Reference-Based Image Super-Resolution with Deformable Attention Transformer

Jiezhang Cao[1], Jingyun Liang[1], Kai Zhang[1], Yawei Li[1], Yulun Zhang[1(✉)], Wenguan Wang[1], and Luc Van Gool[1,2]

[1] Computer Vision Lab, ETH Zürich, Zürich, Switzerland
{jiezhang.cao,jingyun.liang,kai.zhang,yawei.li,yulun.zhang,
wenguan.wang,vangool}@vision.ee.ethz.ch
[2] KU Leuven, Leuven, Belgium
https://github.com/caojiezhang/DATSR

Abstract. Reference-based image super-resolution (RefSR) aims to exploit auxiliary reference (Ref) images to super-resolve low-resolution (LR) images. Recently, RefSR has been attracting great attention as it provides an alternative way to surpass single image SR. However, addressing the RefSR problem has two critical challenges: (i) It is difficult to match the correspondence between LR and Ref images when they are significantly different; (ii) How to transfer the relevant texture from Ref images to compensate the details for LR images is very challenging. To address these issues of RefSR, this paper proposes a deformable attention Transformer, namely DATSR, with multiple scales, each of which consists of a texture feature encoder (TFE) module, a reference-based deformable attention (RDA) module and a residual feature aggregation (RFA) module. Specifically, TFE first extracts image transformation (*e.g.*, brightness) insensitive features for LR and Ref images, RDA then can exploit multiple relevant textures to compensate more information for LR features, and RFA lastly aggregates LR features and relevant textures to get a more visually pleasant result. Extensive experiments demonstrate that our DATSR achieves state-of-the-art performance on benchmark datasets quantitatively and qualitatively.

Keywords: Reference-based image super-resolution · Correspondence matching · Texture transfer · Deformable attention transformer

1 Introduction

Single image super-resolution (SISR), which aims at recovering a high-resolution (HR) image from a low-resolution (LR) input, is an active research topic due to its high practical values [9,13–16,18,20,21,41,46,49,51]. However, SISR is a highly ill-posed problem since there exist multiple HR images that can degrade to

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-19797-0_19.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13678, pp. 325–342, 2022.
https://doi.org/10.1007/978-3-031-19797-0_19

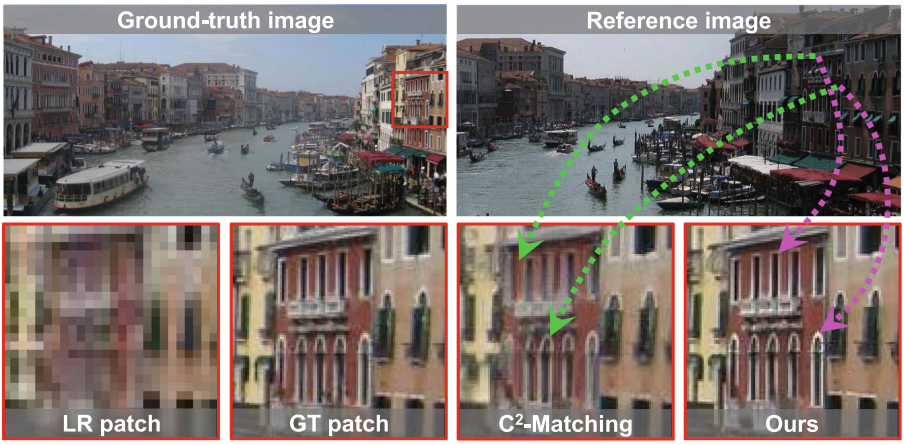


Fig. 1. Comparison with the state-of-the-art RefSR method C^2-Matching [12]. When the brightness of LR and Ref image is different, our method performs better than C^2-Matching [12] in transferring relevant textures from the Ref image to the SR image, which is closer to the ground-truth image.

the same LR image [8,38]. While real LR images usually have no corresponding HR ground-truth (GT) images, one can easily find a high-quality image as a reference (Ref) image with high-frequency details from various sources, such as photo albums, video frames, and web image search, which has similar semantic information (such as content and texture) to the LR image. Such an alternative SISR method is referred to as reference-based super-resolution (RefSR), which aims to transfer HR textures from the Ref images to super-resolved images and has shown promising results over SISR. Although various RefSR methods [12, 27,45,47] have been recently proposed, two challenges remain unsolved for SR performance improvement.

First, it is difficult to match the correspondence between the LR and Ref images especially when their distributions are different. For example, the brightness of the Ref images is different from that of the LR images. Existing methods [48,56] mostly match the correspondence by estimating the pixel or patch similarity of texture features between LR and Ref images. However, such similarity metric is sensitive to image transformations, such as brightness and color of images. Recently, the state-of-the-art (SOTA) method C^2-Matching [12] trains a feature extractor, which demonstrates strong robustness to scale and rotation. However, it neglects to explore the effects of brightness, contrast, and color of images. As a result, this method may transfer inaccurate textures from the Ref image, when the Ref images have different brightness from the LR image, as shown in Fig. 1. Based on the observation and analyses, we can see that the quality of correspondence is affected by the similarity metric and the distribution gap between the LR and Ref images.

On the other hand, some methods [34,57] adopt optical flow or deformable convolutions [3,4,42,59] to align spatial features between the Ref and LR images. However, these methods may find an inaccurate correspondence when the

distance between the LR and Ref images is relatively large. With the inaccurate correspondence, their performance would deteriorate seriously since the irrelevant texture cannot provide meaningful details. Therefore, how to accurately match the correspondence between the Ref and LR images is a challenging problem as it affects the quality of super-resolved results.

Second, it is also challenging to transfer textures of the high-quality Ref images to restore the HR images. One representative work CrossNet [57] estimates the flow from the Ref image to the LR image and then warp the features based on the optical flow. However, the optical flow may be inaccurate, since the Ref and LR images could be significantly different. In addition, most existing methods [27,48,56] search the most similar textures and the corresponding position, and then swap the texture features from the Ref image. As a result, these methods may transfer irrelevant textures to the output and have poor SR performance, when the original estimated flow or position is inaccurate. Hence, it is important and necessary to explore a new architecture to adaptively transfer texture and mitigate the impact of inaccurate correspondence in the Ref image.

To address the above two challenges, we propose a novel deformable attention Transformer, namely DATSR, for reference-based image super-resolution. DATSR is built on the U-Net and consists of three basic modules, including texture feature encoders, deformable attention, and residual feature aggregation. Specifically, we first use texture feature encoders to extract multi-scale features with different image transformations. Then, we propose a reference-based deformable attention to discover the multiple relevant correspondences and adaptively transfer the textures. Last, we fuse features and reconstruct the SR images using residual feature aggregation. We conduct extensive comparisons with recent representative SOTA methods on benchmark datasets. The quantitative and visual results demonstrate that our DATSR achieves the SOTA performance.

