International Journal of Computational Intelligence Systems
Vol. **14(1)**, 2021, *pp.* 1577–1588
DOI: https://doi.org/10.2991/ijcis.d.210503.002; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

Research Article

A Single Historical Painting Super-Resolution via a Reference-Based Zero-Shot Network

Hongzhen Shi[1,2,*], Dan Xu[1,*], Hao Zhang[1], YingYing Yue[1]

[1]*School of Information Science and Engineering, Yunnan University, Kunming, 650500, China*
[2]*School of Electric and Informative Engineering, Yunnan Minzu University, Kunming, 650500, China*

ARTICLE INFO

Article History
Received 16 Oct 2020
Accepted 20 Apr 2021

Keywords
Historical paintings
Super-resolution Zero-shot Deep learning

ABSTRACT

As an important part of human cultural heritage, many ancient paintings have suffered from various deteriorations that have led to texture blurring, color fading, etc. Single image super-resolution (SISR) which aims to recover a high-resolution (HR) version from a low-resolution (LR) input is actively engaged in the digital preservation of cultural relics. Currently, only traditional super-resolution is widely studied and used in cultural heritage, and it is difficult to apply learning-based SISR to unique historical paintings because of the absence of both ground truth and datasets. Fortunately, the recently proposed ZSSR method suggests that it is feasible to generate a small dataset and extract self-supervised information from a single image. However, when applied to the preservations of historical paintings, the performance of ZSSR is highly limited due to the lack of image knowledge. To address the above issues and to unleash the great potential of learning-based methods in heritage conservation, we present Ref-ZSSR, which is the first attempt to combine zero-shot and reference-based methods to achieve SISR. In our model, both global and local multi-scale similar information is fully exploited from the painting itself. In an end-to-end manner, this information provides consistent artistic style image knowledge and helps synthesize SR images with sharp texture details. Compared with the ZSSR method, our approach shows improvement in both precision (approximately 4.68 dB for scale ×2) and visual perception. It is worth mentioning that all image knowledge required in our method can be extracted from the painting itself, i.e., external examples are not required. Therefore, this approach can be easily generalized to any damaged historical paintings, broken murals, noisy old photos, incomplete art works, etc.

1. INTRODUCTION

Throughout history, historical paintings have been susceptible to multiple degradation processes due to natural decay, improper preservation, willful sabotage, etc. Such degradations can manifest themselves as, e.g, color degradation, texture blurring and even the absence of large areas of the painting. Furthermore, the degradation inevitably becomes worse with the passage of time. However, since these artworks are normally of a fragile nature, physical restoration is often impractical or too risky to implement. Instead, restoration of paintings are normally performed by scanning or photocopying them into digital images, and their damaged or lost parts can then be renewed via image processing and computer vision. These digital protection methods are effective not only in removing the existing destruction without causing secondary damage but also in protecting the paintings from future degeneration. One of the most popular of these methods is single image super-resolution (SISR), which reproduces high-resolution (HR) paintings with rich and clear texture details from a rough low-resolution (LR) image. SISR can infer the lost and blurred texture and oxidized and shed color from several potential spots. In other words, appropriate SISR is expected to renew relic paintings to as close to their original state as possible. Furthermore, research on SISR is important for not only artwork restoration but also computer vision.

Current SISR methods can be broadly classified into three categories: interpolation-based, model-based optimization and learning-based methods. The interpolation-based methods, such as the nearest neighbor interpolation [1,2], bilinear interpolation [3,4] and bicubic interpolation [5,6], are simple and fast to implement. However, because such approaches only consider the degradation of downsampling and ignore the decay caused by other factors, their performance is often limited. Model-based optimization methods can flexibly achieve super-resolution reconstruction either by employing Hopfield neural networks [7] or by using global self-similar information [8], sparse priors [9], denoising priors [10], etc. However, these methods are time-consuming and require many manual priors, and they are not an end-to-end learning strategy. Since Dong [11] successfully applied convolutional neural networks (CNNs) to image super-resolution, learning-based SISR methods, which have attracted increasing interest recently, have been greatly improved in terms of effectiveness and efficiency [12–14]. In comparison, the alternative interpolation based and model-based optimization methods usually show poor

*Corresponding authors. Email: *yingbinggan@126.com*

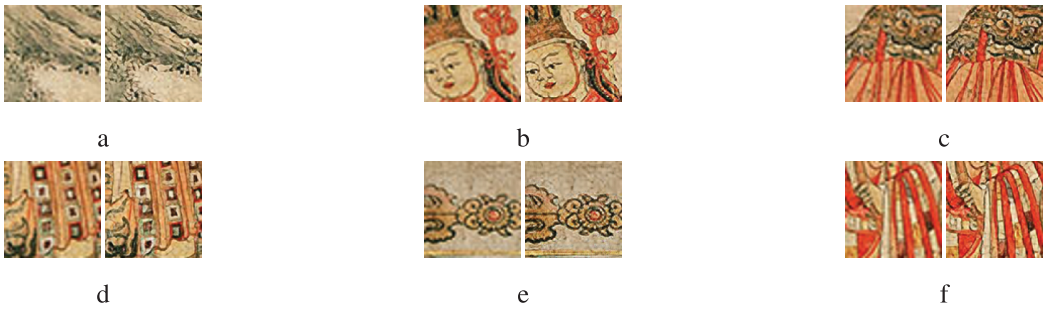

Figure 1 | Input paintings and SR results

Figure 2 | Input image and reference images of different spaces.

performance in super-resolution with large up-scaling factors. Therefore, in this work, we will restrict our investigation to the super-resolution of relic paintings using a learning-based approach.

For natural images, learning-based SISR has made great progress in model performance, network design and training strategies [15–20], which usually require public datasets and a huge amount of prior training. Unfortunately, the collection of thousands of HR-LR image pairs is still very challenging, particularly for relic paintings. Because the majority of famous paintings are unique works rather than mass-produced, the number of obtainable images is often far from adequate for deep learning. For the ground truth, since it is impossible to travel back in time, the original form of the existing relic paintings will remain an unsolvable mystery. It is almost impossible to acquire HR paintings. Therefore, generating a dataset with many images based on a single image and finding effective supervision information based on the existing LR paintings are two of the pressing issues for historical paintings. Traditional algorithms usually apply the self-example methods to realize unsupervised (self-supervised) SISR using redundant information of the image itself. Examples of this formalism can be found in Refs. [21,22], where the authors achieved self-supervised SISR by specifying the block size (usually 5×5) and then used Euclidean distance and k-nearest neighbor searches to find similar blocks. However, these methods are often inefficient in finding repeating structures of different sizes, making it difficult to generalize them to the implicit similarity measure learned by deep learning networks. Therefore, it is not feasible to simply copy the above ideas in learning-based SISR for paintings.

