DOI: 10.1111/cgf.14679
Pacific Graphics 2022
N. Umetani, E. Vouga, and C. Wojtan
(Guest Editors)

Volume 41 (2022), Number 7

Ref-ZSSR: Zero-Shot Single Image Superresolution with Reference Image

Xianjun Han[†1], Xue Wang[1], Huabin Wang[1], Xuejun Li[1] and Hongyu Yang[2]

[1] School of Computer Science and Technology, Anhui University, China
[2]College of Computer Science, Sichuan University, China

Abstract

Single image superresolution (SISR) has achieved substantial progress based on deep learning. Many SISR methods acquire pairs of low-resolution (LR) images from their corresponding high-resolution (HR) counterparts. Being unsupervised, this kind of method also demands large-scale training data. However, these paired images and a large amount of training data are difficult to obtain. Recently, several internal, learning-based methods have been introduced to address this issue. Although requiring a large quantity of training data pairs is solved, the ability to improve the image resolution is limited if only the information of the LR image itself is applied. Therefore, we further expand this kind of approach by using similar HR reference images as prior knowledge to assist the single input image. In this paper, we proposed zero-shot single image superresolution with a reference image (Ref-ZSSR). First, we use an unconditional generative model to learn the internal distribution of the HR reference image. Second, a dual-path architecture that contains a downsampler and an upsampler is introduced to learn the mapping between the input image and its downscaled image. Finally, we combine the reference image learning module and dual-path architecture module to train a new generative model that can generate a superresolution (SR) image with the details of the HR reference image. Such a design encourages a simple and accurate way to transfer relevant textures from the reference high-definition (HD) image to LR image. Compared with using only the image itself, the HD feature of the reference image improves the SR performance. In the experiment, we show that the proposed method outperforms previous image-specific network and internal learning-based methods.

CCS Concepts
• ***Computing methodologies*** → *Reconstruction;*

1. Introduction

The aim of SISR [BWKN21,CHQ*22] recovers an HR image from its degraded LR counterpart. Compared with previous nondeep SR methods, deep-learning-based methods have received an immense boost in performance [KBN*20, PW20] in this field. Although these methods have better visual quality and can eliminate undesired artifacts, they require image pairs [LTH*17] in supervised learning mode or demand a large scale of training data in unsupervised learning mode [LDT20,JW21]. However, a large quantity of paired images does not always exist. In addition, these methods are vulnerable to real images because these LR images are obtained with an ideal downscaling kernel (usually a Gaussian kernel) from HR images.

To address this issue, several internal learning-based methods train an image-specific network [KJK20, BKSI19] and employ a single input image to avoid the dilemma caused by the lack of training data. InGAN [SBII19] trains on a single input image and learns its internal distribution of patches. This method can remap the input to any size or shape in a single feedforward pass while preserving the same internal patch distribution. Consistent with this approach, SinGAN [SDM19] captures the internal distribution of patches within the image and can generate high-quality samples that carry the same visual content as the input image. ZSSR [SCI18] exploits the internal recurrence of information inside a single image and allows the acquisition process to be unknown or nonideal.

For zero-shot blind SR approaches, many excellent image-specific network architectures exist. KernelGAN [BKSI19] estimates a downscaling kernel for blind SR based on internal learning with linear CNNs within a generative adversarial framework. DBPB [KJK20] proposes that blind SR can be modeled as a two-stage optimization problem, which conducts downscaling kernel estimation followed by SR network training with the estimated kernel. Similarly, DualSR [EPC21] proposes a dual-path architecture that learns low-to-high mapping with a downsampler and an upsampler. In the DualSR architecture, the upsampler and downsam-

[†] hxj@ahu.edu.cn

pler are trained simultaneously, and they improve each other using cycle consistency losses.

However, the final SR images in the abovementioned methods are obtained by only using information about the images themselves or by estimating the downscaling kernel by downsampling the input images, which limits the effect of image superresolution. Reference-based image superresolution (RefSR) [SH12, TSG13] can fully utilize the information of the HR reference image, which transfers HR textures from a given reference HR image to produce visually pleasing results. Nevertheless, the extraction of HD texture details and high-level semantic features also requires large-scale training data.

To summarize, in terms of whether to use the external dataset and the reference image, the task of SR can be divided into four quadrants. As shown in Fig. 1, distinguishing whether using the reference and external datasets helps us to understand the dependence of this method on a priori information. After categorizing existing approaches into these classes, one remaining research gap naturally reveals itself: single image with the reference image. We argue that this direction is promising in terms of generating HD real-world images, and we will also attempt to propose a network architecture in this direction.

To address these problems and fill the gap in the field of image-specific SR with reference images, we propose a novel, image-specific network for SISR. Specifically, three network modules optimized for image generation tasks are proposed. First, we introduce a pyramid of fully convolutional generative adversarial networks (GANs) to produce high-quality results that preserve the internal patch statistics of the reference image. Second, a dual-path architecture that learns low-to-high resolution mapping of the input image is introduced. The dual-path architecture contains a downsampler that learns the degradation process and an upsampler that learns the superresolution process, which are trained simultaneously and improved by cycle-consistency losses. Finally, we introduce another generator to obtain an HD texture similar to the reference image. We formulate this generator to have the same structure as the reference-image generator and inherit its network parameters after training. In addition, this generator has an attention mechanism, which enables our approach to learn a more powerful feature representation and texture extraction. Meanwhile, a discriminator is also employed to fit the distribution of HD texture details of the reference image.

The main contributions of our work are as follows:

- To the best of our knowledge, we are one of the first researchers to introduce RefSR into an image-specific network.
- We design a three-stage learning pipeline, including Ref-learning, Self-learning and Alliance-learning. Such a design enables our approach to achieve a better visual result with only one input image and one reference image in the training process, which improves the current SR method based on internal learning.
- We introduce a texture transmission mode with a balanced attention mechanism, which encourages a simple and accurate way to transfer relevant textures from the reference image to the LR image.

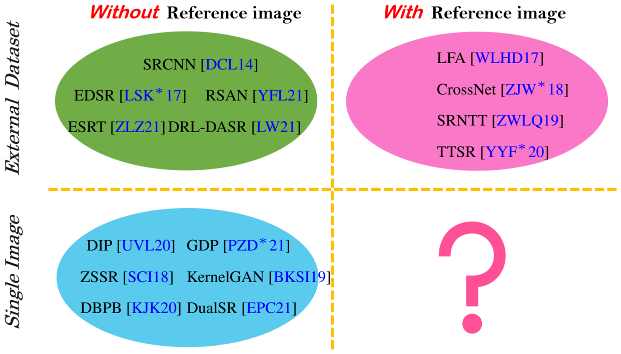

Figure 1: *The taxonomy of RefSR and the corresponding representative methods. This taxonomy distinguishes the methods of reference images and data used for solving SR models.*

It should be noted that almost all individual components of our framework have appeared in previous work, although the specific instantiations may be different. However, the superiority of our framework relative to previous work is not explained by any single design choice but by their composition. Additionally, such a design fills the gap in Fig.1.

2. Related Work

We review previous works of SISR and image-specific networks for superresolution that are the most relevant to our work.

Single Image Superresolution. Due to the powerful model-fitting capabilities of convolutional neural networks, the visual effect of SISR [HSU20, XSGW21, WLL*21, LKLE21] has made extraordinary progress compared to traditional methods. SRCNN [DCL14] proposes a three-layer convolutional network to learn the nonlinear mapping from LR images to HR images. EDSR [LSK*17] proposes a very deep and wide network in which batch normalization layers are removed and a residual scaling technique is utilized. Meta-SR [HMZ19] applies meta-learning to predict the weights of filters for different scale factors; however, it does not exploit scale information during feature learning. To solve this problem, the RSAN [YFL21] introduces a residual scale attention network that is employed as prior knowledge to learn discriminative features.