The main contributions are summarized as follows:

- We propose a novel reference-based image super-resolution with deformable attention transformer (DATSR), which is end-to-end trainable by incorporating Transformer into RefSR. Compared with existing RefSR methods, our DATSR performs more robust correspondence matching and texture transfer and subsequently achieves SOTA performance quantitatively and visually.
- We design a new reference-based deformable attention module for correspondence matching and texture transfer. Different from existing transformer-based methods, our transformer is built on U-Net with multi-scale features and alleviates the resolution gap between Ref and LR images. Moreover, our transformer relieves the correspondence mismatching issue and the impact of distribution gap between LR and Ref images.
- We conduct extensive experiments on benchmark datasets to demonstrate that our DATSR achieves SOTA performance and is also robust to different image transformations (*e.g.*, brightness, contrast and hue). Moreover, we find that our DATSR trained with a single Ref image outperforms existing Multi-RefSR methods trained with multiple Ref images. In addition, our DATSR still shows good performance even in some extreme cases, when the Ref images have no texture information.

2 Related Work

We will briefly introduce two related super-resolution paradigms, including single image super-resolution and reference-based image super-resolution.

Single Image Super-Resolution (SISR). The goal of SISR is to recover high-resolution (HR) images from the low-resolution (LR) images. Recent years have witnessed significant achievements of using deep neural networks to solve SISR [6,55]. SRCNN [6] is the pioneer work of exploiting deep convolutional networks to map LR image into HR image. To further improve SR performance, researchers resort to employing deeper neural networks with attention mechanisms and residual blocks [5,19–23,31,33,36,50,54,55]. However, it is difficult for traditional SISR methods to produce realistic images when the HR textures are highly degraded. To relieve this, some SR methods [11,17,40,43,44,53,58] adopt generative adversarial networks (GANs) to further improve the perceptual quality of the super-resolved outputs.

Reference-Based Image Super-Resolution (RefSR). Different from SISR, RefSR has auxiliary HR images and aims to super-resolves images by transferring HR details of Ref images. Such auxiliary information can be extracted from the reference images which are similar to HR ground-truth images. CrossNet [57] estimates the optical flow (OF) between Ref and LR images and then performs the cross-scale warping and concatenation. Instead of estimating OF, SRNTT [56] calculates the similarity between the LR and Ref images and transfer the texture from the Ref images. Similarly, SSEN [34] proposes a similarity search and extraction network and it is aware of the best matching position and the relevancy of the best match. To improve the performance, TTSR [48] proposes a hard and soft attention for texture transfer and synthesis. Instead of using the features of a classifier, E2ENT^2 [45] transfers texture features by using a SR task-specific features. To improve the efficiency of matching, MASA [27] proposes a coarse-to-fine correspondence matching module and a spatial adaptation module to map the distribution of the Ref features to that of the LR features. Recently, a strong RefSR method C^2-Matching [12] first proposes a contrastive correspondence network to learn correspondence, and then adopts a teacher-student correlation distillation to improve LR-HR matching, and last uses a residual feature aggregation to synthesize HR images.

It should be noted that RefSR can be extended to the case of multiple reference images, called **Multi-RefSR**, which aims to transfer the texture features from multiple Ref images to the SR image. Recently, a content independent multi-reference super-resolution model CIMR-SR [47] is proposed to transfer the HR textures from multiple reference images. To improve the performance, AMRSR [32] proposes an attention-based multi-reference super-resolution network to match the most similar textures from multiple reference images. Different from RefSR, Multi-RefSR can exploit more training information as it has multiple Ref images. In this paper, we mainly study RefSR and train the model with single Ref image. Nevertheless, we still compare our model with the above Multi-RefSR methods to further demonstrate the effectiveness of our DATSR.

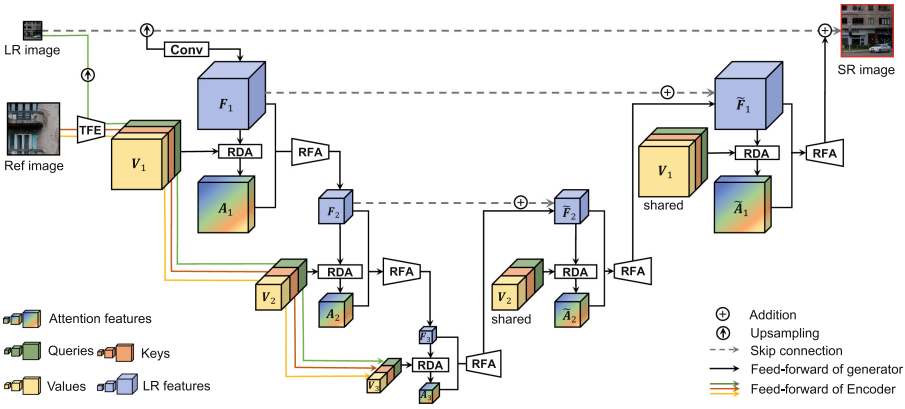


Fig. 2. The architecture of our DATSR network. At each scale, our model consists of texture feature encoders (TFE), a reference-based deformable attention (RDA) module and a residual feature aggregation module (RFA).

3 Proposed Method

Due to the the intrinsic complexity of RefSR, we divide the problem into two main sub-tasks: correspondence matching and texture transfer. To address these, we propose a multi-scale reference-based image SR with deformable Transformer, as shown in Fig. 2. Specifically, we first use TFE to extract multi-scale texture features of Ref and LR images, then propose RDA to match the correspondences and transfer the textures from Ref images to LR images, and last use RFA to aggregate features and generate SR images.