Fortunately, ZSSR [23] can effectively overcome the above deficiencies by training an image-specific CNN and leveraging the power of cross-scale recursive information inside a single image. Our work draws on the ZSSR [23] to generate thousands of HR-LR relations from an LR image and its downscaled versions. Different from ZSSR, in our work the HR-LR image pairs are no longer different downscaled painting copies but rather a large number of HR-LR block pairs cropped from downsampled versions. This operation not only increases the number of images in the dataset but also allows for a more detailed treatment for the subregions. We can train a painting-specific CNN and then apply the learned relations to the original LR painting to produce an SR image. However, due to the lack of image knowledge, ZSSR is not effective when applied for ancient paintings super-resolution. A feasible solution to obtain more prior knowledge and improve the ZSSR performance is the reference-based method, which uses HR reference images with rich texture details to compensate for the lost details in the LR input image. Using the high-frequency information provided in the reference image, reference-based super-resolution usually achieves better performance than that with one single input. Therefore, in this

work we make the first attempt to combine reference-based super-resolution with zero-shot learning and propose a reference-based zero-shot super-resolution model (Ref-ZSSR) to improve the super-resolution performance for relic paintings. It should be noted that the use of reference images acquired from an external database is not the ideal choice for ancient paintings. Since the artworks in question are normally unique in style, finding reference images of similar artistic styles from a massive external database can be very time-consuming and fruitless. In fact, rigorous external images not only can turn out to be useless but also are very likely to contaminate the original style. On the other hand, a large number of recurring crops (subregions) with different scales have been proven to exist in natural images [24], and one can expect that this is also true for ancient paintings. Some previous studies [8,21,25–27] have shown that global similarities can be used to effectively improve model performance. Therefore, we will use global similar blocks from the painting itself as the reference images to achieve super-resolution. This strategy can effectively extract repetitive and similar global information while ensuring that the reference image and the input image correspond to the same author and style. The same root high-frequency information extracted in this manner can be expected to provide a faithful representation for the missing textures. To extract more image information, we also introduce correlation discovery modules [28] in our model to explore local similar structures, the information of which is then fused with that for global similarities at multiple scales. In short, without relying on external images, Ref-ZSSR is expected to extract as much image information as possible to improve both visual perception and precision. Actually, the painting-specific dataset and reference-based CNN proposed in this paper are easily generalized to other low-level vision tasks, such as image de-raining, de-hazing, de-blurring, denoising, etc. The contribution of our work presented in this paper can be summarized as follows:

I. Drawing on the zero-shot approach, we obtained HR-LR relations from a single LR painting and its down-scaled versions and successfully applied learning-based SISR to a unique painting.

II. This work is the first attempt to combine a reference-based method and zero-shot super-resolution, where global similarity blocks serving as the reference images largely compensate for the scarce prior knowledge in ZSSR.

III. An end-to-end model is designed to fully explore global and local similar information in order to achieve super-resolution.

2. RELATED WORK

2.1. Learning-Based SR

With the help of a large number of HR-LR training samples provided in public datasets, learning-based SR has made remarkable progress. Since the advent of SRCNN [11], which has a three-layer structure, network architectures are constantly being improved by increasing depth or sharing weights. VDSR [29], which is the first work to introduce residual learning into super-resolution to accelerate the convergence in a deep network, later extended the network to 20 layers. DRCN [12] is the first method to introduce recursive learning in SISR and reuse the features of each layer in a CNN. By increasing the dense connection between the convolutional layers, DenseNet [30] contributed to reducing the gradient disappearance, enhancing feature propagation and minimizing the number of parameters. EDSR [13] removed unnecessary modules in the traditional remaining network. The residual in the residual (RIR) structure was proposed to construct very deep trainable networks in RCAN and [14] introduced a channel attention mechanism to find interdependencies among feature channels. Apart from improving the network structures, SROBB [31] further confirms that the choice of training strategies can affect the performance of optimization-based methods.

Whereas the majority of works on learning-based SR involves the pixel distance between the SR and ground truth, it has been confirmed that a high PSNR is not necessarily consistent with human perception. Therefore, images produced by methods that focus exclusively on the PSNR tend to be excessively smooth and lacking in high-frequency textures. Based on the idea of perceptual similarity [32] and perceptual loss [33], pretrained feature extraction models, such as VGG [34], can be used to extract the features of a certain layer or layers. By minimizing the error in the feature space, the perceptual-driven methods can remarkably improve the visual quality. This optimization has been used in some works to generate images that rely on high-level features [35–37]. SRGAN [38] proposed a perceptual loss function composed of adversarial loss and content loss based on perceptual similarity to replace pixel-level similarity. This deep residual network can recover realistic texture details from greatly downsampled images. However, it should be noted that although the CNN-based SISR method can achieve good performance in terms of network structure, training strategy and objective function, all the aforementioned approaches require LR-HR pairs from open datasets. Due to the lack of a ground truth, it is difficult to simply apply the above methods to relic paintings.