Recently, because a transformer can capture long-term information among sequence elements, it has been successfully applied in vision tasks [CHT20,WLX*21]. ViT [AD20] is the first work to use a transformer in image tasks. ViT flattens the 2D image patches in a vector and feeds them into the transformer. DETR [CMS*20] further discards certain complex, handcrafted operations and models the prediction of a set of objects. ESRT [ZLZ21] proposes a hybrid transformer, where a CNN-based SR network is designed in the front to extract deep features. The network uses a lightweight CNN backbone to extract deep SR features at low computational cost and employs an efficient transformer with novel, efficient multihead attention. This module has a low computational cost and achieves competitive results.

To fill the gap of the degradation process from an HR image

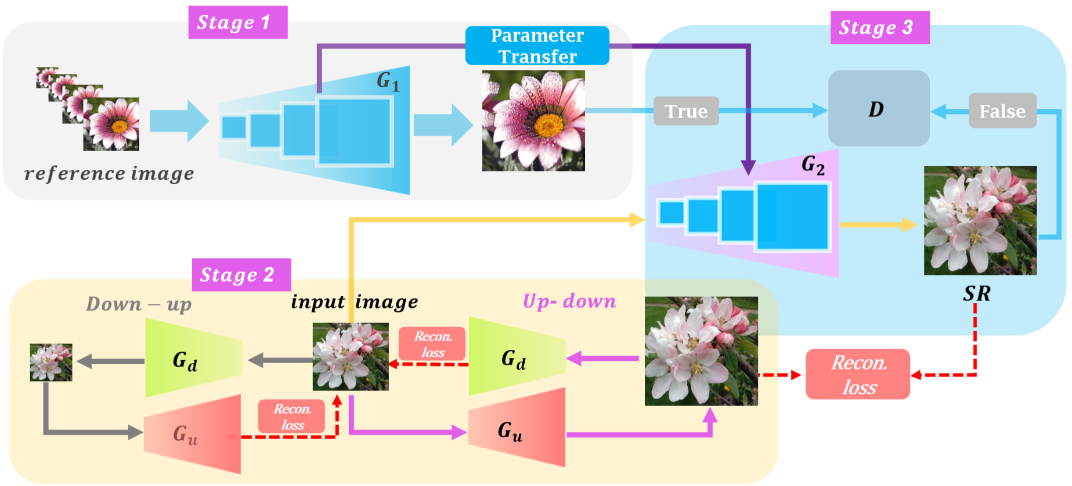

Figure 2: *Network architecture of the proposed Ref-ZSSR. **Stage 1(Ref-learning)**: The HD reference image is first fed into G_1 to capture the internal statistics. **Stage 2(Self-learning)**: The upscaling network G_u and the downscaling network G_d construct the dual back-projection architecture. The "Down-up" procedure is used to train G_u, which refers to how the input image passes through G_d to generate the downscaled image and then passes through G_u to reconstruct the input image. The "Up-down" procedure is an opposite mapping process to train G_d. **Stage 3(Alliance-learning)**: G_2 inherits the network parameters of G_1 and combines the discriminator D to generate the final SR image. The SR results generated by G_1 guides the distribution and the upscaling network G_u restricts the reconstruction.*

to an LR image, blind SR [LLG*21] has been proposed to process the unknown degradations. DRL-DASR [LW21] estimates the degradation information with a trainable encoder in the latent feature space, and the degradation encoder is trained with contrastive learning. Such a framework can achieve satisfactory SR results with a single forward pass. AMNet-RL [HLWG21] incorporates kernel estimation into the SR network and optimizes the blind SR model with undifferentiable perceptual metrics under the reinforcement learning framework. KOALAnet [KSK20] employs a dynamic kernel strategy that adapts the SR network to a specific degradation and extends the noniterative framework to spatially variant degradation for local kernel estimation.

Superresolution with Reference Images or Image-Specific Networks. Since the reference image can provide similar HD textures, reference image-based SR can harvest more accurate details from the reference image. Wang et al. [WLHD17] proposes the recurrent application of nonuniform warping before feature synthesis to the reference image. CrossNet [ZJW*18] adopts optical flow to align the LR and the reference images at different scales and concatenates them into the corresponding layers of the decoder. SRNTT [ZWLQ19] applies patch matching to search for proper reference information between the input image and the reference image; however, this method ignores the relevance between original features and swapped features. To solve this problem, TTSR [YYF*20] proposes a texture transformer network in which the LR and the reference images are formulated as queries and keys in a transformer. The network includes a learnable texture extractor, relevance embedding module, hard-attention module for texture transfer, and soft-attention module for texture synthesis. Such a design achieves significant improvements over state-of-the-art approaches on both quantitative and qualitative evaluations.

Although existing SISR and RefSR networks have achieved promising results, they are trained for SR with a large scale of training data. To overcome this limitation, an image-specific network is proposed for SR using only the input image itself. DIP [UVL20] assumes that the generator network is sufficient to capture many low-level image statistical priors. This work replaces the regularization term with an implicit prior captured by the convolutional network. Hence, such a design is effective for various image-restoration tasks, including reconstructing the SR image. Inspired by this finding, GDP [PZD*21] provides an effective way to exploit the image prior captured by a generative adversarial network and allows the generator to be fine-tuned on the fly in a progressive manner.

ZSSR [SCI18] and KernelGAN [BKSI19] attempt to train image-specific CNNs for superresolving each input LR without any pretraining step. The CNNs trained with the image pairs themselves will be capable of inferring specific relationships across different scales. The idea of self-supervision with internal statistics requires no effort to gather a large external training dataset. Nevertheless, it is difficult to exploit recurring information across scales to robustly perform SR with this kind of input image. Hence, these approaches can only produce favorable SR outputs for a very limited set of images with frequently recurring content across scales.

Therefore, to address these problems, we combine the reference image learning module and the image-specific network architecture to train a GAN. Such a design can generate the SR image with the details of the HR reference image and only the input image itself. Moreover, the performance of our approach can be further improved by adjusting the appropriate network structure and embedding the transformer into the architecture.

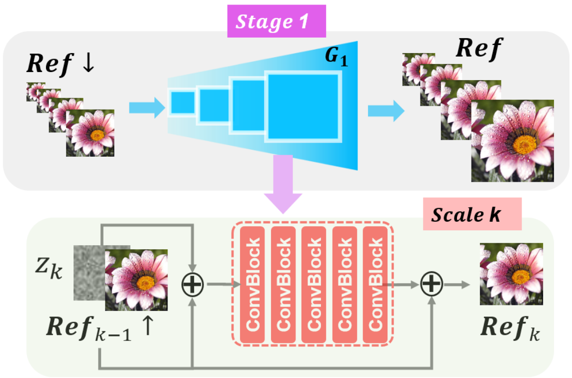

Figure 3: *Network architecture of the reference image internal learning in the first stage. Ref ↓ represents the downsampled version of the reference image. $Ref_{k-1} \uparrow$ represents the upsampled version of the reference image at scale $k-1$. At scale k, G_1 is fed the noise Z_k and $Ref_{k-1} \uparrow$ to learn the internal statistics by reconstructing the reference image.*

3. Approach

In this section, we introduce the proposed Ref-ZSSR. We integrate the RefSR architecture into the image-specific network to enhance the performance of SR by transferring relevant texture information from the reference image. We provide an overall description of the proposed Ref-ZSSR accompanied by an optimization target and further analyze its modules in detail.

3.1. Overall network framework

As shown in Fig. 2, the proposed Ref-ZSSR pipeline consists of three stages: the reference image internal learning (Ref-learning) stage, the input-image-itself internal learning (Self-learning) stage, and the alliance-learning stage (ALS).

In the first stage, we use G_1 to capture the internal statistics of the reference image. G_1 is a pyramid of fully convolutional GANs; each GAN is responsible for learning the patch distribution at a different scale. Once trained, G_1 stores the parameters that can construct HD texture details and complex structures. In the second stage, we employ two networks to minimize a dual back-projection loss. The upscaling G_u network is trained to reconstruct a given input image from the LR image generated by the downscaling G_d network, and G_d is trained to downscale the HR image generated by the G_u to be as similar as possible to the given input image. In the last stage, we combine Ref-learning and Self-learning into new GANs, where the initialization of parameters in G_2 directly inherited from G_1. In addition, the SR image provided by G_u is used to reconstruct the output of G_2. The reference image and final SR image are fed to the discriminator, which is responsible for fitting the distribution of the SR and reference images.

