Neural Networks 143 (2021) 400–412

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Enhanced image prior for unsupervised remoting sensing super-resolution[☆]

Jiaming Wang[a], Zhenfeng Shao[a,*], Xiao Huang[b], Tao Lu[c], Ruiqian Zhang[d], Jiayi Ma[e]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430079, China
[b] Department of Geosciences, University of Arkansas, Fayetteville, AR, 72701, USA
[c] Hubei Key Laboratory of Intelligent Robot, School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, 430073, China
[d] The School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, 430079, China
[e] The Electronic Information School, Wuhan University, Wuhan, 430079, China

ARTICLE INFO

Article history:
Received 27 February 2021
Received in revised form 26 April 2021
Accepted 3 June 2021
Available online 18 June 2021

Keywords:
Latent space
Satellite imagery
Unsupervised learning
Prior enhancement
Super resolution

ABSTRACT

Numerous approaches based on training low-high resolution image pairs have been proposed to address the super-resolution (SR) task. Despite their success, low-high resolution image pairs are usually difficult to obtain in certain scenarios, and these methods are limited in the actual scene (unknown or non-ideal image acquisition process). In this paper, we proposed a novel unsupervised learning framework, termed Enhanced Image Prior (EIP), which achieves SR tasks without low/high resolution image pairs. We first feed random noise maps into a designed generative adversarial network (GAN) for satellite image SR reconstruction. Then, we convert the reference image to latent space as the enhanced image prior. Finally, we update the input noise in the latent space with a recurrent updating strategy, and further transfer the texture and structured information from the reference image. Results on extensive experiments on the Draper dataset show that EIP achieves significant improvements over state-of-the-art unsupervised SR methods both quantitatively and qualitatively. Our experiments on satellite (SuperView-1) images reveal the potential of the proposed approach in improving the resolution of remote sensing imagery compared with the supervised algorithms. Source code is available at https://github.com/jiaming-wang/EIP.

1. Introduction

Recently, remote sensing satellites that enable uninterrupted target observation have drawn widespread attention in various practical applications. Compared with satellites that obtain static images (*e.g.*, the Jilin-1, He, He, Mei, & Hu, 2019 and Zhuhai OVS-1 A/B, Cao et al., 2018), high temporal satellites are able to acquire time-series images, facilitating the continuous monitoring of moving targets. Such satellites have a wider range of applications, including extraction of urban impervious surfaces (Shao, Fu, Li, Altan, & Cheng, 2019), land cover classification (Lv et al., 2018), event detection (Gu, Wang, Jin, & Gao, 2020), and resource exploration (Li, Shao, Zhang, Huang, & Zhang, 2021). It is widely acknowledged that optical remote sensors own the contradictory nature between spatial and spectral resolution. In addition, due to limitations in bandwidth and hardware cost, the spatial resolution of satellites with a high temporal resolution often fails to meet the demand of high precision applications that require precise monitoring. Therefore, improving the spatial resolution of satellite images with large compression ratios has become an urgent issue in remote sensing applications.

One effective method to improve spatial resolution is to employ longer focal ratios or equip charge-coupled devices (CCD) with a smaller pixel sizes and increased pixel density (Merino & Nunez, 2007). However, both methods require renovations of orbital platforms, which is often costly. Another solution is to implement super-resolution (SR) that aims to reconstruct the high resolution (HR) image from observed low resolution images (LR) (Park, Park, & Kang, 2003). In real-world remote sensing scenarios, SR problems are characterized by the following properties (Haut et al., 2018): (1) HR video images are often unavailable; (2) HR image datasets that share the same content with LR video images widely exist and can serve as a good reference (Ref); (3) because of the different sensors and imaging band, the imaging environment greatly differs between the LR and Ref-HR images (more details at Section 2.3).

Existing supervised SR methods often generate SR images/ patches from the prior information provided by LR/HR pairs. The

[☆] This work was supported in part by the National Key R&D Program of China under Grant 2018YFB0505401, the National Natural Science Foundation of China under Grants 41890820, 41771452, 41771454, 62072350, the Key R&D Program of Yunnan Province under Grant 2018IB023.