2.2. Zero-Shot

The application of zero-shot in super-resolution, i.e, the ZSSR prescription [23], is among the most widely used models in super-resolution and has gained increasing interest recently. In addition, the majority of zero-shot methods that have provided a large number of excellent results in recent years are mainly based on segmentation [39], emotion recognition [40], object detection [41], image retrieval [42–45], image classification [46–48] and intelligent learning in machines or robots [49]. In the ZSSR formalism, LR images are downsampled to generate many lower-resolution images ($I = I_0, I_1, I_2, ..., I_n$), which serve as the HR supervision information called "HR fathers," then, each HR father is downscaled by the required scale factor s to obtain the corresponding "LR sons." The image-specific CNN can be trained on the LR-HR dataset, and after training convergence, the model produces a SR image from the input LR image. In short, ZSSR trains the image-specific CNN with the LR-HR pairs generated by a single image, eliminating the need for any external examples and training priors. ZSSR is the first unsupervised (self-supervised) SISR based on a CNN. However, since this model is mainly focused on the overall visual performance, treatment of the subregions is still lacking in terms of texture details. The recently developed MZSR [50] model utilizes transfer learning to implement model pretraining on an external dataset in order to obtain general initial parameters. Compared

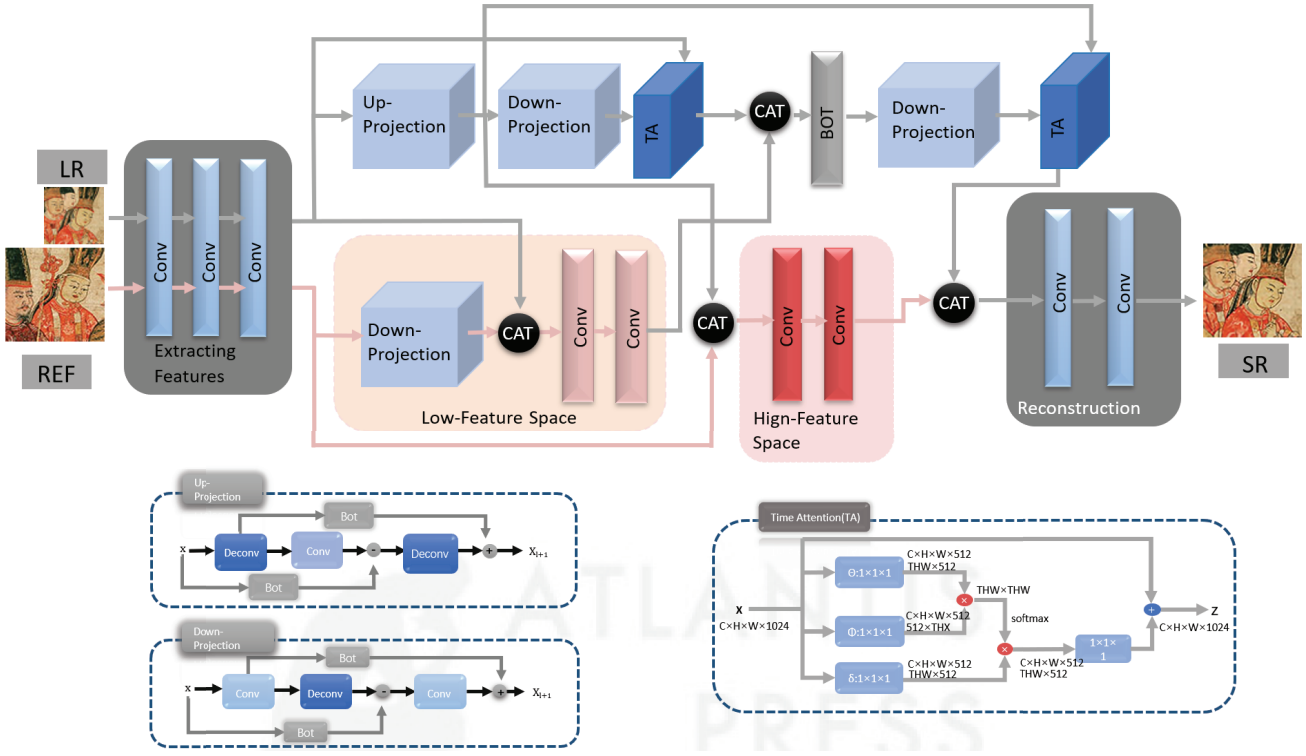

Figure 3 Network structure of our proposed Ref ZSSR. In contrast to the CNN-based SISR methods that only take the LR image as input, Ref-ZSSR provides both an HR reference image and an LR input image to the network to enrich the high-frequency texture. Low and high global similar features are extracted from the reference image by low and high feature space blocks, while the local similar feature is extracted by the time attention (TA) modules. Enhanced up (down)-sampling back-projection blocks are employed to facilitate up- and down-sampling.

with ZSSR [23], MZSR accelerates the super-resolution process for single image by reducing the number of gradient updates. Although MZSR compensates for the lack of training prior and significantly reduces the training time, the effectiveness of the training prior depends on the correlation between the external examples and the target image. For relic paintings, an external dataset with strong correlations is still difficult to obtain. In addition, detailed texture reproduction is usually more important than a reduction in the training time. Therefore, in this work, we extract LR-HR pairs of different scales and partitions from one representative artwork, such that the model can reproduce lost texture details in small areas without compromising the overall visual perception. In addition, to meet the delicate repair needs of precious paintings, global similar blocks from the painting itself are introduced as reference images to fully extract image prior knowledge.

2.3. Reference-Based Super-Resolution

The Ref-SR model, first proposed by Freeman *et al.* [22], incorporated the idea of using more than one LR image as input. Compared to the traditional SISR, which only uses one LR image as the input, the reference-based SISR model uses both an HR reference image and an LR input image, thereby compensating for the missing texture in the LR input by using the high-frequency texture from the HR reference. These reference images can be obtained from an external database [51], web image searches [52] and video materials [38,53]. However, in cases where the high-frequency information provided in external reference pictures is not well aligned to the LR images, the performance of the Ref-SR model is very likely to be compromised. To alleviate this problem, attempts have been made in order to produce more detailed and realistic textures while allowing for input and reference images to be irrelevant [14,54]. This approach involves using global scene descriptors to find similar scenes from the internet-scale image database, which are expected to provide ideal example textures that can be used to alleviate the image up-sampling issues. For example, Ref. [52] shows that the performance of example-based SR can be improved by standardizing different illuminations, focal lengths and lens angles of the image materials retrieved from the internet. These methods can lower the requirements for external reference images, but finding reference images from external data is still a very time-consuming task. To solve this problem, a possible solution is to select the reference image from the input image itself [55,56]. Specifically, geometric variations can be used to expand the internal patch search space, detected perspective geometry can then be utilized to guide the patch search process and find repeated statistical priors. In self-similarity-based super-resolution [56], an encoder is used to extract and combine multi-scale feature maps from both the LR and reference images, thereby producing SR images. Rather than matching content in the raw pixel space, these methods mainly use internal similar information to fuse the HR patches at pixel level. The model proposed in Ref. [57] can significantly enrich the SR details by

Figure 4 | Visual comparison of different reference images for image super-resolution.

adaptively transferring the texture from the reference images according to their textural similarity. Considering the pros and cons of the above methods, we intend to search global similar blocks from the painting itself as the reference images.