3.2. Ref-learning

In the stage of reference image internal learning, we aim to store relevant network parameters that can relate to the HD texture of the reference image. We utilize the pyramid network architecture to learn the image's patch statistics across multiple scales. Similar to SinGAN [SDM19], the network architecture is deployed as a GAN. Given the reference image, it is downsampled to different scales to form an image pyramid. Formally, we have

$$Ref \downarrow = downsampling(Ref), \tag{1}$$

where *downsampling* denotes the downsampling operation and Ref represents the reference image. Thus, $Ref \downarrow$ and Ref constitute $LR \longleftrightarrow HR$ in the form of supervision.

As shown in Fig. 3, the model consists of the pyramid of generator G_1 and the corresponding discriminator D. At each scale k of the image pyramid, adversarial training is employed through G_1, which learns to fool the associated discriminator D that attempts to distinguish whether an image is original or generated by the generator.

To ensure that the generator has not only the function of reconstructing the original image but also the ability of SR, at every scale, Gaussian noise and the downsampled image are concurrently sent to the generator. Thus, noise has the role of injecting HD texture details into the image through internal learning. The process of generation starts at the coarsest scale and sequentially passes to the scale of the original image size, where the generators and discriminators have the same receptive field.

Hence, at scale k, the generator G_1 adds details by accepting an upsampled image and Gaussian noise Z_k with the same size. This process can be expressed as:

$$Ref_k = G_{1_k}(Z_k, (Ref_{k-1}) \uparrow) + (Ref_{k-1}) \uparrow, \tag{2}$$

where ↑ denotes the upsampling and $Ref_{k-1} \uparrow$ represents the upsampled version of the reference image at scale $k-1$. The training loss at the k scale is an adversarial term and a reconstruction term:

$$\min_{G_{1_k}} \max_{D_k} L_{adv}(G_{1_k}, D_k) + \lambda \parallel Ref_k - Ref_{k-1} \uparrow \|_2, \tag{3}$$

where the adversarial loss L_{adv} is

$$\begin{aligned} \min_{G_{1_k}} \max_{D_k} L_{adv}(G_{1_k}, D_k) =& \mathbb{E}[logD_k(Ref_{real_k})+ \\ & \mathbb{E}[log(1-D_k(Ref_k))]. \end{aligned} \tag{4}$$

Here, Ref_{real_k} is the real reference image with the size of scale k, and Ref_k is generated by Eqn.2 with G_{1_k}. L_{adv} penalizes the distribution of patches in Ref_k and generated samples. The reconstruction term measuring the error from the upsampled version ensures that the final generated image is closer to the original image rather than the reference image.

The generator is a fully convolutional network with 5 conv-blocks of the form Conv(3 × 3)-BatchNorm-LeakyReLU [IS15] with 32 kernels per block at each scale. The fully convolutional network contributes to generating reference images of arbitrary size by adjusting the dimensions of the noise maps. Thus, G_1 stores the parameters that can construct the HD texture details of the reference image.

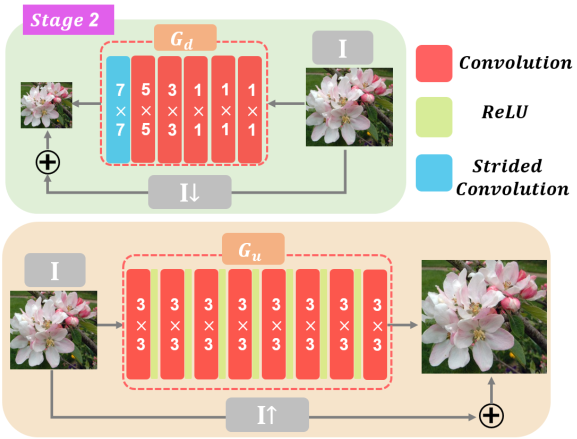

Figure 4: *Network architecture of the input image internal learning in the second stage. I ↑ and I ↓ represent the upsampled and downsampled versions of the input image, respectively.*

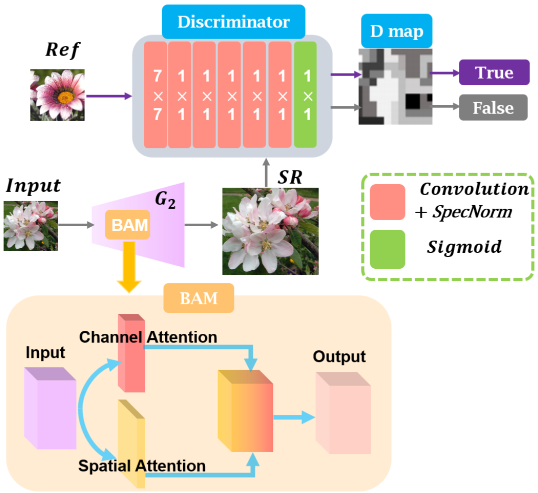

Figure 5: *Network architecture of alliance learning in the third stage. The BAM does not change the size of the feature map and can adapt to any position between the ConvBlocks. In this article, we insert this structure in front of the last ConvBlock of G_1 to form G_2.*

3.3. Self-learning

In this stage, the input image internal learning jointly comprises two image-specific networks. The first network involves reconstructing the input image from the LR image generated by the downscaling network, and the second network reconstructs the input image from the HR image generated by the upscaling network.

Similar to DBPB [KJK20] and DualSR [EPC21], we introduce dual back-projection loss into the opposite-direction training, from the superresolved image to the input image and from the downscaled image to the input image. As shown in stage 2 in Fig. 2, during the process of downup, the input image is downscaled with the downscaling network G_d, and then the upscaling network G_u upscales the downscaled image, generating the downup image. Hence, the difference between the input image and the downup image is applied to train the upscaling network G_u.

In parallel, in the process of updown, the input image is superresolved by the upscaling network G_u, and the downscaling network G_d downscales the superresolved image to generate an updown image. For almost the same reason, the downscaling network G_d is trained to minimize the difference between the input image and the updown image.

The network architectures of G_u and G_d are shown in Fig. 4 in detail. As shown, G_d consists of one strided convolution layer and five convolution layers without nonlinear activation. As conducted in DBPB [KJK20], this process imitates a convolution with a large kernel followed by subsampling. The upscaling network G_u includes eight convolution layers, and ReLU activation is performed for them all, except for the last layer. The network is equivalent to the network employed in KernelGAN [BKSI19].

Specifically, the upscaling network G_u is trained to project the downscaled image onto the input image, and the downscaling network G_d is trained to reconstruct the input image from the superresolved image. This process can be derived as

$$L_{cycle} = \frac{1}{i^2}|G_u(G_d(I)) - I|_1 + \frac{1}{i^2}|G_d(G_u(I)) - I|_1, \qquad (5)$$

where I is a given $i \times i$ image patch. G_d and G_u are the downscaling network and upscaling network, respectively.

By minimizing the loss, the upscaling and downscaling networks are trained to revert LR and HR images generated by each other to the input image. This finding indicates that the downscaling network and upscaling network are inverse functions. In addition, complementary training can improve downscaling kernel estimation, and the upscaled image generated by the upscaling network G_u can be fed to the next stage to improve the SR performance.

3.4. Alliance learning

As the reference network G_1 and the upscaling network G_u have been trained, we use these two trained networks to assist a GAN network to obtain an HD texture similar to that of the reference image. More specifically, we formulate the generator G_2 in the GAN to have the same structure as G_1. In addition, the parameter of G_2 inherits the network parameters of G_1, which has less time consumption. We introduce a discriminator that attempts to distinguish texture in the generated samples from that in the reference image. As shown in stage 3 in Fig. 2, such a design encourages a simple and accurate way to transfer relevant textures from the Ref image to the LR image.