3.1 Texture Feature Encoders

In the RefSR task, it is important to discover robust correspondence between LR and Ref images. However, there are some underlying gaps between LR and Ref images, *i.e.*, the resolution gap and the distribution gap (*e.g.*, brightness, contrast and hue). To address this, we propose texture feature encoders to extract robust features of LR and Ref images. For the resolution gap, we propose to use pre-upsampling in the LR image and extract multi-scale features of LR and Ref images. Specifically, given an LR image $\boldsymbol{X}_{LR}$ and a reference image $\boldsymbol{X}_{Ref}$, we upsample the LR image to the resolution of the Ref image, denoted as $\boldsymbol{X}_{LR\uparrow}$. Then, we calculate multi-scale features of the LR and Ref images, *i.e.*,

$$\boldsymbol{Q}_l = E_l^q(\boldsymbol{X}_{LR\uparrow}), \quad \boldsymbol{K}_l = E_l^k(\boldsymbol{X}_{Ref}), \quad \boldsymbol{V}_l = E_l^v(\boldsymbol{X}_{Ref}), \tag{1}$$

where E_l^q, E_l^k and E_l^v are feature encoders at the l-th scale. In our architecture, we use three scales in the texture feature encoders. With the help of the multi-scale features in U-Net, we are able to alleviate the resolution gap between the Ref and LR images since they contain the complementary scale information.

For the distribution gap, we augment images with different image transformations (*e.g.*, brightness, contrast and hue) in the training to improve the robustness of our model. In addition to data augmentation, we use contrastive learning to train the encoder be less sensitive to different image transformations, inspired by [12]. To estimate the stable correspondence between $\boldsymbol{X}_{LR\uparrow}$ and $\boldsymbol{X}_{Ref}$, the feature encoders E_l^q and E_l^k are the same, and the feature encoder E_l^r is pre-trained and fixed in the training. In contrast, TTSR [48] directly uses a learnable feature encoder, resulting in limited performance since the textures are changing during training and the correspondence matching is unstable. For C^2-Matching [12], it neglects to improve the robustness to brightness, contrast and hue. To address these, we propose to learn robust multi-scale features $\boldsymbol{Q}_l, \boldsymbol{K}_l, \boldsymbol{V}_l$, which can be regraded as Query, Key, and Value, and can be used in our attention mechanism conditioned on the LR features.

3.2 Reference-Based Deformable Attention

Existing attention-based RefSR methods (*e.g.*, [48]) tend to suffer from limited performance when the most relevant features between LR and Ref images are inaccurate, *i.e.*, the learned LR features may not well match the Ref features. To address this, we propose a new reference-based attention mechanism, called RefAttention, as shown in Fig. 3. Formally, given Query $\boldsymbol{Q}_l$, Key $\boldsymbol{K}_l$, Value $\boldsymbol{V}_l$, and LR features $\boldsymbol{F}_l$, the attention feature $\boldsymbol{A}_l$ is defined as follows:

$$\boldsymbol{A}_l = \text{RefAttention}(\boldsymbol{Q}_l, \boldsymbol{K}_l, \boldsymbol{V}_l, \boldsymbol{F}_l) = \mathcal{T}\left(\sigma\left(\boldsymbol{Q}_l^\top \boldsymbol{K}_l\right), \boldsymbol{V}_l, \boldsymbol{F}_l\right). \tag{2}$$

Different from existing attention mechanism [39], our attention is conditioned on the LR features and designed for the RefSR task. In Fig. 3, we denoted by $\boldsymbol{A}_l$ and $\boldsymbol{F}_l$ in the downscaling process, and $\tilde{\boldsymbol{A}}_l$ and $\tilde{\boldsymbol{F}}_l$ in the upscaling process. $\sigma(\cdot)$ is a correspondence matching function to calculate the relevance between the Ref and LR images. Based on the relevance, we propose a texture transfer function $\mathcal{T}(\cdot)$ to transfer the textures from the Ref to the LR image.

Correspondence Matching. The first important sub-task in RefSR is to match correspondences between LR and Ref images. Most existing methods [48,56] are sensitive to different image transformations (*e.g.*, brightness, contrast and hue) and may match inaccurate correspondences. To relieve this issue, we propose a correspondence matching module in our RefAttention, as shown in Fig. 3. Specifically, we estimate the relevance between $\boldsymbol{X}_{LR\uparrow}$ and $\boldsymbol{X}_{Ref}$ by calculating similarity between $\boldsymbol{Q}_l \in \mathbb{R}^{C\times H_1\times W_1}$ and $\boldsymbol{K}_l \in \mathbb{R}^{C\times H_2\times W_2}$. First, we unfold $\boldsymbol{Q}_l$ and $\boldsymbol{K}_l$ into patches $\boldsymbol{Q'}_l = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{H_1W_1}] \in \mathbb{R}^{C\times H_1W_1}$ and $\boldsymbol{K'}_l = [\boldsymbol{k}_1, \ldots, \boldsymbol{k}_{H_2W_2}] \in \mathbb{R}^{C\times H_2W_2}$. Then, for the given query $\boldsymbol{q}_i$ in $\boldsymbol{Q'}$, the top K relevant positions in $\boldsymbol{K'}$ can be calculated by normalized inner product,

$$\boldsymbol{P}_i = \left[\sigma\left(\boldsymbol{Q'}_l^\top \boldsymbol{K'}_l\right)\right]_i = \text{TopK}_j\left(\widetilde{\boldsymbol{q}}_i \cdot \widetilde{\boldsymbol{k}}_j\right), \tag{3}$$

where $\widetilde{\boldsymbol{q}}_i = \boldsymbol{q}_i/\|\boldsymbol{q}_i\|$ and $\widetilde{\boldsymbol{k}}_j = \boldsymbol{k}_j/\|\boldsymbol{k}_j\|$ are normalized features, and $\text{TopK}(\cdot)$ is a function and returns top K relevant positions $\boldsymbol{P}_i = \left\{p_i^1, \ldots, p_i^K\right\}$. Here, $\boldsymbol{P}_i$ is

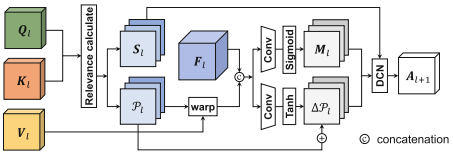


Fig. 3. The architecture of RDA.

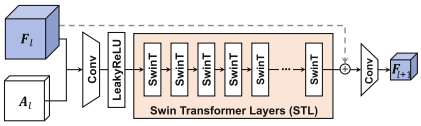


Fig. 4. The architecture of RFA.

the i-th element of $\mathcal{P}_l$, and the position p_i^1 is the most relevant position in the Ref image to the i-th position in LR. When $K > 1$, it helps discover multiple correspondences, motivated by KNN [24]. For fair comparisons with other RefSR methods, we set $K = 1$ and exploit the most relevant position in the experiments.