3. APPROACH

Here, we introduce the Ref-ZSSR model aimed at producing SR images with rich and authentic details from an LR painting. To apply learning-based SISR to digital preservation of relic paintings, we adopted zero-shot and data augmentation technologies to generate a dataset from a single painting. To improve the performance of ZSSR [23], we combine VGGLOSS [58] and a pretrained VGG-19 [34] to search global semantic similarity blocks as reference images. An overview of this model is shown in Figure 1. This section is organized as follows: we discuss the zero-shot-based painting-specific dataset in Section 3.1 and review global similar reference images in Section 3.2; Section 3.3 describes the structure of Ref-ZSSR.

3.1. Zero-Shot-Based Painting-Specific Dataset

In the absence of any external supervision information, zero-shot uses the low entropy of the image internal information to build a painting-specific CNN, which is trained on the examples extracted from the painting itself and is easy to extend to any painting. The subject painting I_0 with original size W × H is downscaled according to the required scale factor s to obtain I_1, which is then downsampled by the same factor to obtain I_2. By analogy, after n downsampling iterations, we obtain the set I = (I_0, I_1, I_2, ..., I_n), where $I_n$is the smallest downsampled version, which has a size of w × h given by

$$w \times h = \frac{W \times H}{s^n} \tag{1}$$

Similar to Huang [55], we employ vertical and horizontal directions flip and rotations at different angles to enlarge the dataset by a factor of 8 to obtain an enhanced image set I_E, which serves as the HR supervision and is referred to as "HR fathers" in the subsequent texts. All the "HR fathers" images are downsampled again with the same sampling factor s to generate $I_{i\downarrow}$s, called "LR sons." Note that due to the limited size of the original painting subject and the desired downscale factor s, only a total of 8n "HR-LR" image pairs can be generated. Clearly, a large gap is still present between the CNN's data requirements and 8n image pairs. Many works [11,59–61] have shown that an increase in the number of training data can improve the model performance, which requires further data augmentation. We apply a sliding window to the whole enhanced image set $I_E$and slice all elements in this set into the same size (w × h) as that of I_n, thereby producing an augmented dataset. Note that a certain overlap is maintained between the adjacent slices during this process. Based on the obtained dataset, we can train a small painting-specific CNN that explores more potential information to compensate for the lost texture in the subject painting. The main advantages of the combination of zero-shot, data augmentation and sliding window can be summarized as follows: I. the number of images grows exponentially and the dataset scale can be moderately controlled by changing the overlapping ratio; II. all crops share a uniform size of w × h; III. subregions (denoted as crops) can now be dealt with in greater detail.

3.2. Global Similar Blocks

To achieve good results, we intend to introduce HR reference images to enrich prior knowledge. According to Ref. [24], the internal entropy of the patches from the same image is much smaller than the external entropy of the patches from general images. It is further confirmed that the internal statistics usually have stronger predictive power than the external statistics. Therefore, we use global similar blocks from the painting itself as the reference images.

In terms of the measure of similarity, some works [62–64] showed that pixel-level similarity criteria, such as the PSNR, are better suited for identifying pixel-level image differences rather than the high-frequency texture details. While human eyes tend to perceive two images offset from each other by one pixel as identical, the images can be very different in pixel space. Therefore, in this work, global feature similar blocks are measured using the VGGLOSS [38]. Then, we define the VGGloss as the Euclidean distance between the feature representations of an input image $I^{IN}$and the reference image I^{REF}:

$$l^{SR}_{VGG/i,j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\phi_{i,j}\left(I^{IN}\right)_{x,y} - \phi_{i,j}\left(I^{REF}\right)_{x,y} \right)^2 \tag{2}$$

Here, $\phi_{i,j}$ represents the feature map obtained by the j $-th$ convolutional layer (after ReLU activation and without max-pooling) within the pretrained 19-layer VGG network described in Simonyan and Zisserman [34], $W_{i,j}$ and $H_{i,j}$ describe the dimensions of the corresponding feature maps within the VGG network. The application of Equation (2) to the whole dataset will generate a new reference dataset with the same size, containing the most similar, crops for input blocks. A comparison experiment with pixel similarity, described in detail in Section 4, shows that feature similarity can indeed provide more image knowledge.

3.3. End-to-End Network Structure

Compared to MZSR [50], our work aims to extract as much image knowledge as possible. Using the carefully designed model, we obtain the global and local similar reference and input images respectively. As shown in Figure 1, our Ref-ZSSR model consists of 5 modules: the feature extraction block, the SR main module, the low feature space block, the high feature space block and the image reconstruction module.

The feature extraction module extracts feature maps from both LR input and HR reference and prepares for multi-scale transformation. This module employs three convolutional layers with small 3 × 3 kernels, stride 1 and padding 1. To capture as many original features as possible while limiting model complexity, 256 filters were used in the first convolution layer, and the second and third convolution layers adopted 128 and 64 filters, respectively.

The low feature space module consists of an enhanced downsampling back-projection block (EDBP, which will be introduced later) and two convolution layers. The bottleneck layer (BOT) is used to calculate the weights. The EDBP reduces the HR reference feature

maps to the LR feature maps, the low feature space block then performs spatial alignment at the LR feature level. After two convolution layers with small 3 × 3 kernels and 64 feature maps, this module produces hybrid LR feature maps.

In the high feature space block, HR reference feature maps and the up-sampled input feature maps are merged spatially, two convolution layers with small 3 × 3 kernels and 64 filters are then employed to generate HR mixed feature maps. The low and high feature space blocks jointly provide multi-scale global similar features. Then, up- and downsampling blocks in the SR main module are provided with global similar features of the appropriate scale in order to improve super-resolution performance.

The SR main module is composed of two EDBP blocks, one enhanced up-sampling back-projection (EUBP) block [65], two time attention (TA) modules [28] and a convolution layer that spatially aligns feature maps together. The EDBP and EUBP blocks receive cross-scale hybrid feature maps and allow for the use of global similarity in super-resolution. Two TA modules are employed to extract the local similarity from the HR/LR two scales so that multi-scale local similar structures can also facilitate super-resolution.