The specific details are shown in Fig. 5, where G_2 loads the parameters of G_1, which has the potential to reconstruct HR images to some extent. By reconstructing the generated image by G_u and fine-tuning the G_2 network parameters, G_2 can generate HD textures similar to those of the reference image while maintaining the essential characteristics of the input image.

As shown in Fig. 5, before the input image is fed to G_2, it is first enlarged to the desired resolution with the resize operation. In addition, we introduce a lightweight and efficient balanced attention mechanism (BAM) [WHS21], which does not change the size of the feature map and can adapt to any position between the *ConvBlocks*. We plug this structure in front of the last *ConvBlock* of G_1 to form G_2. Such a design can avoid error accumulation and enhance the feature extraction ability. The BAM consists of a channel attention module and a spatial attention module. The channel attention information is extracted by *Avgpool*, refined by a multilayer perceptron with the bottleneck architecture, and then activated by the *Sigmoid* function. The spatial attention information is extracted by *Maxpool*, refined by a convolutional layer and then activated by the *Sigmoid* function. The BAM module is equivalent to that in [WHS21]. With BAM, G_2 has a better ability to capture image details and ultimately improves superresolution performance.

In addition, D tries to distinguish real reference patches from those generated by G_2(fake). G_2 learns to generate the SR image while fooling D and maintains the same distribution as the reference image. D trains to output a heatmap, which is referred to as the D-map. The adversarial loss is the pixelwise *MSE* difference between the output D-map and the label map. The labels for training D entail a map of all those for crops extracted from the reference image and a map of all zeros for crops extracted from the generated image by G_2. Combined with the reconstruction loss, the process can be expressed as:

$$\min_{G_2}\max_{D} L_{adv}(G_2, D) + \lambda \parallel G_2(I\uparrow) - G_u(I) \parallel_2, \tag{6}$$

where the adversarial loss L_{adv} is similar to Eqn.4 and penalizes the distribution of patches in the reference image and the samples generated by G_2. $\uparrow$ represents the upsampling operation, and the reconstruction loss ensures that the generator G_2 can be consistent with the generated image of G_u.

D is a fully convolutional patch discriminator, as introduced in [BKSI19], with no pooling or strides and a 7×7 convolution filter followed by six 1×1 convolutions, including *Spectral* normalization, *Batch* normalization, *ReLU*, and *Sigmoid* activation. Each pixel in the D-map indicates how likely its surrounding patch is to be drawn from the learned patch distribution.

4. Experiments and results

4.1. Implementation details

For a fair comparison and to verify the effectiveness of the proposed framework, in the Ref-learning and Self-learning stages, the learning parameters are equivalent to those in SinGAN [SDM19] and DBPB [KJK20]. In the last Alliance-learning stage, the GAN was trained with the Adam optimizer for approximately 2000 iterations. The initial learning rate for the networks was set to 10^{-4}. For the scale factor of 4 or higher, we repeated ×2 SR for better performance instead of directly superresolving the input image. In addition, during the experiment, we found that the second stage and the third stage can be simultaneously trained. By adjusting the appropriate hyperparameters, this method can also achieve a satisfactory SR effect. Because the network structures are oriented to image specificity, hyperparameters should be adjusted according to different images.

4.2. Dataset and evaluation method

The reference image and input image are selected from the public RefSR dataset CUFED5 [ZWLQ19]. There are 126 testing images in this testing set, and each testing image is accompanied by 4 reference images with different similarity levels. In addition, we test the network framework on Set14 [ZEP10] and BSD100 [MFTM01].

To evaluate the effectiveness of the proposed method, we compare our model with other state-of-the-art SISRs based on image specificity, which include DIP [UVL20], SinGAN [SDM19], ZSSR [SCI18], KernelGAN [BKSI19] and DBPB [KJK20]. These methods have made remarkable achievements in the field of SISR by using the input image for training. The quantitative comparison uses the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). In all of the experiments, bicubic interpolation is utilized as the upsampling method.

4.3. Qualitative evaluation

For a fair comparison, the parameters applied in the experiment are the values recommended in the related corresponding paper. In addition, we optimized to ensure the highest quality of the final generated image. The final results are shown in Fig. 6. Our model achieved the best performance compared with other excellent SISR methods based on image-specific visual quality.

DIP [UVL20] explores the prior information of the CNN network through reconstruction and fine-tunes the hyperparameters to obtain HR images. Although this method is simple and effective, fine-tuning the parameters of the network is a laborious task. In addition, it is not enough to capture the internal statistical features of the input image only by the reconstructing process. Hence, the effect of this method in superresolution is not desirable.

KernelGAN [BKSI19] estimates the SR kernel that preserves the distribution of patches in the LR image. Its generator is trained to produce a downscaled version of the LR test image, while its discriminator cannot distinguish between the downscaled image and the original LR image. In essence, this method is mainly used to estimate the blur kernel and uses ZSSR [SCI18] to generate the SR results. Thus, except for some small local diversity, there is no obvious difference in visual effect between KernelGAN and ZSSR.

DBPB [KJK20] further assumes that the SR network not only depends on the estimated kernel but can also improves downscaling kernel estimation. Hence, DBPB jointly trains two image-specific networks, resulting in better SR performance. However, DBPB does not use a discriminator to learn the distribution between the real patches and the fake patches generated by the generator. In terms of visual effect, the performance of DBPB is slightly superior to that of KernelGAN.

As mentioned above, ZSSR [SCI18] trains a model to infer complex image-specific HR-LR relations and then applies ZSSR to these learned relations on the LR input image to produce the HR

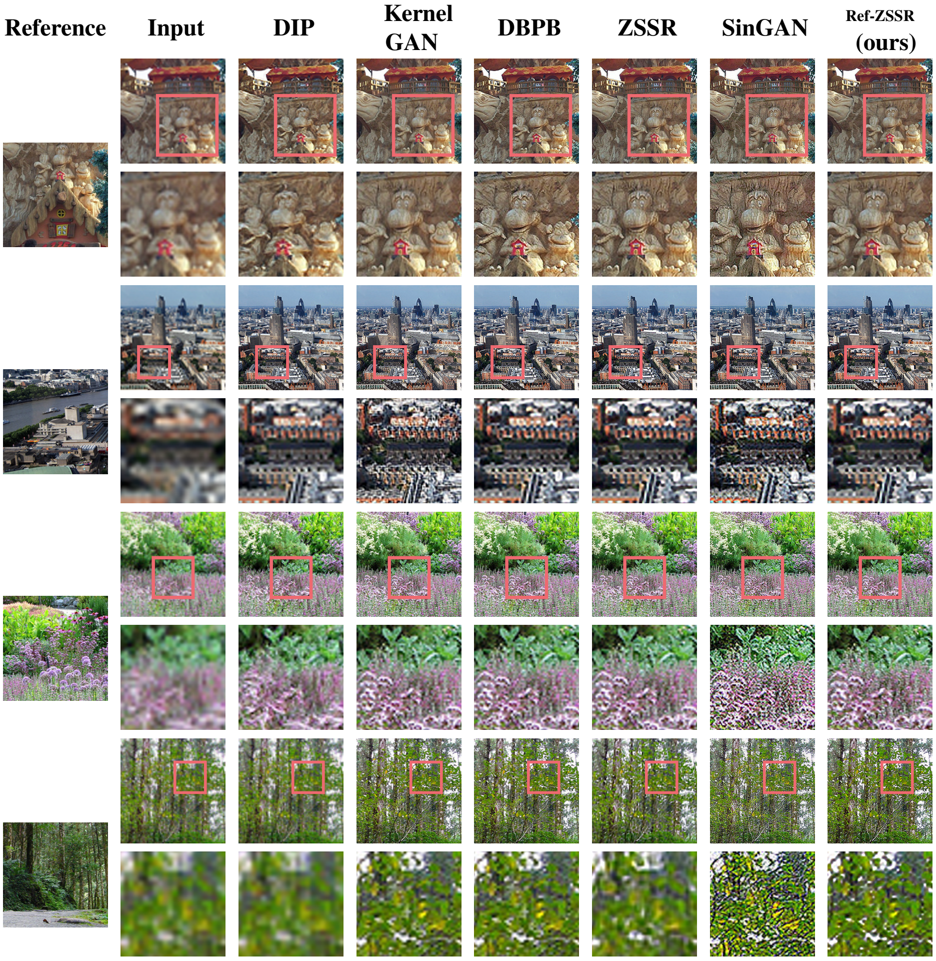

Figure 6: ***Visual comparison among different SISRs based on image-specific methods.***

output. Although ZSSR exploits the internal recurrence of information within a single image and trains a small image-specific CNN at test time, its texture details are not very clear.