Similarity-Aware Texture Transfer. The second important sub-task in RefSR is to transfer textures from Ref images to LR images based on the matched correspondence. Most existing RefSR methods [48,56] directly swap the most relevant texture from Ref image. However, it may degrade the performance when the most relevant texture is inaccurate. To address this, we propose to improve the deformable convolution (DCN) [4,59] to transfer the texture around every position p_i^k of Ref images. Specifically, let Δp_i^k be the spatial difference between the position p_i and the k-th relevant position p_i^k, *i.e.*, $\Delta p_i^k = p_i^k - p_i$. Then, we calculate a feature at the position p using modified DCN, *i.e.*,

$$\boldsymbol{A}_l(p_i) = \sum\nolimits_{k=1}^{K} s_i^k \sum\nolimits_j w_j \boldsymbol{V}_l(p_i + \Delta p_i^k + p_j + \Delta p_j)\, m_j, \tag{4}$$

where $p_j \in \{(-1,1),(-1,0),\ldots,(1,1)\}$, s_i^k is the cooperative weight to aggregate the K textures from the Ref image, *i.e.*, $s_i^k = \exp(\widetilde{\boldsymbol{q}}_i \cdot \widetilde{\boldsymbol{k}}_{p_i^k}) / \sum_{j \in \boldsymbol{P}_i} \exp(\widetilde{\boldsymbol{q}}_i \cdot \widetilde{\boldsymbol{k}}_j)$, w_j is the convolution kernel weight, Δp_j is the j-th learnable offset of $\Delta\mathcal{P}_l$, and m_j is the j-th learnable mask of $\boldsymbol{M}_l$, which can be calculated as follows,

$$\begin{cases} \Delta\mathcal{P}_l = r \cdot \mathrm{Tanh}\left(\mathrm{Conv}([\boldsymbol{F}_l;\ \omega(\boldsymbol{V}_l, \mathcal{P}_l)])\right), \\ \boldsymbol{M}_l = \mathrm{Sigmoid}\left(\mathrm{Conv}([\boldsymbol{F}_l;\ \omega(\boldsymbol{V}_l, \mathcal{P}_l)])\right), \end{cases} \tag{5}$$

where ω is a warping function, $[;]$ is a concatenation operation, Conv is convolutional layers. Sigmoid and Tanh are activation functions, r is the max magnitude which is set as 10 in default, and $\boldsymbol{F}_l$ is the feature of upsampled LR images at the l-th scale. With the help of the mask, we can adaptively transfer textures even if LR and Ref images are significantly different. When the Ref image has irrelevant texture or no information, our model is able to guild whether to transfer the textures in Ref images. In this sense, it can relieve the correspondence mismatching issue. In this paper, we mainly compare with RefSR methods with single Ref image. Thus, we transfer one relevant textures from the Ref image for fair comparison. With the help of our architecture, the proposed RDA module is able to improve the RefSR performance by transferring textures at each scale in both downscaling and upscaling, which is different from C^2-Matching [12].

3.3 Residual Feature Aggregation

To aggregate the multi-scale LR features at different layers and the transferred texture features, we propose a residual feature aggregation module (RFA) to perform feature fusion and extraction. As shown in Fig. 4, RFA consists of CNNs and Swin Transformer layers (STL) [25] which gain much attention in many tasks [2,19,26]. Specifically, we first use a convolution layer to fuse the LR feature $\boldsymbol{F}_l$ and attention features $\boldsymbol{A}_l$, *i.e.*, $\boldsymbol{F}'_{l+1} = \text{Conv}(\boldsymbol{F}_l, \boldsymbol{A}_l)$, where Conv is convolutional layers. Then, we use Swin Transformer and a residual connection to extract deeper features of the LR and transferred features,

$$\boldsymbol{F}'_{l+1} = \text{STL}(\boldsymbol{F}'_{l+1}) + \boldsymbol{F}_l, \tag{6}$$

where the details of STL are put in the supplementary materials. At the end of RFA, we use another convolutional layer to extract the features of STL, $\boldsymbol{F}_{l+1} = \text{Conv}(\boldsymbol{F}'_{l+1})$. Based on the aggregated features $\boldsymbol{F}_L$ at the last scale, we synthesize SR images with a skip connection as

$$\boldsymbol{X}_{SR} = \boldsymbol{F}_L + \boldsymbol{X}_{LR\uparrow}. \tag{7}$$

3.4 Loss Function

In the training, we aim to i) preserve the spatial structure and semantic information of LR images; ii) discover more texture information of Ref images; iii) synthesize realistic SR images with high quality. To this end, we use a reconstruction loss, a perceptual loss and an adversarial loss, which is the same as [12,48]. The overall loss with the hype-parameters λ_1 and λ_2 is written as:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{per} + \lambda_2 \mathcal{L}_{adv}. \tag{8}$$

Reconstruction Loss. In order to make the SR image $\boldsymbol{X}_{SR}$ to be close to the HR ground-truth image $\boldsymbol{X}_{HR}$, we adopt the following reconstruction loss

$$\mathcal{L}_{rec} = \|\boldsymbol{X}_{HR} - \boldsymbol{X}_{SR}\|_1, \tag{9}$$

where $\|\cdot\|_1$ is the ℓ_1-norm.

Perceptual Loss. To enhance the visual quality of SR images, the perceptual loss is widely used in SR models [12,56]. The perceptual loss is defined as:

$$\mathcal{L}_{per} = \frac{1}{V} \sum\nolimits_{i=1}^{C} \|\phi_i(\boldsymbol{X}_{HR}) - \phi_i(\boldsymbol{X}_{SR})\|_F, \tag{10}$$

where $\|\cdot\|_F$ is the Frobenius norm, and V and C are the volume and channel number of the feature maps, respectively. The function ϕ_i is the i-th intermediate layer in VGG19 [35], and we use the relu5_1 layer of VGG19 in the experiment.

Adversarial Loss. To improve the visual quality of SR images, many SR methods [17,44] introduce GANs [1,7] which have achieved good performance for SR. Specifically, we use WGAN [1] loss as follows,

$$\mathcal{L}_{adv} = \mathbb{E}_{\boldsymbol{X}_{SR} \sim \mathbb{P}_{SR}}[D(\boldsymbol{X}_{SR})] - \mathbb{E}_{\boldsymbol{X}_{HR} \sim \mathbb{P}_{HR}}[D(\boldsymbol{X}_{HR})], \tag{11}$$

where $D(\cdot)$ is a discriminator, $\mathbb{P}_{SR}$ is the distribution of the generated SR images, and $\mathbb{P}_{HR}$ is the distribution of the real data.