[*] Corresponding author.
E-mail address: shaozhenfeng@whu.edu.cn (Z. Shao).

https://doi.org/10.1016/j.neunet.2021.06.005
0893-6080/© 2021 Elsevier Ltd. All rights reserved.

development of machine learning greatly promotes the progress of SR. The supervised deep-learning-based SR methods achieve excellent performance via the innovation of networks, *e.g.*., enhanced deep residual networks (Lim, Son, Kim, Nah, & Lee, 2017), residual dense network (Zhang, Tian, Kong, Zhong, & Fu, 2018b), non-local (Mei et al., 2020), and attention mechanism (Dai, Cai, Zhang, Xia, & Zhang, 2019; Niu et al., 2020). The rationale of these algorithms can be summarized as follows: a deep learning model is trained to learn the mapping between the corrupted LR images (degraded from their original versions) and corresponding HR ones via a convolutional neural network (CNN). Based on these methods, reference-based image super-resolution (RefSR) methods, another research hotspot in recent years, introduce additional information from a reference image, such as Dong, Zhang, and Fu (2021), Yue, Zhou, Jiang, Yang, and Hou (2021). The superior performance of supervised learning substantially depends on the pixel-by-pixel supervision of the ground truth. Although these methods are intended to obtain deep features from image prior information, HR training examples needed to provide such information are difficult to obtain, due to the cost constraints and hardware limitations (Haut et al., 2018). In addition, despite the establishment of mapping functions between LR and HR images, such models are only suitable for ideal degradation models (Bicubic kernel of simulated the degradation model), and their performance is found rather limited in actual scenes due to the unknown or non-ideal acquisition processes (Levin, Weiss, Durand, & Freeman, 2009; Shocher, Cohen, & Irani, 2018; Soh, Cho, & Cho, 2020).

In light of the aforementioned limitations of supervised SR, unsupervised SR, an approach to reconstruct images without ground truth, has received more attention in recent years. To exploit the prior structure in an image, these studies (Haut et al., 2018; Soh et al., 2020; Ulyanov, Vedaldi, & Lempitsky, 2018) generate high-resolution images from random noises or LR images via a generative adversarial network (GAN), where fixed input noise tends to cause the model to fall into local optimum (more details in Section 4). However, recovering high-frequency information, especially information regarding the edge and texture, from the existing image prior remains a difficulty for unsupervised SR methods (Lempitsky, Vedaldi, & Ulyanov, 2018).

Inspired by the development of RefSR, we intend to explore the potential of unsupervised reference-based strategies to overcome the aforementioned obstacles. We propose a novel unified framework, termed Enhanced Image Prior (EIP), to solve SR tasks in an unsupervised manner. The designed EIP consists of three major components: (1) a generator, (2) a reference feature extractor, and (3) a recurrent updater. For image generation, we adopt a deep GAN to connect shallow features and deep features. The proposed method applies an encoder–decoder model with spatial transformer networks (Jaderberg, Simonyan, Zisserman, & Kavukcuoglu, 2015) to learn the transformer parameters that facilitate the alignment between LR images and the reference HR image (Ref-HR), which is a necessary step due to their discrepancies introduced from different satellite perspectives. The adopted GAN model uses the noise map as input. The feature map of the reference image is further transformed into the latent space as an enhanced image prior for updating the input noises. We further conduct extensive experiments to demonstrate the superiority of the proposed framework. The experimental results suggest that the proposed method can obtain more realistic images (Chen et al., 2016) and outperforms many state-of-the-art unsupervised algorithms.

The main innovative contributions of this paper are three folds:

1. **To our best knowledge, the proposed method is the first unsupervised SR approach that takes advantage of similar**

   Ref-HR images as the reference. Even without pixel-by-pixel supervision, high-frequency information from Ref-HR images can be translated into LR domains for textured generation, achieving great performance compared with other state-of-the-art unsupervised SR methods.
2. **We propose a novel model that transforms the reference into a latent space as the enhanced image prior via a recurrent update strategy.** Our model demonstrates stronger robustness and achieves global optimization without falling into local optimization.
3. We test the proposed method on the SuperView-1 video satellite images, a real-world scenario without ground truth, and validate the effectiveness and universality of the proposed method qualitatively and quantitatively. **We find that the proposed unsupervised EIP framework even outperforms selected supervised SR algorithms.**

The remainder of this paper is organized as follows. Section 2 provides a brief review of related works on satellite image super-resolution, single image super-resolution, and reference-based image super-resolution. Section 3 describes the proposed method in detail. Sections 4 and 5 report the experimental results and ablation studies, respectively. Section 6 concludes the study.

2. Related work

In this section, we briefly review the single image super resolution (SISR). Specifically, we focus on outlining existing works of unsupervised SR and RefSR.