Unlike ZSSR, SinGAN [SDM19] employs a pyramid network structure when facing cross scales. This structure captures more internal statistics of the training samples at different scales. The visual quality of SinGAN exceeds that of the other comparison methods. Nevertheless, the HD texture details generated by SinGAN are directly learned from Gaussian noise without a reconstruction process to control the generation of SR images, which produces many artifacts in the final image.

Compared with other methods, Ref-ZSSR not only uses the discriminator to fit the distribution of HD images in the internal statistical training process but also improves downscaling kernel estimation. In addition, with the help of the texture characteristics of the HD reference image, the proposed method can transfer more accurate HR textures from the reference image to generate favorable results, ensuring that the final SR image is more reasonable.

4.4. Quantitative evaluation

We selected 100 images in database CUFED5, 50 images in database BSD100, and all images in data SET14 for the quantitative test. Our model achieves the best performance with the quantitative evaluation results. As shown in Tab. 1, Ref-ZSSR significantly outperforms the other methods on all testing datasets in terms of PSNR and SSIM.

PSNR: PSNR measures the quality of the SR results with the real HD images. The higher the PNSR is, the better the image quality. Although the visual effect of SinGAN is relatively good, the generated images have great randomness due to the lack of mapping from the generated image to the original image. The data in the table show that when the SR experiment is executed at ×2, the PSNR value is relatively satisfactory. When the value increases to 4 times, it decreases significantly.

This happens simply because SinGAN lacks the mapping from the generated SR result to the original input image. It can be proven that the PSNR data are not severely changed in KernelGAN and DBPB, no matter which dataset is used. Similarly, although ZSSR has a higher PSNR than DIP, its performance is inferior to DIP in large-scale resolution reconstruction, especially in the $BSD100$ dataset. Thus, the reconstruction operation ensures that the super-resolution process is more stable and reliable. With the cycle loss, the results of Ref-ZSSR are superior to other methods and achieve better PSNR.

SSIM: SSIM is a well-known metric for measuring the structural similarity between the SR results and the real HD images. As the experimental data in Tab.1 prove, reconstruction can ensure that the image generation process is more stable. However, if only reconstruction is employed, such as DIP, which obtains the results by adjusting the network parameters, the quality of the SR result is not satisfactory. The $SSIM$ test value in the $CUFED5$ and $SET14$ datasets describes this situation. KernelGAN estimates the SR kernel based on the LR image with an internal GAN, which learns the internal distribution of the input image patches with the discriminator. Thus, the $SSIM$ test results are better than DIP, especially in the $BSD100$ dataset, which increased by nearly 5 percentage points whether ×2 or ×4 SR reconstruction.

DBPB and ZSSR have similar principles. ZSSR trains an SR network with an input image and a downscaling kernel. First, the input image is downscaled with the downscaling kernel, and the SR network is trained to reconstruct the input image from the downscaled image. Second, the input image is fed to the SR network to generate an HR image. In this way, ZSSR can also achieve relatively good results, but the ability to improve the clarity of images is limited due to the lack of a discriminator to fit the distribution of real HD images. In most experimental results, its $SSIM$ index is weaker than that of DBPB, since DBPB simultaneously conducts downscaling kernel estimation and SR network training, taking advantage of the joint training framework and the dual back-projection loss. This model uses the advantages of their respective network architecture to improve the quantitative evaluation value of SR. Hence, the $SSIM$ values are relatively high.

Table 1: *Comparison of SR results for benchmark datasets in terms of quantitative evaluation.*

Methods	Scale	PSNR / SSIM (%)		
		CUFED5	SET14	BSD100
DIP	×2	28.28 / 74.15	30.52 / 77.69	29.32 / 75.07
	×4	27.55 / 70.65	29.29 / 76.04	28.02 / 76.78
KernelGAN	×2	30.36 / 82.65	31.78 / 83.45	31.22 / 80.07
	×4	29.01 / 76.45	30.22 / 79.92	29.72 / 79.38
DBPB	×2	30.77 / 81.95	31.55 / 83.25	31.72 / 81.98
	×4	28.91 / 75.30	31.02 / 80.53	30.08 / 80.40
ZSSR	×2	30.05 / 80.37	33.00 / 91.08	31.65 / 89.20
	×4	29.97 / 76.28	28.01 / 76.51	27.12 / 72.11
SinGAN	×2	29.35 / 76.04	30.01 / 76.52	29.03 / 75.29
	×4	28.07 / 70.55	27.92 / 72.11	26.28 / 71.45
Ref-ZSSR	×2	30.98 / 82.04	31.88 / 92.95	32.01 / 89.29
	×4	30.17 / 81.62	30.38 / 85.43	30.55 / 82.95

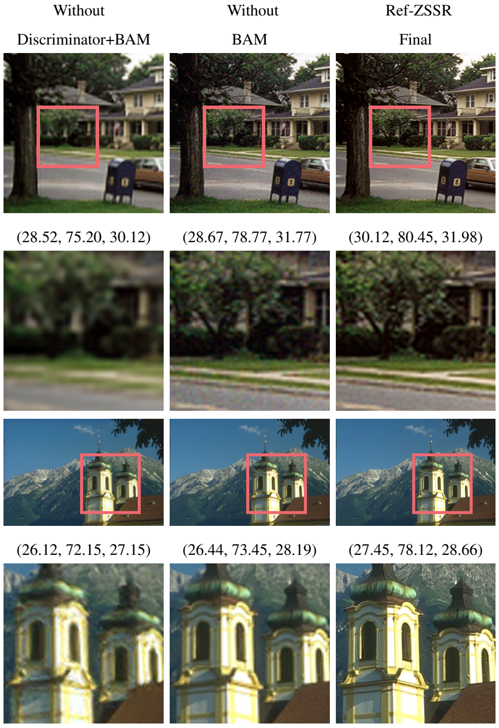

Figure 7: ***4x SR results between different variations of Ref-ZSSR with/without the discriminator or BAM. (PSNR, SSIM[%], NMI[%])***

Ref-ZSSR combines the advantages of all the abovementioned methods. Ref-ZSSR uses not only the discriminator to fit the distribution of HD image patches but also reconstruction loss and cycle loss to ensure the accuracy of the image. In addition, the generator inherits the generation parameters of the HD reference image and has the potential to generate HD textures. The BAM attention mechanism is introduced to further improve the performance of the network. Such a design enables the proposed method to transfer relevant textures from Ref images to LR images and is superior to other image-specific networks. Tab.1 shows that the highest $SSIM$ value gained by Ref-ZSSR proves that this architecture effectively protects the texture details during SR reconstruction on three databases, further indicating that the SR results are the closest to the real HD image.

Table 2: *Average PSNR, SSIM and NMI results with and without the discriminator or BAM by factor of* ×4.

Methods	PSNR, SSIM (%), NMI (%)		
	CUFED5	SET14	BSD100
– (D+BAM)	29.02, 76.60, 27.28	28.02, 80.32, 29.44	27.87, 76.51, 28.45
– BAM	30.05, 80.77, 30.75	30.32, 84.95, 31.22	29.16, 81.78, 29.15
Ref-ZSSR	30.17, 81.62, 31.25	30.38, 85.43, 32.60	30.55, 82.95, 31.98

– (D+BAM) represents the model without the BAM module and the discriminator. – BAM indicates the lack of BAM module.