4 Experiments

Datasets. In the experiment, we consider the RefSR dataset, *i.e.*, CUFED5 [56], which consists of a training set and a testing set. The CUFED5 training set contains 11,871 training pairs, and each pair has an original HR image and a corresponding Ref image at the size of 160 × 160. The CUFED5 testing set has 126 input images and each image has 4 reference images with different similarity levels. For fair comparisons, all models are trained on the training set of CUFED5. To evaluate the generalization ability, we test our model on the CUFED5 testing set, Urban100 [10], Manga109 [30], Sun80 [37] and WR-SR [12]. The Sun80 and WR-SR datasets contain 80 natural images, and each paired with one or more reference images. For the Urban100 dataset, we concatenate the LR and random sampled HR images as the reference images. For the Manga109 dataset, we randomly sample HR images as the reference images since there are no the reference images. All experiments are conducted for 4× SR.

Evaluation Metrics. Existing RefSR methods [12,48,48] mainly use PSNR and SSIM to compare the performance. Here, PSNR and SSIM are calculated on the Y channel of YCbCr color space. In general, larger PSNR and SSIM correspond to better performance of the RefSR method. In addition, we compare the model size (*i.e.*, the number of trainable parameters) of different models.

Implementation Details. The input LR images are generated by bicubicly downsampling the HR images with scale factor 4. For the encoders and discriminator, we adopt the same architectures as [12]. We use a pre-trained relu1_1, relu2_1 and relu3_1 of VGG19 to extract multi-scale features. we augment the training data with randomly horizontal and vertical flipping or different random rotations of 90°, 180° and 270°. Besides, we also augment the training data by randomly changing different brightness, contrast and hue of an image by using ColorJitter in pytorch. In the training, we set the batch size as 9, *i.e.*, each batch has 9 LR, HR and Ref patches. The size of LR images is 40 × 40, and the size of HR and Ref images is 160 × 160. Following the training of [12], we set the hype-parameters λ_1 and λ_2 as 1×10^{-4} and 1×10^{-6}, respectively. We set the learning rate of the SR model and discriminator as 1×10^{-4}. For the Adam optimizer, we set $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We provide more detailed network architectures and training details in the supplementary material.

4.1 Comparison with State-of-the-Art Methods

We compare with the SISR methods (SRCNN [6], EDSR [22], RCAN [55], SwinIR [19], SRGAN [17], ENet [33], ESRGAN [44], and RankSR-GAN [53]) and RefSR methods (CrossNet [57], SRNTT [56], SSEN [34], TTSR [48], E2ENT2 [45], and MASA [27]). For fair comparisons, the above models are trained on CUFED5 training set, and tested on CUFED5 testing set, Urban100, Manga109, Sun80 and WR-SR. In this experiment, we train our model on two cases only with reconstruction loss (denoted as '-rec'), and with all loss functions.

Table 1. Quantitative comparisons (PSNR and SSIM) of SR models trained with only reconstruction loss (with the suffix '-rec'). We group methods by SISR and RefSR. We mark the best results **in bold**.

SR paradigms	Methods	CUFED5 [56]		Urban100 [10]		Manga109 [30]		Sun80 [37]		WR-SR [12]	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SISR	SRCNN [6]	25.33	0.745	24.41	0.738	27.12	0.850	28.26	0.781	27.27	0.767
	EDSR [22]	25.93	0.777	25.51	0.783	28.93	0.891	28.52	0.792	28.07	0.793
	ENet [33]	24.24	0.695	23.63	0.711	25.25	0.802	26.24	0.702	25.47	0.699
	RCAN [55]	26.06	0.769	25.42	0.768	29.38	0.895	29.86	0.810	28.25	0.799
	SwinIR [19]	26.62	0.790	26.26	0.797	30.05	0.910	30.11	0.817	28.06	0.797
RefSR	CrossNet [57]	25.48	0.764	25.11	0.764	23.36	0.741	28.52	0.793	–	–
	SRNTT-rec [56]	26.24	0.784	25.50	0.783	28.95	0.885	28.54	0.793	27.59	0.780
	TTSR-rec [48]	27.09	0.804	25.87	0.784	30.09	0.907	30.02	0.814	27.97	0.792
	SSEN-rec [34]	26.78	0.791	–	–	–	–	–	–	–	–
	E2ENT^2-rec [45]	24.24	0.724	–	–	–	–	28.50	0.789	–	–
	MASA-rec [27]	27.54	0.814	26.09	0.786	30.24	0.909	30.15	0.815	28.19	0.796
	C^2-Matching-rec [12]	28.24	0.841	26.03	0.785	30.47	0.911	30.18	0.817	28.32	0.801
	DATSR-rec (Ours)	**28.72**	**0.856**	**26.52**	**0.798**	**30.49**	**0.912**	**30.20**	**0.818**	**28.34**	**0.805**

Table 2. Quantitative comparisons (PSNR and SSIM) of SR models trained with all losses. We mark the best results **in bold**.

SR paradigms	Methods	CUFED5 [56]		Urban100 [10]		Manga109 [30]		Sun80 [37]		WR-SR [12]	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SISR	SRGAN [17]	24.40	0.702	24.07	0.729	25.12	0.802	26.76	0.725	26.21	0.728
	ESRGAN [44]	21.90	0.633	20.91	0.620	23.53	0.797	24.18	0.651	26.07	0.726
	RankSRGAN [53]	22.31	0.635	21.47	0.624	25.04	0.803	25.60	0.667	26.15	0.719
RefSR	SRNTT [56]	25.61	0.764	25.09	0.774	27.54	0.862	27.59	0.756	26.53	0.745
	TTSR [48]	25.53	0.765	24.62	0.747	28.70	0.886	28.59	0.774	26.83	0.762
	SSEN [34]	25.35	0.742	–	–	–	–	–	–	–	–
	E2ENT^2 [45]	24.01	0.705	–	–	–	–	28.13	0.765	–	–
	MASA [27]	24.92	0.729	23.78	0.712	27.26	0.847	27.12	0.708	25.74	0.717
	C^2-Matching [12]	27.16	0.805	25.52	0.764	29.73	0.893	29.75	0.799	27.80	0.780
	DATSR (Ours)	**27.95**	**0.835**	**25.92**	**0.775**	**29.75**	**0.893**	**29.77**	**0.800**	**27.87**	**0.787**

Quantitative Comparison. We provide quantitative comparisons of SR models trained with only reconstruction loss and all losses in Tables 1 and 2, respectively. In Table 1, our model has the best PSNR and SSIM on all testing sets and significantly outperforms all SISR and RefSR models. It implies that our Transformer achieves the state-of-the-arts and good generalization performance. For the SISR setting, our method performs better than the state-of-the-art SISR method [19]. It is difficult for these SISR methods to synthesize since the high-frequency information is degraded. In contrast, our model is able to adaptively discover the useful information from a reference image on the Urban100 and Manga109 datasets even if it is a random image. For the RefSR setting, our proposed DATSR significantly outperforms all methods with the help of the cooperative transfer with deformable convolution module.