Back-projection is based on the assumption that an estimated LR image downscaled from a good SR image is as close to the original LR image as possible. This method was first proposed in Ref. [66], and its latest modified version was introduced in Ref. [65]. Following the basic assumptions, we use ABPN to carry out the up- and down-sampling.

The TA module [28], which does not involve the location distance, directly captures long-range dependencies by calculating the interactions between two arbitrary positions. In this work, local crops from the painting are used as the input images, such that the TA module actually explores local similar information.

In summary, with the help of low and high feature space modules, the reference images can effectively provide multi-scale global similar structures, while the two TA modules comprehensively capture cross-scale local similar information. Through the interaction of the above blocks, painting-specific image knowledge is fully excavated to optimize the SR performance. Furthermore, Ref-ZSSR is an end-to-end network.

4. EXPERIMENTS

4.1. Single Painting Dataset and Reference Dataset Construction

Based on the idea of zero-shot, we repeatedly downsampled the subject painting (600 × 600 in size) with a scale factor of 1.1. Using 8-times data augmentation and 96 × 96 sliding window, more than 26K crops are obtained, and the ZSW96 dataset is generated. The training set and the testing set are divided according to the ratio of 9:1. For the reference dataset, three steps are taken to ensure global similarity. First, each 96 × 96 crop is removed from its corresponding scaled painting and filled with zeros to obtain a global painting. Second, the sliding window is used to crop the global painting into a total of 96 × 96 blocks. Finally, VGGLOSS [58] is used to find semantic global similar blocks. In this segment, feature layer relu5-1 and Equation (2) are used to calculate the perceptual similarity. Because the feature extraction is based on the last feature layer of VGG-19, the whole process can be highly computationally intensive. To accelerate the matching process, the step of the sliding window is set to 30. In addition, for comparison purposes, PSNR is adopted to obtain the pixel-similar reference dataset. Both the feature-level and pixel-level reference datasets have the same number of images. As clearly observed from Figure 2, the VGGLOSS-based reference can supply more high-frequency similar texture than the PSNR-based reference.

Table 1 Comparison of different reference images. The best results are highlighted in red.

Method	Scale	ZSW-Testset	
		PSNR	SSIM
No reference	2	32.96	0.9362
Nature images	2	32.98	0.9378
PSNR	2	33.08	0.9383
Vggloss	2	33.70	0.9451

4.2. Training Details

Except for the EDBP, EUBP and TA modules, all the convolution layers use the Prelu activation function [67]. To make the training more stable, the input image is converted to [0, 1]. We use Adam [68] as our optimization method, the learning rate is initially set to 10^{-4} and is alternately decreased by 5 and 2 every 400 epochs. All experiments are implemented using Pytorch [69] and evaluated on the NVIDIA TITAN X GPU devices.

4.3. Loss Function

There are several loss functions for SISR, such as L1 loss, L2 loss, adversarial loss [15] and perceptual loss [33]. This work intends to reconstruct the SR images as close as possible to the "HR supervision" in the training session, the learned "LR-HR" relationship is then applied to the existing LR historical painting to generate SR images with rich details. Thus, pixel-wise losses are better choices, additionally, the whole dataset is derived from a single painting so that there is a lack of data diversity and quantity. To make the training more stable and the network more convergent, L1 loss will be the most appropriate choice [70]. Therefore, we employ the L1 loss.

4.4. Model Analyses

To prove that reference images can provide useful prior knowledge and that feature-level similarity measures can explore more image priors, we also test the zero images and nature images selected from the DIV2K dataset as the reference. The qualitative and quantitative results are shown in Figure 3 and Table 1, respectively.

Table 1 lists the PSNR and SSIM results on different reference images for a scale factor of 2. As expected, the quantitative results show that valid reference images outperform zero images and nature images at both the pixel level and feature level, while that of the natural images is only 0.02 dB higher than the zero images.

Table 2 Performance comparison with state-of-the-art methods. The number in red indicates the best result and the number in blue indicates the second best result.

Method	Scale	ZSW96-Testset	
		PSNR	SSIM
BICUBIC	2	24.37	0.6682
ZSSR	2	29.02	0.8259
EDSR	2	30.34	0.8308
RCAN	2	31.52	0.8411
OURS	2	33.70	0.9451

Table 3 Comparisons of the time complexities for super-resolution of 48 × 48 LR images with scaling factor ×2. The table shows the average time required for each LR image, and the result for the fastest model is highlighted in red.

Method	ZSSR	RCAN	EDSR	OURS
Time (sec)	-	0.1051	0.2107	0.0160

These comparisons effectively confirm that irrelevant reference images provide little effective image knowledge to improve the super-resolution performance. Table 1 also shows that different similarity measures will explore different numbers of image priors. VGGLOSS-based reference images show superior results compared to those PSNR-based method by a large margin of 0.62 dB. Examples in [32–34] further demonstrate that reference images based on feature similarity can provide more similar textures.

Furthermore, the visual results in Figure 3 show that our model has an absolute advantage over the bicubic method and that the feature-level similar reference can reconstruct sharper and clearer images than the pixel-level similar reference, zero image and natural image. Moreover, using reference images from different spaces leads to significantly different results, demonstrating the importance of the similarity measures. The PSNR and SSIM of the visual results in Table 1 strongly support this conclusion.

In conclusion, Figure 3 and Table 1 jointly show that valid reference images are indeed quite helpful for improving the super-resolution performance and that the VGGLOSS-based reference images produce the best result in both precision and visualization. These results convincingly demonstrate that the reference-based method can effectively compensate for the insufficient prior knowledge in zero-shot super-resolution.

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

Ref-ZSSR mainly aims to handle a single LR painting whose degradation process is complex, diverse and unknown. Since there is no ground truth for ancient paintings, we will visually show the final SR results (as shown in Figure 4). To quantitatively measure the performance of Ref-ZSSR, we performed some comparative experiments and tested ZSSR [23] and some other typical SISR methods, including EDSR [13] and RCAN [14]. Our Ref-ZSSR and ZSSR [23] are both based on the zero-shot approach to achieve super-resolution for a single real LR image. Both methods do not rely on any external instances and are trained in an unsupervised or self-supervised manner, but compared with ZSSR [23], Ref-ZSSR is superior in mining image knowledge to improve the super-resolution performance. As mentioned in Section 3.3, Ref-ZSSR introduces the reference images and combines global and local similar internal structures in a cross-scale manner to facilitate super-resolution. We trained ZSSR [23] using the painting-specific dataset ZSW96, and the comparison results are shown in Table 2. Compared to ZSSR, our method shows an improvement of 4.68 dB at a scale ×2 . As demonstrated in Figure 5, it is clear that the Ref-ZSSR method can generate richer texture details and better visual perception SR images, while ZSSR tends to produce over-smoothed results. For a single LR painting, the SR result is greatly improved after full exploration of the internal statistics at multiple scales by introducing the reference.