2.1. Supervised single image super resolution via deep learning

Inspired by deep learning theories, CNNs received much attention in computer vision tasks (Wang et al., 2021; Xu, Ma, Jiang, Guo, & Ling, 2020) in recent years. Many CNN-based SISR methods have been developed, achieving decent improvements in the SR domain. The CNN-based SISR method was first proposed by Dong, Loy, He, and Tang (2016), who constructed a non-linear mapping function of LR and HR images/patches with an end-to-end three-layers CNN framework, opening up a new avenue in SR tasks. Kim, Lee, and Lee (2016) proposed a very deep CNN network with skip connection, providing stable and fast convergence to avoid the disappearance of the gradient. Further, ESPCNN (Shi et al., 2016) replaced the up-sampling version of input LR patches with the subpixel convolution layer to reduce the checkerboard artifact. With the success of the residual block in recognition tasks (He, Zhang, Ren, & Sun, 2016), various works (Cao, Yao, & Liang, 2020; Lim et al., 2017; Liu & Cao, 2020; Zhang et al., 2018b) have focused on designing deeper networks and combining multi-level feature maps to fully exploit the hierarchical features. For example, the attention mechanism was proposed to improve the CNN performance for various tasks. Zhang et al. (2018) developed a deep residual channel attention network (RCAN) to capture the adaptively re-scale channel-wise features. In the most recent efforts, Guo et al. (2020) introduced a dual regression scheme by introducing additional supervision.

Recently, Luo, Zhou, Wang, and Wang (2017) improved the SR performance for satellite images by proposing a mirroring reflection method that considers HR satellite video data to avoid the loss of border information in images. Lu, Wang, Zhang, Wang, and Jiang (2019) developed a framework to simulate receptive fields in varying sizes by fusing multi-scale information in the residual domain, which effectively enhances the high-frequency information. Jiang et al. (2018) proposed a deep distillation recursive network to extract features through the addition of rich interactive links in and between multiple-path units in each ultra-dense residual block. Jiang et al. (2019) generated clean and real image details by proposing a GAN-based edge-enhancement method, consisting of an ultra-dense sub-network and an edge-enhancement sub-network. Lei, Shi, and Zou (2019) investigated a pair of images to better discriminate its inputs for the low-frequency regions in a remote sensing image. Zhang, Chen, Ma, and Zhang (2020) introduced the varying saliency maps of different areas as the additional prior information and constraints. Geng, Liu, Wang, and Sun (2021) proposed a shearlet-transform-based residual network for an optimal sparse approximation. Despite the great performance of the aforementioned methods, they are intrinsically designed for supervised SR tasks that require known HR/LR pairs and their performances largely rely on massive training samples.

2.2. Unsupervised single image super resolution

Despite that efforts have been made to solve SR problems with a known imaging acquisition process, few works tried to solve SR problems without ground truth. In traditional unsupervised SR, searching self-examples and reconstruction (Freedman & Fattal, 2011; Huang, Singh, & Ahuja, 2015) are two common strategies. Michaeli and Irani (2013) presented an algorithm that is able to estimate the optimal blur kernel. However, the shallow networks usually failed to recover high-frequency information, especially in non-linear scenarios, leading to their limited performance.

From the perspective of training mode, existing unsupervised SR methods can be divided into zero-shot and data-based methods. However, CinCGAN (Yuan et al., 2018), a representative of the data-based unsupervised SR methods, requires on a large amount of training data to obtain superior performance, which cannot be directly applied to the zero-shot scenes. Haut et al. (2018), for the first time, proposed an unsupervised hourglass model to super-resolved LR remote sensing images from random noise in a zero-shot framework. DIP (Ulyanov et al., 2018) exploited the prior information in LR images and generated SR. Recently studies (Assaf Shocher, 2018; Soh et al., 2020) focused on down-sampled kernel learning for internal learning in the actual scene. USISResNet (Prajapati et al., 2020) employed a subnetwork with the capability of conducting quality assessments to encourage the network to produce texture with excellent perceived quality. Ahn, Yoo, and Sohn (2020) built a set of pseudo pairs and cast the original unsupervised task into a supervised learning task. Despite the above efforts, recovering the missing high-frequency information, especially information regarding the edge and texture, remains a challenging task.

2.3. Reference-based image super-resolution

Different from SISR methods, RefSR algorithms render more accurate and realistic details, which are transferred from the reference images. A reference image is similar to an LR image in terms of content but different in focal lengths and shot perspectives. Video frames (Liu & Sun, 2011), multi-view images (Zhu, Zhang, & Yuille, 2014), and web-retrieval images (Yue, Sun, Yang, & Wu, 2013) are the common sources for reference images.