In Table 2, our DATSR also achieves the much higher PSNR/SSIM values than other RefSR methods with a large margin. Our DATSR trained with adversarial loss reduces PSNR and SSIM but increases the visual quality. Still, it has

the best performance over all compared methods. The above quantitative comparison results on different SR paradigms demonstrate the superiority of our Transformer over state-of-the-art SISR and RefSR methods.

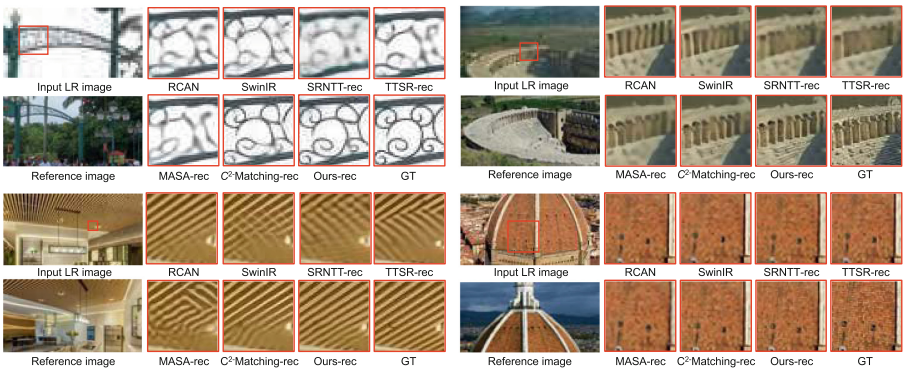


Fig. 5. Qualitative comparisons of SISR and RefSR models trained with the reconstruction loss.

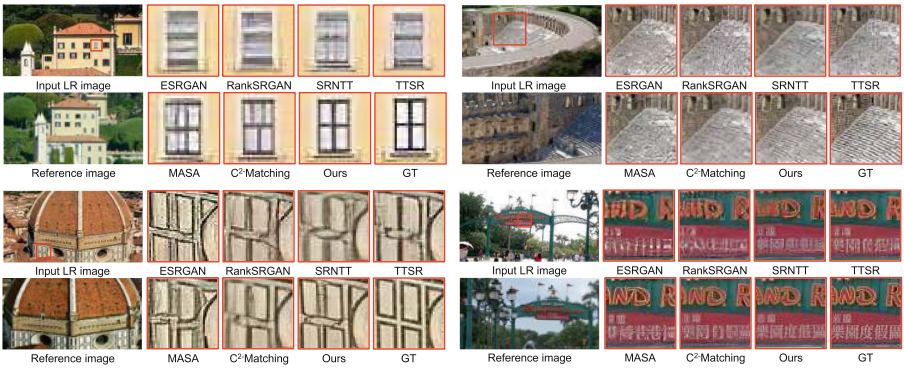


Fig. 6. Qualitative comparisons of SISR and RefSR models trained with all loss.

Qualitative Comparison. The visual results of our method are shown in Figs. 5 and 6. In these figures, our model also achieves the best performance on visual quality when trained with the reconstruction loss and all loss. These results demonstrate that our proposed method is able to transfer more accurate textures from the Ref images to generate SR images with higher quality. When trained with the reconstruction loss, our medel can synthesize SR images with sharp structure. Moreover, our method is able to search and transfer meaningful texture in a local regions even if the Ref image is not globally relevant to the input image. When trained with the adversarial loss, our model is able to restore the realistic details in the output images which are very close to the

HR ground-truth images with the help of the given Ref images. In contrast, it is hard for ESRGAN and RankSRGAN to generate realistic images without the Ref images since the degradation is severely destroyed and high frequency details of images are lost. For RefSR methods, our model is able to synthesize more realistic texture from the Ref images than SRNTT [56], TTSR [48], MASA [27], and C^2-Matching [12]. For example, in the top of Fig. 6, our model is able to recover the "window" with sharper edge and higher quality than C^2-Matching, but other methods fail to restore it even if they have a Ref image.

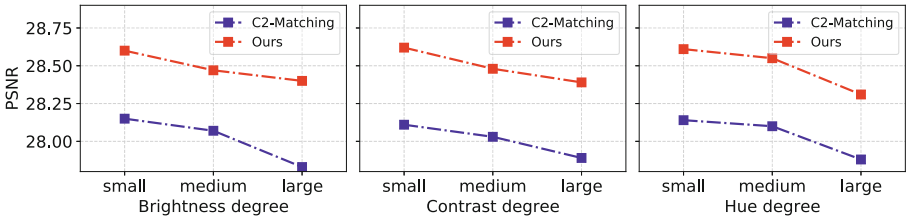
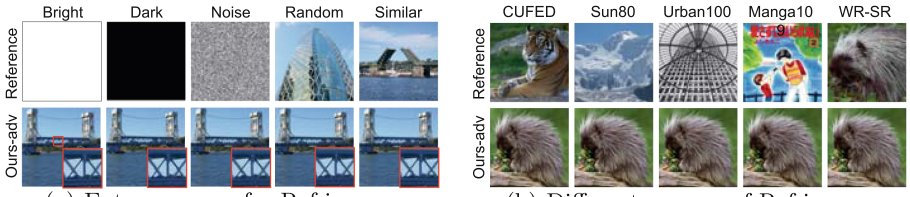


Fig. 7. Robustness to different image transformations. Our DATSR is more robust than C^2-Matching [12] under different image transformations.



(a) Extreme cases for Ref images. (b) Different sources of Ref images.

Fig. 8. Investigation on different types of reference images.

4.2 Further Analyses

Robustness to Image Transformations. We analyze the robustness of our model to different kinds of image transformations. Specifically, we use ColorJitter to augment the CUFED5 testing set by randomly change the brightness, contrast and hue of Ref images into three group: small, medium and large. The detailed settings are put in the supplementary materials. In Fig. 7, our model is more robust than C^2-Matching [12] under different image transformations. Note that the medium and large transformations are not included during training but our model still has superior performance.

Effect on Type and Number of Ref Images. We test our model on different Ref images, such as extreme images (*i.e.*, may have only one color or noise without any information) and random images from different testing sets. In Fig. 8, our method has robust performance and high visual quality even if the

Ref images have no useful texture information. In addition, our model has better performance when increasing #Ref images in Fig. 9. Table 3 shows the results of four similarity levels ("L1" to "L4") where L1 is the most relevant level. Our method achieves the best performance across all similarity levels.