4.5. Ablation Study

To study the contribution of each block in the Alliance-learning stage for the Ref-ZSSR network architecture, we compare Ref-ZSSR with ablations of the full version. In addition, we add an-

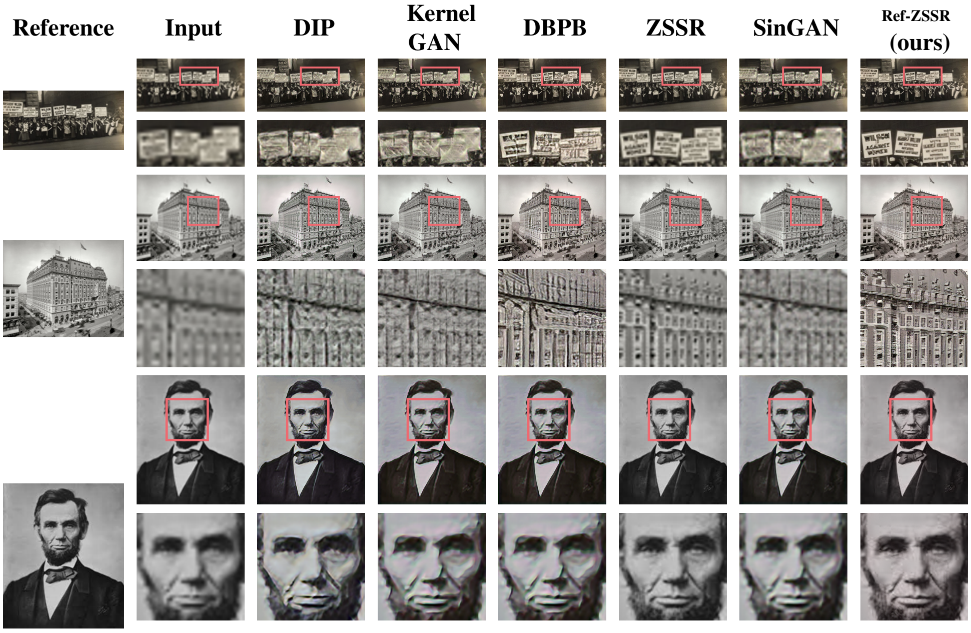

Figure 8: ***Visual comparison among different methods on the image under a nonideal environment.***

other indicator to test the proposed model in addition to PSNR and SSIM. Normalized mutual information (NMI) is an excellent measurement index with which to measure the similarity of the generated images and corresponding ground truths. The higher NMI is, the better the SR result quality.

We first evaluate our method in the absence of the discriminator and BAM. As shown in Tab. 2, without the BAM module and the discriminator, the PSNR, SSIM and NMI are at their lowest. In the Alliance-learning stage, the discriminator tries to distinguish real reference patches from those generated by the generator. The adversarial loss penalizes the distribution of patches in the reference image and the generated SR samples. By fitting the distribution of HD reference images, the generated image has better texture features. Therefore, this module contributes to the improvement of relevant indicators.

In addition, when lacking the BAM module, the PSNR, SSIM and NMI are still inferior to Ref-ZSSR. This module can avoid error accumulation and enhance the feature-extraction ability. Hence, it captures more image details, improves SR performance and finally enhances the relevant image-quality evaluation indicators.

Fig. 7 further illustrates the visual comparison of different structures. If the model lacks the discriminator and BAM, the SR result has the lowest image quality compared to the other two methods. Next, we remove the BAM and test the visual effect. This model suffers from artifacts in the output image. In both cases, the results are not as sharp as the proposed full Ref-ZSSR output. Having the adversarial loss and plugging the BAM module makes the Ref-ZSSR network capable of generating realistic natural images without unwanted artifacts.

4.6. The SR results of the image under a nonideal environment and the image itself as the reference image

To further verify the performance of the proposed SR model, a nonideal case, that is, poor-quality LR images with unknown degradation, is conducted. The purpose of this experiment is to test more realistic blur kernels. We randomly selected 20 LR images from the test database and downscaled these images by using random Gaussian kernels. In addition, due to the different results in searching for different reference images toward these poor-quality images, in this experiment, we use the image itself (LR) as the reference image.

Table 3: *The SR comparision results under non-ideal environment by factor of* $\times 2$.

	Methods					
Quantity	DIP	KernelGAN	DBPB	ZSSR	SinGAN	Ref-ZSSR
PSNR	21.02	23.23	24.03	23.76	22.02	24.56
SSIM	61.24	65.75	64.70	61.24	60.46	65.98
NMI	22.45	24.96	25.75	22.23	21.72	26.01

Tab.3 shows the results comparison between Ref-ZSSR and 5 peer methods in terms of 3 metrics, namely, PSNR, SSIM and NMI. Although the specific degraded kernels are unknown, Ref-ZSSR has the best metrics. As seen from the table, compared to other methods, SinGAN and ZSSR do not reconstruct the original image and yield inferior metrics, which proves that the reconstruction contributes to enhancing the SR performance, especially for low-quality images.

Without the guidance of the HD reference image, the results of Ref-ZSSR are still superior to DIP, KernelGAN and DBPB. This is because Ref-ZSSR has a self-learning process, which learns the

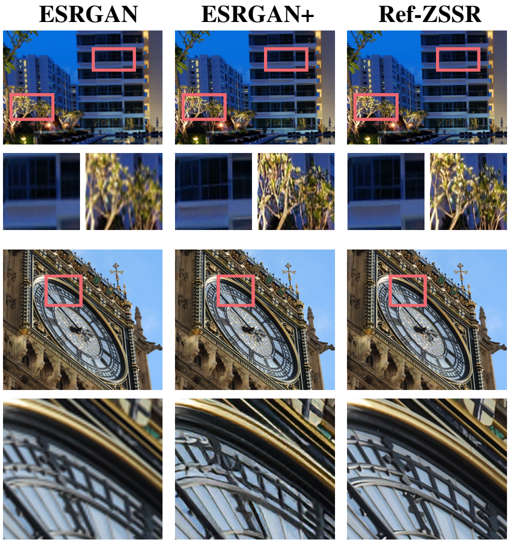

Figure 9: ***The effectiveness verification of Ref-ZSSR in terms of visual effects.*** *The SR images generated by ESRGAN are used as reference images. We employ the self-learning and alliance-learning of Ref-ZSSR to improve the visual performance.*

image's patch statistics and makes reasonable inferences in the subsequent process SR. This experiment also demonstrated the applicability of the proposed model even without the HD reference image.

The visualization results of some samples are presented in Fig. 8. As shown, under a nonideal environment, Ref-ZSSR has satisfactory visual effects using only the image information. Regardless of the characters in the image or the textures of the house, even facial wrinkles, the SR results of the proposed model can be clearly displayed. Compared with other methods, three-stage learning enables the network to better capture the internal statistical features of the input image, indicating that Ref-ZSSR has a better capacity to upsample poor-quality images.

4.7. The effectiveness verification of the proposed architecture

To confirm the effectiveness of the proposed framework, we use some excellent SR networks as the prior information provider. This means that the SR image generated by these networks is used as the reference image. Hence, the outstanding method ESRGAN [WXW19] is selected as G_1 in Ref-ZSSR, and its advanced edition ESRGAN+ [RR20] is selected for comparison.

We randomly select some images from the database for testing. After feeding into the two excellent frames, the generated SR images achieve very high values. As shown in Tab. 4, ESRGAN introduces the residual-in-residual dense block and allows the discriminator to predict relative realness instead of the absolute value. Thus, these values in ESRGAN are higher than those methods in Tab. 1. ESRGAN+ is extended to further improve the perceptual quality of the images. Hence, the image quality generated with ESRGAN+ achieves better performance. Additionally, we use ESRGAN as G_1 to generate the reference image and then perform the self-learning and alliance-learning of Ref-ZSSR. It can be seen that the proposed method is on par with the advanced version in terms of these indicators. Even on the SET14 database, the SR results of Ref-ZSSR completely surpass those of ESRGAN+. This implies the effectiveness of our framework and has the potential to further improve the performance of existing SR models.