Several RefSR algorithms (Wang, Liu, Heidrich, & Dai, 2017; Yue et al., 2013; Zheng, Ji, Wang, Liu, & Fang, 2018) achieved great performance when LR and reference images were tightly aligned. However, a reduction of performance occurs when LR and reference images present considerable discrepancies, given their inability to effectively coupling LR and reference images. In light of this issue, Zhang, Wang, Lin, and Qi (2019) proposed a deep model that adopts local texture matching for long-distance dependency. Most recently, Yang, Yang, Fu, Lu, and Guo (2020) introduced a new approach for searching and transferring relevant textures from the reference images to LR images. SSEN (Shim, Park, & Kweon, 2020) was further proposed to align the reference images and LR images in the feature domain to capture similarity-aware. To handle large displacement between LR and reference images, Yue et al. (2021) proposed a coarse-to-fine dense warping strategy that can deal with both large- and small-scale misalignments well. Despite their better performances compared with SISR methods, they cannot be applied in scenarios without ground truth images.

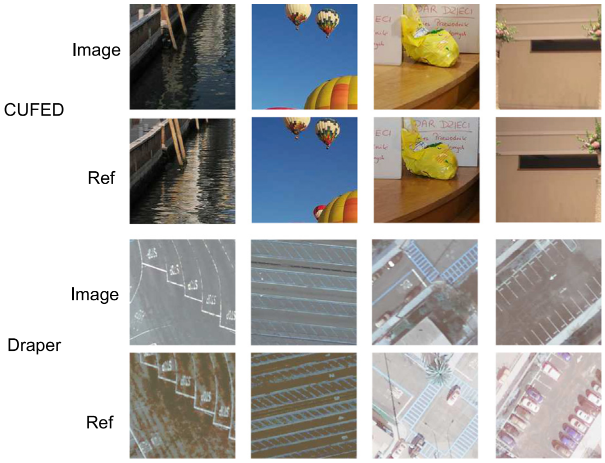

Fig. 1. Comparison between CUFED (the traditional reference dataset) with Draper (the satellite image dataset used in this paper). Images in the Draper dataset present heterogeneous visual characteristics due to a variety of reasons that include the difference in synthetic bands (the second row of Draper in Fig. 1)) and overexposure issues (the next two columns of Draper in Fig. 1)), causing great difficulty for traditional algorithms to transform features between the reference and the input image.

Images in the traditional reference dataset, i.e., CUFED (Wang et al., 2016), are taken within a relatively short time interval (3 h), suggesting a generally homogeneous imaging environment. Fig. 1 presents the comparison between selected images in CUFED and Draper (the satellite image dataset used in this paper). We can observe that the input and reference images in CUFED are similar both in the content and imaging environment, and therefore feature fusion can be performed with traditional algorithms within the feature domain. However, satellite sample images from Draper show considerably different visual characteristics due to the difference in synthetic bands (the second row of Draper in Fig. 1) and overexposure issues (the next two columns of Draper in Fig. 1), causing great difficulty for traditional algorithms to transform features from the reference image to the input image.

3. Our method

In this section, we provide detailed information regarding the problem formulation, loss function, and network architecture that consists of three major components: a generator, a reference feature extractor, and a recurrent updater.

3.1. Problem formulation

To recover the missing high-frequency textures along with other details from LR images $\boldsymbol{I}^{LR} \in \Re^{C\times H\times W}$ and reconstruct an

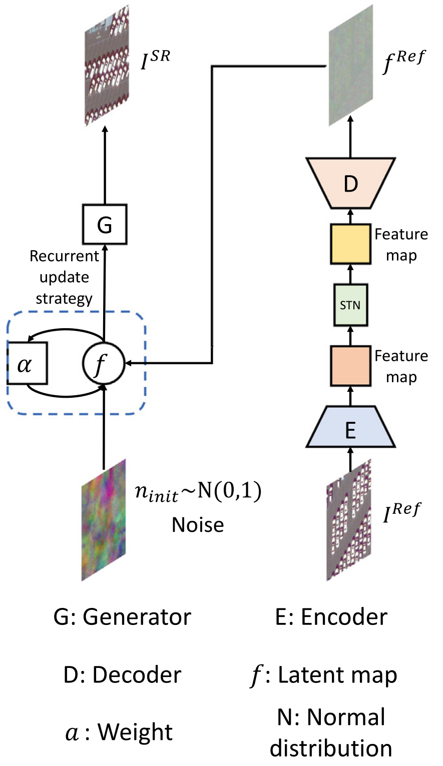

Fig. 2. Illustration of the EIP framework that includes three major components: the generator, the reference feature extractor, and the recurrent updater.

