the others.

4.3. Ablation Study

In this section, we verify the effectiveness of Multi-reference Attention Module (MAM) and Spatial Aware Filtering Module (SAFM). Besides, we demonstrate the benefit of large-resolution training images of LMR. At last, we also investigate the impact of number of reference images.

The effectiveness of MAM and SAFM. As shown in Table 2, with C^2-Matching-LMR as the baseline, our MAM achieves a PSNR improvement of 1.06 dB on LMR and 0.20 dB on CUFED5. The reason why the improvement on LMR is larger than that on CUFED5 is that C^2-Matching-LMR cannot use multiple reference images on LMR due to the limitation of GPU memory, and MAM greatly solves this problem. More importantly, our MAM supports an arbitrary

Figure 6. Visual comparisons of ablation study on MAM and SAFM. With MAM and SAFM, the eyes on the face are more obvious and the lines of the windows are clearer.

Table 3. The effect of different number of reference images on CUFED5.

$\#Num$	C^2-Matching-LMR	Ours
$n=1$	28.474	28.663
$n=2$	28.615 (+0.141)	28.869 (+0.206)
$n=3$	28.651 (+0.036)	28.920 (+0.051)
$n=4$	28.649 (−0.002)	28.932 (+0.012)
$n=5$	28.650 (+0.001)	28.935 (+0.003)
Δ	+0.176	+0.272

number of reference images, making it more flexible and practical. On the basis of MAM, SAFM is used to adjust the fused reference features and the PSNR scores on LMR and CUFED5 increase to 31.81 dB and 28.94 dB, respectively. From Figure 6, we can see with MAM and SAFM, the output SR images have clearer textures. Furthermore, thanks to the larger image size of the LMR training data, MRefSR with large-patch (300×300) finetuning strategy (LPF) can consistently improve the performance by roughly 0.1 dB on LMR and CUFED5. This result reflects the advantage of the large training images of the LMR dataset.

The effect of the number of reference images. To study the influence of number of reference images, we conduct experiments on the testing set of CUFED5, in which each LR input image has five reference images. As shown in Table 3, as the number of reference images increases, although C^2-Matching-LMR has a slight improvement with the stitching testing strategy, the gap is still smaller than the improvement of MRefSR. What's more, when the number of reference images is greater than 3, the results are worse than the case of 3 reference images, which indicates that the stitching testing strategy neglects the interaction among references, so the information from the fourth reference doesn't explore with that from the first three reference effectively in this case. In contrast, it can be seen that with the increase of reference images, MRefSR has a stable positive gain. Last but no least, MRefSR with five references has a PSNR increase of 0.272 dB than that with one reference, whereas C^2-Matching-LMR only has a PSNR increase of 0.176 dB, which further demonstrates the superiority of modeling the relationship among multiple references.

4.4. Computational Cost

Here, we present the computational cost comparisons between the proposed MRefSR and previous single-reference RefSR methods, including MASA [21] and C^2-Matching [12]. The computational cost is computed on CUFED5 [42, 33] using one NVIDIA V100 GPU. In specific, for the single-reference RefSR methods on CUFED5, we stitch five reference images into a 2500×500 image as the reference image for testing. Certainly, our MRefSR can

Table 4. Computational cost and performance comparisons on the testing set of CUFED5. C^2-Matching-LMR means C^2-Matching-rec is trained on our LMR dataset.

Model	MASA-rec	C^2-Matching-rec	C^2-Matching-LMR	MRefSR-rec
GPU Memory (GB)	21.98	8.37	8.37	3.42
Runtime (s)	0.417	2.29	2.29	0.875
PSNR↑	27.54	28.40	28.65	28.94
SSIM↑	0.814	0.846	0.853	0.860

directly utilize all the reference images for testing. Table 4 reports the GPU memory, runtime and performance for each method. Our MRefSR consumes the least GPU memory and achieves the best performance with acceptable runtime.

5. Conclusion

In this paper, we propose a large-scale multi-reference RefSR dataset: LMR. Unlike CUFED5, the only training RefSR dataset available before, LMR has 5 reference images for each LR input image. What's more, LMR contains 112,142 groups of 300×300 training images, 10 times the number of CUFED5, and the image size is also larger. Besides, we propose a new multi-reference baseline RefSR method, named MRefSR. We use a multi-reference attention module (MAM) for feature fusion of an arbitrary number of reference images, and a spatial aware filtering module (SAFM) for the fused feature selection. With LMR enabling multi-reference RefSR training, our method effectively models the relationship among multiple references, thus achieving significant improvements over state-of-the-art approaches on both quantitative and qualitative evaluations. And our method solves the mismatch problem of previous methods using a single reference image for training but testing with multiple reference images.

Acknowledgements We thank Qing Chang, Jiawei He, and anonymous reviewers for helpful discussions. This work was supported in part by the Major Project for New Generation of AI (No.2018AAA0100400), the National Natural Science Foundation of China (No. 61836014, No. U21B2042, No. 62072457, No. 62006231).