Table 4: *The effectiveness verification of Ref-ZSSR in terms of PSNR, SSIM and NMI by factor of* $\times 2$.

Methods	PSNR, SSIM (%), NMI (%)		
	CUFED5	SET14	BSD100
ESRGAN	31.32, 77.80, 32.55	29.01, 81.14, 30.05	28.01, 77.45, 29.88
ESRGAN+	31.77, 81.77, 32.99	30.22, 84.99, 31.40	30.32, 82.07, 31.61
Ref-ZSSR	31.72, 82.06, 33.87	30.87, 85.60, 32.44	30.06, 81.21, 30.20

Fig.9 displays the visual comparison results. As shown, the visual effects of ESRGAN+ are superior to those of ESRGAN. Since it designs a novel block to replace the one used by the original ESRGAN and introduces noise inputs to the generator network to exploit stochastic variation, the resulting images present more realistic textures. Meanwhile, we also find that, compared to ESRGAN, more detailed textures are displayed in Ref-ZSSR. The proposed method takes advantage of self-learning to learn more input image internal statistical characteristics. In addition, alliance-learning further improves the visual performance by fine-tuning the network parameters of ESRGAN. These operations make the final generated SR image have a visual perception similar to ESRGAN+.

Table 5: *The influence of the similarity between the reference image and the input image on the SR result.*

PSNR	SSIM (%)			
	≤ 20	$20 \sim 40$	$40 \sim 60$	≥ 60
increment	≤ 0.85	$0.85 \sim 1.2$	$1.2 \sim 1.5$	≥ 1.5

4.8. The SR results with different reference images

We also explored the influence of the reference images on the final SR results. If the reference image is utterly distant from the input image, these extracted features in the Ref-learning stage are difficult to guide the G_2 to generate HR textures in the Alliance learning stage, which leads to the SR results being similar to KernelGAN [BKSI19].

According to TTSR [YYF*20], the recommended approach is to select an image from different perspectives of the same scene as the reference image. As shown in Fig.10, Ref_2 is of this type and is superior to Ref_1 with similar content in terms of visual effect and numerical indices. Ref_3 has a closer perspective to the input image, which brings the SR result the best performance. This is because the closer view angle provides more surrounding pixel information. As a powerful prior, these statistics assist the final image to be well filled.

In addition, we randomly selected several images to quantitatively analyze the influence of the similarity between the reference

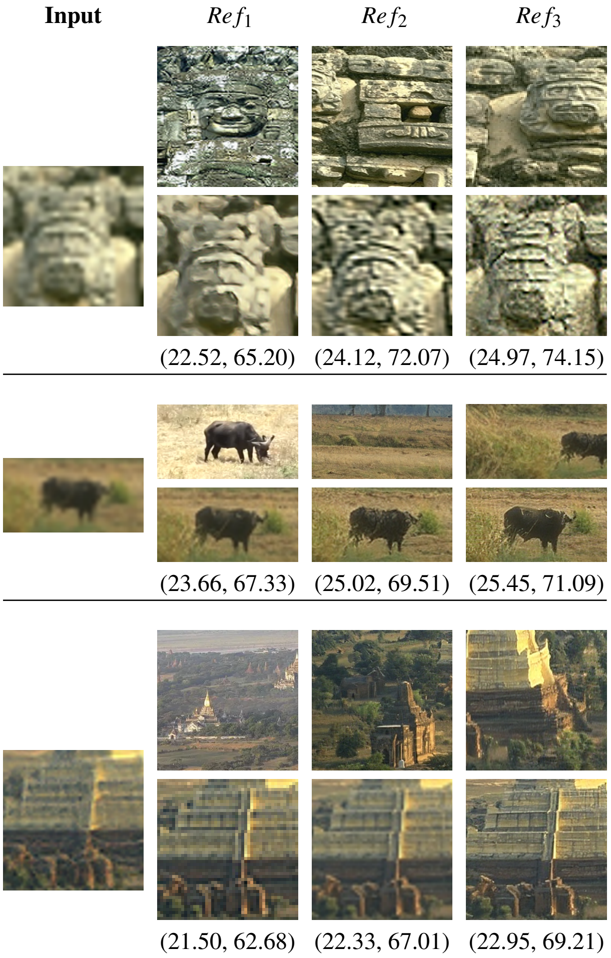

Figure 10: ***Visual comparison among different reference images.*** *Ref_1, Ref_2 and Ref_3 represent different reference images, respectively. These numbers represent PSNR and SSIM[%].*

image and the original image on the SR results. As shown in Tab.5, we used *SSIM* to measure similarity. When the value is lower than 20%, the improvement of *PSNR* is quite limited. With the growing similarity, the *PSNR* continues to increase accordingly. If the similarity exceeds 60%, more similar areas have an obvious effect on improving *PSNR*.

5. Conclusion

We proposed Ref-ZSSR, an image-specific SISR network architecture, which performs SR by referring to the texture of a single HD image. Ref-ZSSR consists of the Ref-learning stage, Self-learning stage, and Alliance-learning stage. First, we use a pyramid of fully convolutional GANs to produce the HD texture of the reference in the Ref-learning stage. Second, a dual-path architecture that includes a downsampler and an upsampler is introduced to learn the degradation process and superresolved process, which are trained simultaneously and improved using cycle-consistency losses. Finally, we combine the reference-image learning module and dual-path architecture module to train a new GAN model with a BAM to generate an SR image with the details of the HR reference image. Such a design encourages a simple and accurate way to transfer relevant textures from Ref images to LR images. The SR results outperform previous image-specific SISR methods. Our future work will aim to extend the architecture of Ref-ZSSR by designing a more effective network structure.

References

[AD20] Alexey Dosovitskiy Lucas Beyer A. K.: An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR abs/2010.11929* (2020). URL: https://arxiv.org/abs/2010.11929, arXiv:2010.11929.

[BKSI19] Bell-Kligler S., Shocher A., Irani M.: Blind super-resolution kernel estimation using an internal-gan. In *Advances in Neural Information Processing Systems* (2019), Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., (Eds.), vol. 32, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/5fd0b37cd7dbbb00f97ba6ce92bf5add-Paper.pdf. 1, 2, 3, 5, 6, 10

[BWKN21] Bashir S. M. A., Wang Y., Khan M., Niu Y.: A comprehensive review of deep learning- based single image super-resolution. *PeerJ Computer Science 7* (07 2021), e621. doi:10.7717/peerj-cs.621. 1

[CHQ*22] Chen H., He X., Qing L., Wu Y., Ren C., Zhu C.: Real-world single image super-resolution: A brief review. *Information Fusion 79* (2022), 124–145.

[CHT20] Chen Hanting W. Y., Tianyu G.: Pre-trained image processing transformer. *arXiv e-prints* (Dec. 2020). 2

[CMS*20] Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S.: End-to-End Object Detection with Transformers. *arXiv e-prints* (May 2020), arXiv:2005.12872. arXiv:2005.12872.