We also compare our model with two other typical supervised SISR methods, i.e., EDSR [13] and RCAN [14]. For fairness, our ZSW96 dataset is used to train EDSR and RCAN. The values of the obtained results are summarized in Table 2, and a visual comparison of the results is shown in Figure 5. We found that our method outperforms EDSR by more than 3.36 dB [13] and RCAN by more than 2.18 dB [14]. Moreover, as shown in Figure 5, our method produces significantly better qualitative results with much clearer high-frequency details. Compared with EDSR, RCAN has better quantitative and visual performance for LR real painting, Generally, we found that our model is more effective for LR paintings with unknown damage processes. On the other hand, EDSR [13] and RCAN [14] outperform Ref-ZSSR for images with a ground truth.

4.6. Complexity

Table 3 shows a comparison of time complexities between Ref-ZSSR and other state-of-the art methods. All comparisons are performed in the environment of NVIDIA TITAN X GPU devices. We adopt the average time required for super-resolution of each 48 × 48 LR image with a scaling factor of ×2 in order to measure the time complexities. Our model requires only one-sixth of the time consumed by RCAN and one-thirteenth of that required by EDSR. It is clear that our model has a large advantage in terms of the time complexity. By contrast, ZSSR combines training and testing and requires thousands of forward and backward iterations to obtain a super-resolution image. For a fair comparison, we do not compare the time complexity with ZSSR. We also compare model complexities of our approach with EDSR, RCAN and ZSSR. The numbers of parameters are shown in Figure 6. Both EDSR and RCAN are supervised methods, while ZSSR and our network are totally unsupervised. Compared with supervised learning models, zero-shot based unsupervised models require much fewer parameters. Due to the addition of reference images, Ref-ZSSR has more parameters than the ZSSR method.

5. CONCLUSION

In this work, we have extended the powerful deep learning approach to the field of cultural heritage preservation. We make the first

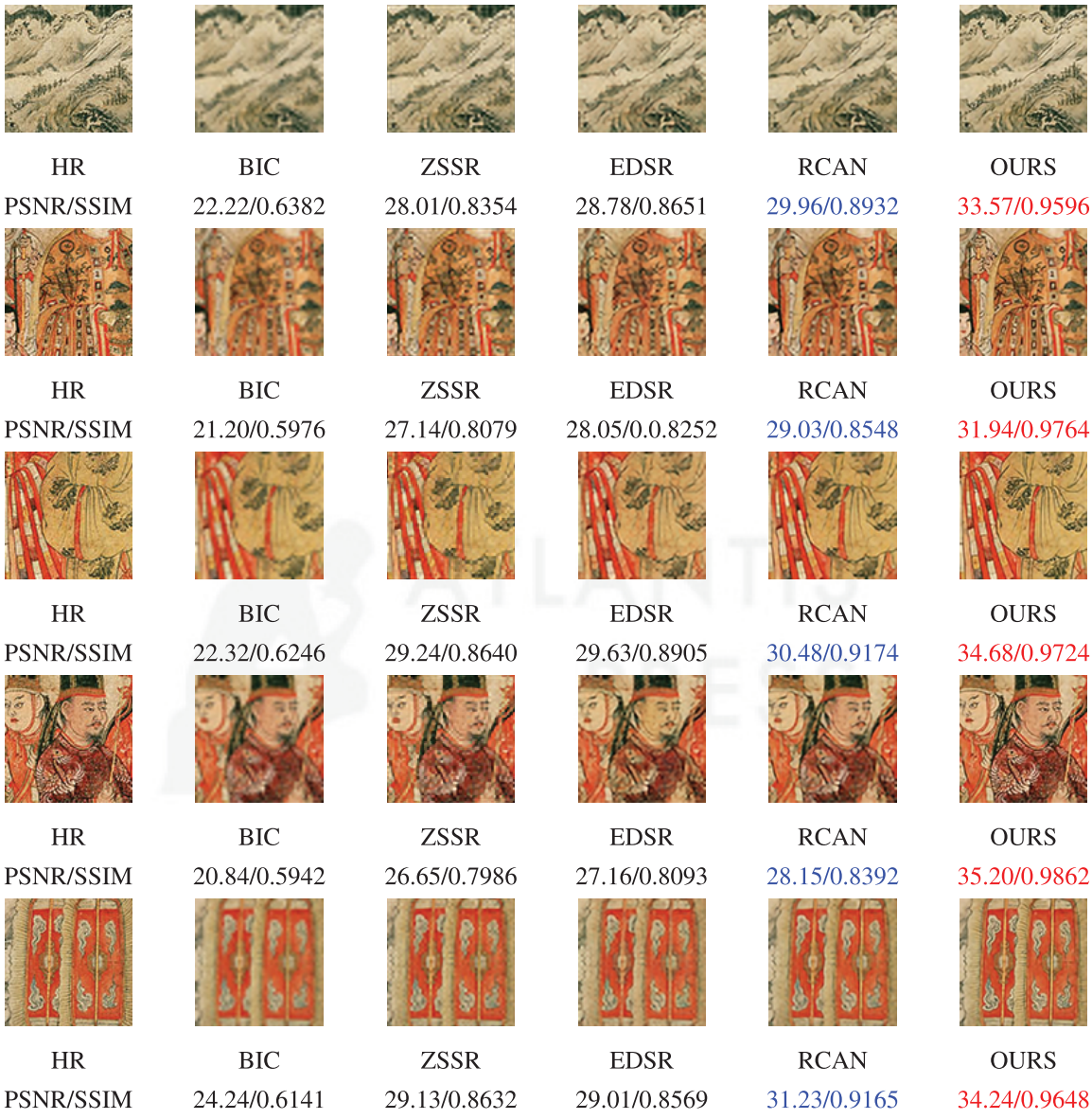

Figure 5 | Performance comparison of the images produced by the state-of-the-art methods. The best result and the second best result are marked in red and blue, respectively. It is clear that our Ref-ZSSR scheme can produce sharper images than the other methods.