Comparisons with Multi-RefSR Methods. We compare our model with multi-RefSR methods, *i.e.*, CIMR-SR [47] and AMRSR [32]. Note that these multi-RefSR methods are trained with a collection of reference images. In Table 4, our model trained with single reference image performs better than CIMR-SR and AMRSR with many reference images, which further demonstrate the superiority of our proposed DATSR.

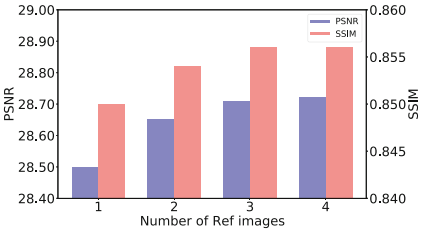


Fig. 9. Effect on #Ref images.

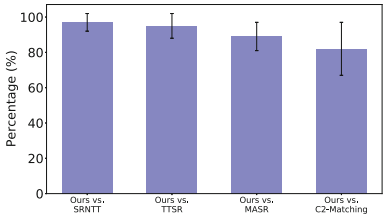


Fig. 10. User study.

Table 3. Performance in terms of different similarity levels on CUFED5 test set.

Similarity levels	L1		L2		L3		L4		Average	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CrossNet [57]	25.48	0.764	25.48	0.764	25.47	0.763	25.46	0.763	25.47	0.764
SRNTT-rec [56]	26.15	0.781	26.04	0.776	25.98	0.775	25.95	0.774	26.03	0.777
TTSR-rec [48]	26.99	0.800	26.74	0.791	26.64	0.788	26.58	0.787	26.74	0.792
C^2-Matching-rec [12]	28.11	0.839	27.26	0.811	27.07	0.804	26.85	0.796	27.32	0.813
DATSR-rec (Ours)	**28.50**	**0.850**	**27.47**	**0.820**	**27.22**	**0.811**	**26.96**	**0.803**	**27.54**	**0.821**

4.3 More Evaluation Results

Perceptual Metric. We further use the perceptual metric LPIPS [52] to evaluate the visual quality of the generated SR images on the CUFED5 and WR-SR testing sets. Recently, this metric is also widely used in many methods [28,29]. In general, smaller LPIPS corresponds to the better performance for RefSR. As shown in Table 5, our model achieves smaller LPIPS than C^2-Matching. Thus, our model generates SR images with better quality than C^2-Matching.

User Study. To further evaluate the visual quality of the SR images, we conduct the user study to compare our proposed method with previous state-of-the-art methods, including SRNTT [56], TTSR [48], MASA [27] and C^2-Matching [12] on the WR-SR testing set. The user study contains 20 users, and each user

is given multiple pairs of SR images where one is our result. Then, each user chooses one image with better visual quality. The final percentage is the average user preference of all images. In Fig. 10, over 80% of the users prefer that our results have better quality than existing RefSR methods.

4.4 Discussion on Model Size

To further demonstrate the effectiveness of our model, we also show the comparison of model size (*i.e.*, the number of trainable parameters) with the state-of-the-art model (*i.e.*, C^2-Matching [12]) in Table 6. Our model has a total number of 18.0M parameters and achieves PSNR and SSIM of 28.72 and 0.856, respectively. The results demonstrate that our proposed model outperforms C^2-Matching with a large margin, although our model size is higher than this method. The part of our model size comes from the Swin Transformer in the RFA module. More discussions of other RefSR models are put in the supplementary materials.

Table 4. Comparisons with Multi-RefSR on the CUFED5 testing set.

Methods	CIMR-SR [47]	AMRSR [32]	DATSR-rec
w/ rec. loss	26.35/0.789	28.32/0.839	**28.72/0.856**
w/ all losses	26.16/0.781	27.49/0.815	**27.95/0.835**

Table 5. Comparisons of LPIPS [52] with C^2-Matching.

Methods	CUFED5	WR-SR
C^2-Matching [12]	0.164	0.219
DATSR (Ours)	**0.140**	**0.211**

Table 6. Comparisons of model size and performance with C^2-Matching.

Methods	Params	PSNR	SSIM
TTSR-rec [48]	6.4M	27.09	0.804
C^2-Matching-rec [12]	8.9M	28.24	0.841
DATSR-rec (Ours)	18.0M	**28.72**	**0.856**

Table 7. Ablation study on the RDA and RFA modules.

Methods	PSNR	SSIM
RDA (w/ feature warping)	28.25	0.844
RFA (w/ ResNet blocks)	28.50	0.850
DATSR-rec	**28.72**	**0.856**

4.5 Ablation Study

We first investigate the effectiveness of RDA and RFA in Table 7. Specifically, we replace the texture transfer method in RDA with a feature warping based on the most relevant correspondence, and replace RFA with several convolutional neural networks (CNNs). The model with feature warping or CNNs is worse than original model with RDA or RFA. Therefore, RDA is able to discover more relevant features especially when the correspondence is not inaccurate.

For RFA, our model has better performance than the directly using simple CNNs. Nevertheless, with the help of RDA, training with CNNs still outperforms C^2-Matching with large margin. Therefore, it verifies that the effectiveness of RFA and it is able to aggregate the features at different scales. More discussions on ablation studies are put in the supplementary materials.

5 Conclusion

In this work, we propose a novel reference-based image super-resolution with deformable attention Transformer, called DATSR. Specifically, we use texture feature encoders module to extract multi-scale features and alleviate the resolution and transformation gap between LR and Ref images. Then, we propose reference-based deformable attention module to discover relevant textures, adaptively transfer the textures, and relieve the correspondence mismatching issue. Last, we propose a residual feature aggregation module to fuse features and generate SR images. Extensive experiments verify that DATSR achieves the state-of-the-arts performance as it is robust to different brightness, contrast, and color between LR and Ref images, and still shows good robustness even in some extreme cases, when the Ref images have no useful texture information. Moreover, DATSR trained with a single Ref image has better performance than existing Multi-RefSR methods trained with multiple Ref images.

Acknowledgements. This work was partly supported by Huawei Fund and the ETH Zürich Fund (OK).