HR image $\boldsymbol{I}^{HR} \in \Re^{C\times t\cdot H\times t\cdot W}$, the conventional formulation of SR methods can be represented as $\boldsymbol{I}^{LR} = \boldsymbol{D}\boldsymbol{I}^{HR}$, where $\boldsymbol{D}$ denotes the down-sampling matrix, and t denotes the scale factor. Here, we assume that HR/LR pairs are unavailable and impractical to obtain. Nevertheless, we can obtain a set of LR and Ref-HR images used for unsupervised training. Rather than minimizing the error between the SR images and the ground truth in most supervised methods, the proposed method is to retrieve texture and content information from the reference that benefits the reconstruction of LR images with similar information. Thus, the goal of the proposed method is to fuse heterogeneous information via a unified framework.

In reference images, the information to be preserved includes low-level information (*e.g.*, textures and edges) and high-level information (*e.g.*, content and global structures). As unsupervised methods usually fail to estimate high-frequency information in high-resolution space compared to traditional SR methods, it is difficult to transform information from the reference image into the generated image within the feature domain. Experiments also prove that observable image artifacts are introduced when the transformation process takes place in the feature space (more details in Section 4.4). We solve this problem by referring to the latent code of the reference image that carries implicit information.

The pipeline of EIP is summarized in Fig. 2. We denote $\boldsymbol{I}^{Ref} \in \Re^{C\times t\cdot H\times t\cdot W}$ as the corresponding HR reference image. The random noise maps $\boldsymbol{n}_{init}$ are represented as $C' \times t\cdot H \times t\cdot W$. The proposed method mainly consists of three major components: a generative network, a reference feature extraction network, and an enhanced image prior model. First, a mapping function from noise maps to an HR image is learned. Further, an encoder–decoder model to code and transform the reference image is adopted. Finally, the coded feature maps of the reference image are mapped into the latent space and random noise $\boldsymbol{n}_{init}$ is updated. In the following subsections, we give more details on the generator, the reference feature extractor, the enhanced image prior model.

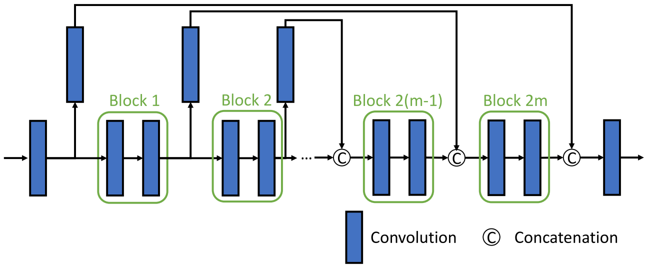

Fig. 3. Illustration of the skip model in the proposed generative network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. The generator

Different from common GAN-based image generation tasks, SR tasks requires that the generated images to be not only high-quality but also as real as possible. Directly applying GAN-based SR models (Goodfellow et al., 2014; Ledig et al., 2017) in an unsupervised framework often causes notable checkerboard phenomenon due to the up-sampling operation (more details in Section 4.4). Given an input HR-size noise map, we first generate an image, which can be formulated as,

$$\boldsymbol{I}^{SR} = H(\boldsymbol{n}), \tag{1}$$

where $\boldsymbol{n}$ denotes the noise map, $H(.)$ denotes the function of the SR network in the proposed method, and $\boldsymbol{I}^{SR}$ refers to the output of the SR network.

In deep CNN, information redundancy often exists between shallow feature maps and deep feature maps. Dense convolutional networks (Huang, Liu, Van Der Maaten, & Weinberger, 2017) that connect each layer to every other layer are proposed to encourage feature re-utilization and thus achieves high performance in image classification tasks. Although no additional parameters are introduced in dense networks, such dense networks lead to a dramatic increase in computation caused by the sharp increase of the channels of feature maps. As for the image SR tasks, it is crucial to exploit and fuse the shallow and deep feature maps in an effective manner. In the proposed method, we adopt stacked skip models for reconstructing purposes, as shown in Fig. 3. The skip generative network consists of $2m(m = 4)$ convolutional blocks (shown in green color). Each convolutional block, with 128 feature maps, is composed of two 3 × 3 convolutions with a stride of 1. Furthermore, each convolutional layer is followed by the Leaky ReLU activation and the Batch Normalization (Ioffe & Szegedy, 2015). For the proposed skip model, we generate deep features by extracting shallow feature maps and further modify the dense skip connection strategy to interval connection.