attempt to combine zero-shot and reference-based methods and train a painting-specific CNN without relying on any external examples or prior training. We use perceptive similarity to search for global similarity blocks, which serve as reference images and provide global similarity. Time attention modules are used to fully explore local similar information. Both global and local image knowledge is used in super-resolution at multiple scales. For the unique paintings with no ground truth, we found that our method (Ref-ZSSR) outperforms the similar ZSSR [23] in both precision and visual experience. Ref-ZSSR also outperforms state-of-the-art SR methods both qualitatively and quantitatively. Our method can be applied to arbitrary paintings or single low-quality images, regardless of the size, degree of damage or the degradation process of the artwork.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

AUTHORS' CONTRIBUTIONS

Hongzhen Shi: writing-original draft; Dan Xu: funding aquisition; Hao Zhang: visualization; YingYing Yue: data curation.

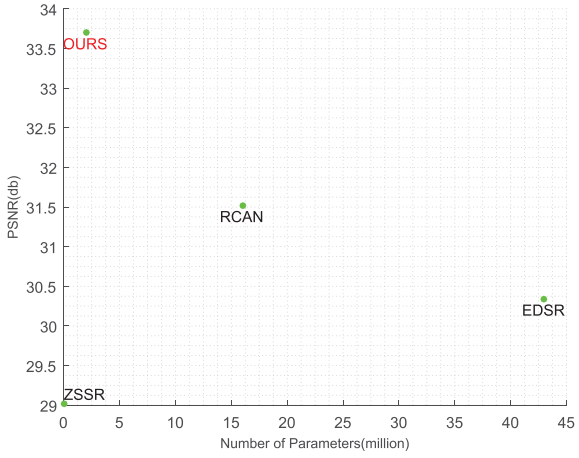

Figure 6 | Comparisons of model complexity (the number of parameters).

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 61761046, Grant 61540062 and Grant 62061049, in part by the Yunnan Province Ten Thousand Talents Program and Yunling Scholars Special Project under Grant YNWR-YLXZ-2018-022, in part by the Yunnan Provincial Science and Technology Department-Yunnan University "Double First Class" Construction Joint Fund Project under Grant No. 2019FY003012.

REFERENCES

[1] D. Hale, Image-guided blended neighbor interpolation of scattered data, in SEG Technical Program Expanded Abstracts 2009, Society of Exploration Geophysicists, 2009, pp. 1127–1131.

[2] N. Jiang, L. Wang, Quantum image scaling using nearest neighbor interpolation, Quantum Inf. Process. 14 (2015), 1559–1571.

[3] D. Suresha, H.N. Prakash, Natural image super resolution through modified adaptive bilinear interpolation combined with contra harmonic mean and adaptive median filter, Int. J. Image Graph. Signal Process. 8 (2016), 1.

[4] K.T. Gribbon, D.G. Bailey, A novel approach to real-time bilinear interpolation, in Proceedings, Second IEEE International Workshop on Electronic Design, Test and Applications (DELTA 2004), IEEE, Perth, Australia, 2004, pp. 126–131.

[5] Z. Huang, L. Cao, Bicubic interpolation and extrapolation iteration method for high resolution digital holographic reconstruction, Optics Lasers Eng. 130 (2020), 106090.

[6] Z. Dengwen, An edge-directed bicubic interpolation algorithm, in 2010 3rd International Congress on Image and Signal Processing, IEEE, Yantai, China, 2010, vol. 3, pp. 1186–1189.

[7] M.Q. Nguyen, P.M. Atkinson, H.G. Lewis, Superresolution mapping using a hopfield neural network with fused images, IEEE Trans. Geosci. Remote Sens. 44 (2006), 736–749.

[8] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Non-local sparse models for image restoration, in 2009 IEEE 12th International Conference on Computer Vision, IEEE, Kyoto, Japan, 2009, pp. 2272–2279.

[9] J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (2010), 2861–2873.

[10] K. Egiazarian, V. Katkovnik, Single image super-resolution via bm3d sparse coding, in 2015 23rd European Signal Processing Conference, IEEE, Nice, France, 2015, pp. 2849–2853.

[11] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2015), 295–307.

[12] J. Kim, J.K. Lee, K.M. Lee, Deeply-recursive convolutional network for image super-resolution, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 1637–1645.

[13] B. Lim, S. Son, H. Kim, S. Nah, K.M. Lee, Enhanced deep residual networks for single image super-resolution, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 2017, pp. 136–144.

[14] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in The IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in Advances in Neural Information Processing Systems, Montreal, CANADA, 2014, pp. 2672–2680.

[16] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, Spatial transformer networks, in Advances in Neural Information Processing Systems, Montreal, CANADA, 2015, pp. 2017–2025.

[17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 770–778.

[18] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature. 521 (2015), 436–444.

[19] I. Izonin, R. Tkachenko, D. Peleshko, T. Rak, D. Batyuk, Learning-based image super-resolution using weight coefficients of synaptic connections, in 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), IEEE, Lviv, Ukraine, 2015, pp. 25–29.

[20] R. Tkachenko, P. Tkachenko, I. Izonin, Y. Tsymbal, Learning-based image scaling using neural-like structure of geometric transformation paradigm, in: Advances in Soft Computing and Machine Learning in Image Processing, Springer, Cham, Switzerland, 2018,pp. 537–565.

[21] D. Glasner, S. Bagon, M. Irani, Super-resolution from a single image, in 2009 IEEE 12th International Conference on Computer Vision, IEEE, Kyoto, Japan, 2009, pp. 349–356.

[22] W.T. Freeman, T.R. Jones, E.C. Pasztor, Example-based super-resolution, IEEE Comput. Graph. Appl. 22 (2002), 56–65.

[23] A. Shocher, N. Cohen, M. Irani, Zero-shot super-resolution using deep internal learning, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 3118–3126.

[24] M. Zontak, M. Irani, Internal statistics of a single natural image, in CVPR 2011, IEEE, Colorado Springs, CO, USA, 2011, pp. 977–984.

[25] A. Buades, B. Coll, J.-M. Morel, A non-local algorithm for image denoising, in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, San Diego, CA, USA, 2005, vol. 2, pp. 60–65.