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning, pp. 214–223 (2017)
2. Cao, H., ET AL.: Swin-Unet: unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
3. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5972–5981 (2022)
4. Dai, J., et al.: Deformable convolutional networks. In: IEEE International Conference on Computer Vision, pp. 764–773 (2017)
5. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 11065–11074 (2019)
6. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)
7. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
8. Guo, Y., et al.: Closed-loop matters: dual regression networks for single image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5407–5416 (2020)
9. Guo, Y., Luo, Y., He, Z., Huang, J., Chen, J.: Hierarchical neural architecture search for single image super-resolution. IEEE Sig. Process. Lett. **27**, 1255–1259 (2020)
10. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5197–5206 (2015)
11. Hui, Z., Li, J., Wang, X., Gao, X.: Learning the non-differentiable optimization for blind super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2093–2102 (2021)

12. Jiang, Y., Chan, K.C., Wang, X., Loy, C.C., Liu, Z.: Robust reference-based super-resolution via c2-matching. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2103–2112 (2021)
13. Jo, Y., Kim, S.J.: Practical single-image super-resolution using look-up table. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 691–700 (2021)
14. Kar, A., Biswas, P.K.: Fast bayesian uncertainty estimation and reduction of batch normalized single image super-resolution network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4957–4966 (2021)
15. Khrulkov, V., Babenko, A.: Neural side-by-side: predicting human preferences for no-reference super-resolution evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4988–4997 (2021)
16. Kong, X., Zhao, H., Qiao, Y., Dong, C.: ClassSR: a general framework to accelerate super-resolution networks by data characteristic. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 12016–12025 (2021)
17. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690 (2017)
18. Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W.: Feedback network for image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3867–3876 (2019)
19. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: image restoration using swin transformer. In: IEEE International Conference on Computer Vision Workshops, pp. 1833–1844 (2021)
20. Liang, J., Lugmayr, A., Zhang, K., Danelljan, M., Van Gool, L., Timofte, R.: Hierarchical conditional flow: a unified framework for image super-resolution and image rescaling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4076–4085 (2021)
21. Liang, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Mutual affine network for spatially variant kernel estimation in blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4096–4105 (2021)
22. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144 (2017)
23. Liu, J., Zhang, W., Tang, Y., Tang, J., Wu, G.: Residual feature aggregation network for image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2359–2368 (2020)
24. Liu, Q., Liu, C.: A novel locally linear KNN model for visual recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9446–9454 (2015)
25. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE International Conference on Computer Vision, pp. 10012–10022 (2021)
26. Liu, Z., et al.: Video Swin transformer. arXiv preprint arXiv:2106.13230 (2021)
27. Lu, L., Li, W., Tao, X., Lu, J., Jia, J.: MASA-SR: matching acceleration and spatial adaptation for reference-based image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6368–6377 (2021)
28. Lucas, A., Lopez-Tapia, S., Molina, R., Katsaggelos, A.K.: Generative adversarial networks and perceptual losses for video super-resolution. IEEE Trans. Image Process. **28**(7), 3312–3327 (2019)

29. Lugmayr, A., Danelljan, M., Timofte, R.: Ntire 2020 challenge on real-world image super-resolution: Methods and results. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 494–495 (2020)
30. Matsui, Y., et al.: Sketch-based manga retrieval using manga109 dataset. Multimedia Tools Appl. **76**(20), 21811–21838 (2016). https://doi.org/10.1007/s11042-016-4020-z
31. Mei, Y., Fan, Y., Zhou, Y.: Image super-resolution with non-local sparse attention. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3517–3526 (2021)
32. Pesavento, M., Volino, M., Hilton, A.: Attention-based multi-reference learning for image super-resolution. In: IEEE International Conference on Computer Vision, pp. 14697–14706 (2021)
33. Sajjadi, M.S., Scholkopf, B., Hirsch, M.: EnhanceNet: single image super-resolution through automated texture synthesis. In: IEEE International Conference on Computer Vision, pp. 4491–4500 (2017)
34. Shim, G., Park, J., Kweon, I.S.: Robust reference-based super-resolution with similarity-aware deformable convolution. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 8425–8434 (2020)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
36. Song, X., et al.: Channel attention based iterative residual learning for depth map super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5631–5640 (2020)
37. Sun, L., Hays, J.: Super-resolution from internet-scale scene matching. In: IEEE International Conference on Computational Photography, pp. 1–12 (2012)
38. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9446–9454 (2018)
39. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
40. Wang, L., Kim, T.K., Yoon, K.J.: EventSR: from asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 8315–8325 (2020)
41. Wang, L., et al.: Exploring sparsity in image super-resolution for efficient inference. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4917–4926 (2021)
42. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: EDVR: video restoration with enhanced deformable convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (2019)
43. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-ESRGAN: training real-world blind super-resolution with pure synthetic data. In: IEEE International Conference on Computer Vision, pp. 1905–1914 (2021)
44. Wang, X., et al.: ESRGAN: enhanced super-resolution generative adversarial networks. In: European Conference on Computer Vision Workshops (2018)
45. Xie, Y., Xiao, J., Sun, M., Yao, C., Huang, K.: Feature representation matters: end-to-end learning for reference-based image super-resolution. In: European Conference on Computer Vision, pp. 230–245 (2020)
46. Xing, W., Egiazarian, K.: End-to-end learning for joint image demosaicing, denoising and super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3507–3516 (2021)

47. Yan, X., Zhao, W., Yuan, K., Zhang, R., Li, Z., Cui, S.: Towards content-independent multi-reference super-resolution: adaptive pattern matching and feature aggregation. In: European Conference on Computer Vision, pp. 52–68 (2020)
48. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5791–5800 (2020)
49. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: IEEE Conference on International Conference on Computer Vision, pp. 4791–4800 (2021)
50. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4791–4800 (2021)
51. Zhang, K., Zuo, W., Zhang, L.: FFDNet: toward a fast and flexible solution for CNN-based image denoising. IEEE Trans. Image Process. **27**(9), 4608–4622 (2018)
52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
53. Zhang, W., Liu, Y., Dong, C., Qiao, Y.: RankSRGAN: generative adversarial networks with ranker for image super-resolution. In: IEEE International Conference on Computer Vision, pp. 3096–3105 (2019)
54. Zhang, Y., Li, K., Li, K., Fu, Y.: MR image super-resolution with squeeze and excitation reasoning attention network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 13425–13434 (2021)
55. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: European Conference on Computer Vision, pp. 286–301 (2018)
56. Zhang, Z., Wang, Z., Lin, Z., Qi, H.: Image super-resolution by neural texture transfer. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7982–7991 (2019)
57. 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: European Conference on Computer Vision, pp. 88–104 (2018)
58. Zhou, R., Susstrunk, S.: Kernel modeling super-resolution on real low-resolution images. In: IEEE International Conference on Computer Vision, pp. 2433–2443 (2019)
59. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9308–9316 (2019)