3.3. Reference feature extraction

Most traditional reference-based supervised approaches aim to design deep networks that align the reference and SR images in the feature domain (Yang et al., 2020; Zhang et al., 2019). The performance of these algorithms is highly dependent on pixel-by-pixel supervision. Considering the great heterogeneity in satellite images (*e.g.*, different shooting angles and different time periods), we advocate an encoder–decoder model to exploit the prior of the reference image, which can be formulated as,

$$\boldsymbol{f}^{Ref} = encode\_decode(\boldsymbol{I}^{Ref}), \tag{2}$$

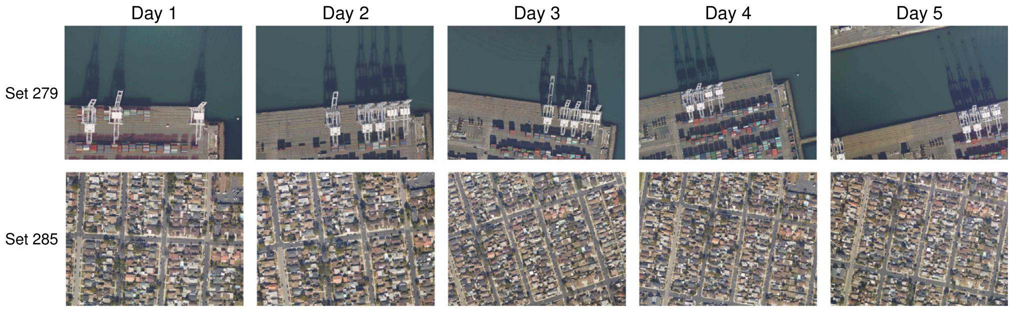

Fig. 4. Example images are randomly selected from Draper. Each group of images was taken at different times. The LR version of the fifth column ("Day 5" in Fig. 4) is used as the high temporal resolution LR image, and the fourth column ("Day 4" in Fig. 4) acts as the corresponding HR reference image in this paper. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

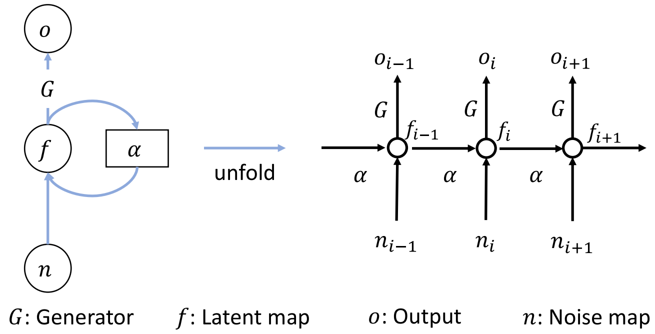

Fig. 5. The proposed recurrent updating strategy.

where *encode_decode*() refers to the function of the encoder–decoder model. Noted that it consists of three parts: encode block *encode*(), feature transformer layer $f_t(.)$, and decode block *decode*(). In particular, the reference feature extractor can be formulated as,

$$\boldsymbol{f}^{Ref} = decode(f_t(encode(\boldsymbol{I}^{Ref}))). \tag{3}$$

In order to compress the reference image into high dimensional feature space, the encoder–decoder network follows a U-shaped design as the U-Net (Ronneberger, Fischer, & Brox, 2015).

Different from the existing studies that transform the local features from the reference image into the SR image, we employ spatial transformer networks (STN) (Jaderberg et al., 2015) to increase the invariance of the affine transformation in a CNN network as the feature transformer layer. Learnable localization and grid can be served as an affine transformation matrix. An image sampling function is applied to sample the feature maps, and sampled feature maps are further merged into a spatial transformer. In this work, we leverage the STN block for the transcoding process.

The spatial transformer network is defined as follows:

$$\begin{bmatrix} \boldsymbol{x}^{output} \\ \boldsymbol{y}^{output} \end{bmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}^{input} \\ \boldsymbol{y}^{input} \\ 1 \end{bmatrix}, \tag{4}$$

where $(\boldsymbol{x}^{input}, \boldsymbol{y}^{input})$ are the coordinates in input feature maps, $(\boldsymbol{x}^{output}, \boldsymbol{y}^{output})$ are the coordinates in the output maps, and θ denotes 2D transformation parameters.

3.4. Enhanced image prior

Texture transferring and alignment from the Ref-HR image are the core problems of reference-based SR approaches (in both supervised and unsupervised manners). Existing RefSR approaches (Shim et al., 2020; Yang et al., 2020) rely on strongly supervised learning strategies to achieve this goal. We call this "hard update". However, these methods generally present limited performance in an unsupervised framework (see Section 5.3 for details). We propose a recurrent updating strategy, termed as "soft update" (Fig. 5). The main idea of the proposed updating strategy is to transform features to the latent space as well as update the initial noise at the same time so as to achieve texture migration while avoiding local optimization.