[DCL14] Dong Chao Loy Chen Change H. K.: Learning a deep convolutional network for image super-resolution. In *Computer Vision – ECCV 2014* (Cham, 2014), Springer International Publishing, pp. 184–199. 2

[EPC21] Emad M., Peemen M., Corporaal H.: Dualsr: Zero-shot dual learning for real-world super-resolution. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2021), pp. 1629–1638. doi:10.1109/WACV48630.2021.00167. 1, 2, 5

[HLWG21] 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, CVPR 2021, virtual, June 19-25, 2021* (2021), Computer Vision Foundation / IEEE, pp. 2093–2102. 3

[HMZ19] Hu X., Mu H., Zhang: Meta-sr: A magnification-arbitrary network for super-resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 1575–1584. doi:10.1109/CVPR.2019.00167. 2

[HSU20] Haris M., Shakhnarovich G., Ukita N.: Deep back-projection networks for single image super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. doi:10.1109/TPAMI.2020.3002836. 2

[IS15] Ioffe S., Szegedy C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37* (2015), ICML'15, JMLR.org, p. 448–456. 4

[JW21] Jiaming Wang Zhenfeng Shao X. H.: Enhanced image prior for unsupervised remoting sensing super-resolution. *Neural Networks 143* (2021), 400–412. 1

[KBN*20] Köhler T., Bätz M., Naderi F., Kaup A., Maier A., Riess C.: Toward bridging the simulated-to-real gap: Benchmarking super-resolution on real data. *IEEE Transactions on Pattern Analysis and Machine Intelligence 42*, 11 (2020), 2944–2959. doi:10.1109/TPAMI.2019.2917037. 1

[KJK20] Kim J., Jung C., Kim C.: Dual back-projection-based internal learning for blind super-resolution. *IEEE Signal Processing Letters 27* (2020), 1190–1194. 1, 2, 5, 6

[KSK20] Kim S. Y., Sim H., Kim M.: KOALAnet: Blind Super-Resolution using Kernel-Oriented Adaptive Local Adjustment. *arXiv e-prints* (Dec. 2020), arXiv:2012.08103. arXiv:2012.08103. 3

[LDT20] Lugmayr A., Danelljan M., Timofte R.: Unsupervised learning for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (2020). 1

[LKLE21] Li B., Keikhosravi A., Loeffler A. G., Eliceiri K. W.: Single image super-resolution for whole slide image using convolutional neural networks and self-supervised color normalization. *Medical Image Anal. 68* (2021), 101938. URL: https://doi.org/10.1016/j.media.2020.101938, doi:10.1016/j.media.2020.101938. 2

[LLG*21] Liu A., Liu Y., Gu J., Qiao Y., Dong C.: Blind image super-resolution: A survey and beyond, 2021. arXiv:2107.03055. 2

[LSK*17] Lim B., Son S., Kim H., Nah S., Lee K. M.: Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2017). 2

[LTH*17] Ledig C., Theis L., Huszár F., Caballero J., Cunningham A., Acosta A., Aitken A., Tejani A., Totz J., Wang Z., Shi W.: Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 105–114. doi:10.1109/CVPR.2017.19. 1

[LW21] Longguang Wang Yingqian Wang X. D.: Unsupervised degradation representation learning for blind super-resolution, 2021. arXiv:2104.00416. 2

[MFTM01] Martin D., Fowlkes C., Tal D., Malik J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001* (2001), vol. 2, pp. 416–423 vol.2. doi:10.1109/ICCV.2001.937655. 6

[PW20] Pengxu Wei Ziwei Xie H. L.: Component divide-and-conquer for real-world image super-resolution. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VIII* (2020), vol. 12353 of *Lecture Notes in Computer Science*, Springer, pp. 101–117. URL: https://doi.org/10.1007/978-3-030-58598-3_7, doi:10.1007/978-3-030-58598-3_7. 1

[PZD*21] Pan X., Zhan X., Dai B., Lin D., Loy C. C., Luo P.: Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1. doi:10.1109/TPAMI.2021.3115428. 2, 3

[RR20] Rakotonirina N. C., Rasoanaivo A.: Esrgan+ : Further improving enhanced super-resolution generative adversarial network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), pp. 3637–3641. doi:10.1109/ICASSP40776.2020.9054071. 10

[SBII19] Shocher A., Bagon S., Isola P., Irani M.: Ingan: Capturing and retargeting the "dna" of a natural image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 4491–4500. doi:10.1109/ICCV.2019.00459. 1

[SCI18] Shocher A., Cohen N., Irani M.: Zero-shot super-resolution using deep internal learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 3118–3126. doi:10.1109/CVPR.2018.00329. 1, 2, 3, 6

[SDM19] Shaham T. R., Dekel T., Michaeli T.: Singan: Learning a generative model from a single natural image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 4569–4579. doi:10.1109/ICCV.2019.00467. 1, 4, 6

[SH12] Sun L., Hays J.: Super-resolution from internet-scale scene matching. In *2012 IEEE International Conference on Computational Photography (ICCP)* (Los Alamitos, CA, USA, apr 2012), IEEE Computer Society, pp. 1–12. URL: https://doi.ieeecomputersociety.org/10.1109/ICCPhot.2012.6215221, doi:10.1109/ICCPhot.2012.6215221. 2

[TSG13] Timofte R., Smet V. D., Gool L. V.: Anchored neighborhood regression for fast example-based super-resolution. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013* (2013), IEEE Computer Society, pp. 1920–1927. URL: https://doi.org/10.1109/ICCV.2013.241, doi:10.1109/ICCV.2013.241. 2

[UVL20] Ulyanov D., Vedaldi A., Lempitsky V.: Deep image prior. *International Journal of Computer Vision 128*, 7 (2020), 1867–1888. doi:10.1007/s11263-020-01303-4. 2, 3, 6

[WHS21] Wang F., Hu H., Shen C.: BAM: A Balanced Attention Mechanism for Single Image Super Resolution. *arXiv e-prints* (Apr. 2021), arXiv:2104.07566. arXiv:2104.07566. 5, 6

[WLHD17] Wang Y., Liu Y., Heidrich W., Dai Q.: The light field attachment: Turning a dslr into a light field camera using a low budget camera ring. *IEEE Transactions on Visualization and Computer Graphics 23*, 10 (2017), 2357–2364. doi:10.1109/TVCG.2016.2628743. 2, 3

[WLL*21] Wang Z., Lu Y., Li W., Wang S., Wang X., Chen X.: Single image super-resolution with attention-based densely connected module. *Neurocomputing 453* (2021), 876–884. URL: https://doi.org/10.1016/j.neucom.2020.08.070, doi:10.1016/j.neucom.2020.08.070. 2

[WLX*21] Wei Y., Liu H., Xie T., Ke Q., Guo Y.: Spatial-temporal transformer for 3d point cloud sequences, 2021. arXiv:2110.09783. 2

[WXW19] Wang Xintao Y. K., Wu S.: Esrgan: Enhanced super-resolution generative adversarial networks. In *Computer Vision – ECCV 2018 Workshops* (ham, 2019), Springer International Publishing, pp. 63–79. 10

[XSGW21] Xiong C., Shi X., Gao Z., Wang G.: Attention augmented multi-scale network for single image super-resolution. *Appl. Intell. 51*, 2 (2021), 935–951. URL: https://doi.org/10.1007/s10489-020-01869-z, doi:10.1007/s10489-020-01869-z. 2

[YFL21] Ying Fu Jian Chen T. Z., Lin Y.: Residual scale attention network for arbitrary scale image super-resolution. *Neurocomputing 427* (2021), 201–211. 2

[YYF*20] Yang F., Yang H., Fu J., Lu H., Guo B.: Learning texture transformer network for image super-resolution, 2020. arXiv:2006.04139. 2, 3, 10

[ZEP10] Zeyde R., Elad M., Protter M.: On single image scale-up using sparse-representations. In *International conference on curves and surfaces* (2010), Springer, pp. 711–730. 6

[ZJW*18] Zheng H., Ji M., Wang H., Liu Y., Fang L.: Crossnet: An end-to-end reference-based super resolution network using cross-scale warping, 2018. arXiv:1807.10547. 2, 3

[ZLZ21] Zhisheng Lu Hong Liu J. L., Zhang L.: Efficient transformer for single image super-resolution. *CoRR abs/2108.11084* (2021). URL: https://arxiv.org/abs/2108.11084, arXiv:2108.11084. 2

[ZWLQ19] Zhang Z., Wang Z., Lin Z., Qi H.: Image super-resolution by neural texture transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 7974–7983. doi:10.1109/CVPR.2019.00817. 2, 3, 6