[26] K. Dabov, A. Foi, V. Katkovnik, K. Egiazarian, Image denoising by sparse 3-d transform-domain collaborative filtering, IEEE Trans. Image Process. 16 (2007), 2080–2095.
[27] S. Gu, L. Zhang, W. Zuo, X. Feng, Weighted nuclear norm minimization with application to image denoising, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 2862–2869.
[28] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7794–7803.
[29] J. Kim, J.K. Lee, K. MuLee, Accurate image super-resolution using very deep convolutional networks, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 1646–1654.
[30] T. Tong, G. Li, X. Liu, Q. Gao, Image super-resolution using dense skip connections, in Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017, pp. 4799–4807.
[31] M.S. Rad, B. Bozorgtabar, V. Marti, M. Basler, H.K. Ekenel, J.-P. Thiran, SROBB: targeted perceptual loss for single image super-resolution, in Proceedings of the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019, pp. 2710–2719.
[32] J. Bruna, P. Sprechmann, Y. LeCun, Super-resolution with deep convolutional sufficient statistics, arXiv preprint arXiv:1511.05666, 2015.
[33] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in European Conference on Computer Vision, Amsterdam, The Netherlands, 2016, pp. 694–711.
[34] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.
[35] L.A. Gatys, A.S. Ecker, M. Bethge, Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks, arXiv preprint arXiv:1505.07376, 2015.
[36] S. Vasu, N.T. Madam, A.N. Rajagopalan, Analyzing perception-distortion tradeoff using enhanced perceptual super-resolution network, in Proceedings of the European Conference on Computer Vision, Munich, German, 2018.
[37] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization, arXiv preprint arXiv:1506.06579, 2015.
[38] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, W. Shi, Real-time video super-resolution with spatio-temporal networks and motion compensation, in The IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017.
[39] Y. Liu, J. Guo, D. Cai, X. He, Attribute attention for semantic disambiguation in zero-shot learning, in Proceedings of the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019, pp. 6698–6707.
[40] C. Zhan, D. She, S. Zhao, M.-M. Cheng, J. Yang, Zero-shot emotion recognition via affective structural embedding, in Proceedings of the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019, pp. 1151–1160.
[41] M. Ye, Y. Guo, Progressive ensemble networks for zero-shot recognition, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019.
[42] Q. Liu, L. Xie, H. Wang, A.L. Yuille, Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval, in Proceedings of the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019, pp. 3662–3671.
[43] S. Dey, P. Riba, A. Dutta, J. Llados, Y.-Z. Song, Doodle to search: practical zero-shot sketch-based image retrieval, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019.
[44] B. Chen, W. Deng, Hybrid-attention based decoupled metric learning for zero-shot image retrieval, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019.
[45] A. Dutta, Z. Akata, Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019.
[46] M. Elhoseiny, M. Elfeki, Creativity inspired zero-shot learning, in Proceedings of the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019, pp. 5784–5793.
[47] K. Li, M.R. Min, Y. Fu, Rethinking zero-shot learning: a conditional visual classification perspective, in Proceedings of the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019, pp. 3583–3592.
[48] J. Li, X. Lan, Y. Liu, L. Wang, N. Zheng, Compressing unknown images with product quantizer for efficient zero-shot classification, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019.
[49] F. Sener, A. Yao, Zero-shot anticipation for instructional activities, in Proceedings of the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019, pp. 862–871.
[50] J.W. Soh, S. Cho, N.I. Cho, Meta-transfer learning for zero-shot super-resolution, arXiv preprint arXiv:2002.12213, 2020.
[51] Y. Zhu, Y. Zhang, A.L. Yuille, Single image super-resolution using deformable patches, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 2917–2924.
[52] H. Yue, X. Sun, J. Yang, F. Wu, Landmark image super-resolution by retrieving web images, IEEE Trans. Image Process. 22 (2013), 4865–4878.
[53] C. Liu, D. Sun, A bayesian approach to adaptive video super resolution, in CVPR 2011, IEEE, Colorado Springs, CO, USA, 2011, pp. 209–216.
[54] L. Sun, J. Hays, Super-resolution from internet-scale scene matching, in 2012 IEEE International Conference on Computational Photography, IEEE, Seattle, WA, USA, 2012, pp. 1–12.
[55] B. Huang, A. Singh, N. Ahuja, Single image super-resolution from transformed self exemplars, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 2015, pp. 5197–5206.
[56] H. Zheng, M. Ji, H. Wang, Y. Liu, L. Fang, Crossnet: an end-to-end reference-based super resolution network using cross-scale warping, in Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018, pp. 88–104.
[57] Z. Zhang, Z. Wang, Z. Lin, H. Qi, Image super-resolution by neural texture transfer, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019.
[58] C. Ledig, L. Theis, F. Huszár, *et al.*, Photo-realistic single image super-resolution using a generative adversarial network, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 4681–4690.

[59] R. Timofte, V. De Smet, L. Van Gool, Anchored neighborhood regression for fast example based super-resolution, in Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 2013, pp. 1920–1927.
[60] R. Timofte, V. De Smet, L. Van Gool, A+: adjusted anchored neighborhood regression for fast super-resolution, in Asian Conference on Computer Vision, Singapore, 2014,pp. 111–126.
[61] C. Dong, C.C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in European Conference on Computer Vision, Zurich, Switzerland, 2014, pp. 184–199.
[62] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, IEEE, Pacific Grove, CA, USA, 2003, vol. 2, pp. 1398–1402.
[63] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004), 600–612.
[64] P. Gupta, P. Srivastava, S. Bhardwaj, V. Bhateja, A modified PSNR metric based on HVS for quality assessment of color images, in 2011 International Conference on Communication and Industrial Application, IEEE, Kolkata, India, 2011, pp. 1–4.
[65] S. Liu, W. Wang, T. Li, W.-C. Siu, Image super-resolution via attention based back projection networks, in IEEE International Conference on Computer Vision Workshop (ICCVW), Seoul, South Korea, 2019.
[66] M. Haris, G. Shakhnarovich, N. Ukita, Deep back-projection networks for super-resolution, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1664–1673.
[67] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015, pp. 1026–1034.
[68] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.
[69] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, 2017.
[70] H. Zhao, O. Gallo, I. Frosio, J. Kautz, Loss functions for neural networks for image processing, arXiv preprint arXiv:1511.08861, 2015.