We first convert feature maps of the reference image into the latent space that carries the structural information. Then, we code the latent space to generate images with the same semantic targets and attributes. We introduce prior information through temporal patterns and update the noise iteration-by-iteration. The proposed updating strategy can be formulated as,

$$\begin{aligned} \boldsymbol{n}_{i+1} &= \Psi(\boldsymbol{n}_i, \boldsymbol{f}_i^{Ref}) \\ &= \boldsymbol{n}_i + \alpha \cdot \boldsymbol{f}_i(x), \, and \; \boldsymbol{n}_1 = \boldsymbol{n}_{init}, \end{aligned} \tag{5}$$

where i is the number of iterations, $\boldsymbol{f}^{Ref}$ is the feature map of the reference image, and $\boldsymbol{f}(x)$ is the hidden space matrix generated by the Gaussian function that transfers the enhanced image prior from the reference image. In particular, the input noise is updated as:

$$\boldsymbol{f}_i(x) = \frac{1}{\sqrt{2\pi}\,\mathrm{std}\left(\boldsymbol{f}_i^{Ref}\right)} \times \exp(-\frac{\left(x-\mathrm{mean}\left(\boldsymbol{f}_i^{Ref}\right)\right)^2}{2\mathrm{std}\left(\boldsymbol{f}_i^{Ref}\right)^2}), \tag{6}$$

where *std*(.) denotes the standard deviation function, and *mean*(.) denotes the mean function.

3.5. Loss function

The goal of supervised SR is to generate an SR image/patch from the corresponding LR one by minimizing the error in HR space (Lu et al., 2019). Existing methods take pixel-wise error as the loss function for higher objective assessment. Among them, mean square error (MSE) and mean absolute error (MAE) are the most widely used loss function. Considering the lack of ground truth in our unsupervised approach, we downsample the SR image $\boldsymbol{I}^{SR}$ via the Lanczos resampling (Turkowski, 1990) function as (Haut et al., 2018), and minimize the MSE between the downsampled version of the SR image with $\boldsymbol{I}^{LR}$ in LR-size domain. This process can be described as follows:

$$\begin{aligned} \mathcal{L}(\theta, S) &= \left\| \boldsymbol{I}^{LR} - down(\boldsymbol{I}^{SR}) \right\|_2 \\ &= \left\| \boldsymbol{I}^{LR} - \boldsymbol{I}^{LR'} \right\|_2, \end{aligned} \tag{7}$$

where $down(.)$ is the Lanczos resampling function. $\boldsymbol{I}^{LR'}$ denotes the LR-size version SR image, θ denotes the parameters in the proposed method, and S denotes the training data.

Table 1
Comparisons of the average number of matching points in a four-level scheme.

Data	Number of matching points	Levels
Day1	26.07	L2
Day2	24.88	L3
Day3	23.10	L4
Day4	27.46	L1

4. Experiments

In this section, we describe the public datasets used in the experiment, detail the configuration of the proposed method, and present the evaluation results of our method.

4.1. Datasets

The Draper dataset[1] is a publicly available benchmark of Kaggle competition for remote sensing images, including a total of 324 scenarios with five images in each scenario. The photographs were captured from a plane, a reasonable facsimile for satellite images taken at different times. The images in the Draper dataset are 3099 × 2329 pixels. We randomly select two sets of images (five images in a set) from this dataset, as shown in Fig. 4, and name them "Day 1", "Day 2" ,"Day 3" ,"Day 4" , and "Day 5". Note that the images may not be at the same time each day.

We take 41 representative images (airport, highway, parking, etc.) from "Day 5" and define a four-level scheme from high to low, i.e., L1, L2, L3, and L4, according to the number of SIFT (Lowe, 1999) feature matches. Table 1 presents the matching results of the Draper dataset, where we regrade the dataset according to the feature similarity based on the number of SIFT matching points. For images in each set (five images in a set), we crop 192 × 192 patches from "Day 5" as the ground truth and the corresponding images for the remaining four days as references.

4.2. Implementation details

All models presented in this paper are trained with Adam (Kingma & Ba, 2015) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1\mathrm{e}-8$. Each mini-batch contains one 192 × 192 noise map and the reference patches with the same size. We initialize the learning rate as $1e-4$. We set the spectral bands of noise $C' = 32$. These experiments run on a desktop with two NVIDIA GTX 2080Ti GPUs, 3.60 GHz Intel Core i7-7820X CPU, and 32 GB memory. We implement the proposed method using PyTorch 1.1.0 library[2] with Python 3.5.6 under Ubuntu 18.04, CUDA 10.1, and CUDNN 7.5 systems. We train the model over 10,000 iterations, until the model converges by minimizing the loss function equation (7).