References

[1] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 325–342, 2022. 2

[2] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 2

[3] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, pages 11065–11074, 2019. 1

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199, 2014. 1, 6

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE TPAMI*, 38(2):295–307, 2015. 1

[6] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *CVPR*, pages 8092–8101, 2019. 4

[7] Hayit Greenspan. Super-resolution in medical imaging. *The computer journal*, 52(1):43–63, 2009. 1

[8] Seamus J Holden, Stephan Uphoff, and Achillefs N Kapanidis. Daostorm: An algorithm for high-density super-resolution microscopy. *Nature methods*, 8(4):279–280, 2011. 1

[9] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, 2015. 3

[10] Yixuan Huang, Xiaoyun Zhang, Yu Fu, Siheng Chen, Ya Zhang, Yan-Feng Wang, and Dazhi He. Task decoupled framework for reference-based super-resolution. In *CVPR*, pages 5931–5940, 2022. 2, 6

[11] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *Graphical models and image processing*, 53(3):231–239, 1991. 1

[12] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *CVPR*, pages 2103–2112, 2021. 1, 2, 3, 4, 5, 6, 8

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016. 1

[14] Yongwoo Kim, Jae-Seok Choi, and Munchurl Kim. 2x super-resolution hardware using edge-orientation-based linear mapping for real-time 4k uhd 60 fps video applications. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(9):1274–1278, 2018. 1

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017. 1

[17] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. In *CVPR*, pages 17824–17833, 2022. 3

[18] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018. 3

[19] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single

image super-resolution. In *CVPRW*, pages 136–144, 2017. 1, 6

[20] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *NeurIPS*, 2018. 1

[21] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *CVPR*, pages 6368–6377, 2021. 1, 2, 6, 8

[22] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 3

[23] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *CVPR*, pages 3517–3526, 2021. 1

[24] Andrew J Patti, M Ibrahim Sezan, and A Murat Tekalp. Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time. *IEEE TIP*, 6(8):1064–1076, 1997. 1

[25] Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 3

[26] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, pages 501–518, 2016. 3

[27] Vida Fakour Sevom, Esin Guldogan, and Joni-Kristian Kämäräinen. 360 panorama super-resolution using deep convolutional networks. In *Int. Conf. Computer Vision Theory and Applications*, 2018. 1

[28] Gyumin Shim, Jinsun Park, and In So Kweon. Robust reference-based super-resolution with similarity-aware deformable convolution. In *CVPR*, pages 8425–8434, 2020. 1, 6

[29] Libin Sun and James Hays. Super-resolution from internet-scale scene matching. In *IEEE Int. Conf. Computational Photography*, pages 1–12, 2012. 1, 3, 5, 6

[30] Yang Tan, Haitian Zheng, Yinheng Zhu, Xiaoyun Yuan, Xing Lin, David Brady, and Lu Fang. Crossnet++: Cross-scale large-parallax warping for reference-based super-resolution. *IEEE TPAMI*, 43(12):4291–4305, 2020. 1

[31] Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, and Qifeng Chen. Dual-camera super-resolution with aligned attention modules. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[32] 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 *ECCVW*, 2018. 1, 6

[33] Yufei Wang, Zhe Lin, Xiaohui Shen, Radomir Mech, Gavin Miller, and Garrison W Cottrell. Event-specific image importance. In *CVPR*, pages 4810–4819, 2016. 2, 3, 5, 6, 8

[34] Yuzhuo Wei, Li Chen, Rong Xie, Li Song, Xiaoyun Zhang, and Zhiyong Gao. Fpga based video transcoding system with 2k-4k super-resolution conversion. In *IEEE Visual Communications and Image Processing*, pages 1–2, 2019. 1

[35] Bin Xia, Yapeng Tian, Yucheng Hang, Wenming Yang, Qingmin Liao, and Jie Zhou. Coarse-to-fine embedded patchmatch and multi-scale dynamic aggregation for reference-based super-resolution. In *AAAI*, 2022. 1, 2, 6

[36] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, pages 5791–5800, 2020. 2, 4, 6

[37] Huanjing Yue, Xiaoyan Sun, Jingyu Yang, and Feng Wu. Landmark image super-resolution by retrieving web images. *IEEE TIP*, 22(12):4865–4878, 2013. 1, 6

[38] Kaipeng Zhang, Zhanpeng Zhang, Chia-Wen Cheng, Winston H Hsu, Yu Qiao, Wei Liu, and Tong Zhang. Super-identity convolutional neural network for face hallucination. In *ECCV*, pages 183–198, 2018. 1

[39] Lin Zhang, Xin Li, Dongliang He, Fu Li, Yili Wang, and Zhaoxiang Zhang. Rrsr: Reciprocal reference-based image super-resolution with progressive feature alignment and selection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 648–664, 2022. 2

[40] Liangpei Zhang, Hongyan Zhang, Huanfeng Shen, and Pingxiang Li. A super-resolution reconstruction algorithm for surveillance images. *Signal Processing*, 90(3):848–859, 2010. 1

[41] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018. 1, 6

[42] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *CVPR*, pages 7982–7991, 2019. 1, 2, 3, 5, 6, 8

[43] 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 *ECCV*, pages 88–104, 2018. 1, 6

[44] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019. 2