Evaluation measures. Seven widely used image quality assessment (IQA) indices are employed to evaluate the performance, including peak signal to noise ratio (PSNR) (Huynh-Thu & Ghanbari, 2008), structural similarity (SSIM) (Wang, Bovik, Sheikh, & Simoncelli, 2004), feature similarity (FSIM) (Zhang, Zhang, Mou, & Zhang, 2011), visual information fidelity (VIF) (Sheikh & Bovik, 2006), erreur relative globale adimensionnelle de synthese (ERGAS) (Veganzones et al., 2015), spectral angle mapper (SAM) (Yuhas, Goetz, & Boardman, 1992), and learned perceptual image patch similarity (LPIPS) (Zhang, Isola, Efros, Shechtman, & Wang, 2018). PSNR, SSIM, and FSIM are three widely adopted quality indices in SR tasks, while SAM and ERGAS are commonly used quantitative image quality indices in fusion tasks. VIF measures information fidelity by computing the distortion between the ground truth and the reconstructed results. LPIPS represents human perception similarity. PSNR, SSIM, FSIM, VIF, and ERGAS are evaluated in the Y channel of transformed YCbCr space, while the others are evaluated in the RGB space. Note that the low resolution images are generated by downsampling the ground truth with a scale factor of ×4 or ×8 via a Bicubic function (the Matlab function imresize).

4.3. Comparison with state-of-the-art unsupervised methods

To verify the effectiveness of the proposed method, we compare the results from EIP with those from state-of-the-art unsupervised SR methods, including DIP (Lempitsky et al., 2018),[3] DIP-ResNet (Lempitsky et al., 2018), URSSR (Haut et al., 2018), and ZSSR (Assaf Shocher, 2018),[4] among which URSSR (Haut et al., 2018) is acknowledged to achieve the state-of-the-art performance in remote sensing image SR. DIP (Lempitsky et al., 2018) achieves state-of-the-art visual quality, even compared with some supervised SR algorithms. All experiments are performed with the scale factors of ×4 and ×8.

Table 2 shows the average performance of EIP and other competing methods in PSNR, SSIM, FSIM, VIF, ERGAS, SAM and LPIPS for ×4 and ×8 on the draper dataset, where the bold represents the best performance. ZSSR (Assaf Shocher, 2018) is an LR-input-based method and thus can achieve stable performance. We can observe that the proposed EIP framework is considerably superior to all selected competitive approaches. On average, the PSNR and SSIM of the proposed EIP framework for scale factor ×4 are 1.26 dB and 0.0293 higher than the second-best method, respectively. With a scale factor ×8, the average PSNR in EIP is 27.10 dB, 0.50 dB higher than URSSR (Haut et al., 2018) and 0.72 dB higher than and ZSSR (Assaf Shocher, 2018).

Fig. 6 presents several reconstructed images and the corresponding error maps of selected representative scenarios (*e.g.*, airplanes, roads, and ships) on the Draper dataset with upsampling factors ×4. Fig. 7 presents results with upsampling factors ×8. From Table 2, we observe that DIP-ResNet (Lempitsky et al., 2018) fails to restore clear and sharp details. Therefore, we do not present their qualitative results. The visual comparison suggests that the results from ZSSR (Assaf Shocher, 2018) contain shape edges (high-frequency information), however, with a lot of artifacts. This may be due to the fact that ZSSR (Assaf Shocher, 2018) directly extends the model of pseudo training pair to the original space. URSSR (Haut et al., 2018) and DIP (Lempitsky et al., 2018) employ a similar learning strategy, while an hourglass-shaped network (URSSR) fills pixels by sampling to effectively alleviate the local optimization problem in DIP. In comparison, our method retains not only sharper edges but also finer details than the other methods (please refer to the regions marked with yellow boxes).

4.4. Performance and model trade-offs

To validate the advantages of the proposed EIP framework, we design several other unsupervised reference-based networks for comparison, as shown in Fig. 8. Their reconstructed images and the corresponding error maps are shown in Fig. 9.

Structure 1 The first designed structure (Fig. 8(a)) generates HR images by fusing the reference and the LR images in the feature domain, the same approach as the majority of existing reference-based methods. We used an encoder–decoder model to directly encode images and use a CNN model to generate

[1] https://www.kaggle.com/c/draper-satellite-image-chronology/data
[2] https://pytorch.org
[3] https://github.com/DmitryUlyanov/deep-image-prior
[4] https://github.com/assafshocher/ZSSR

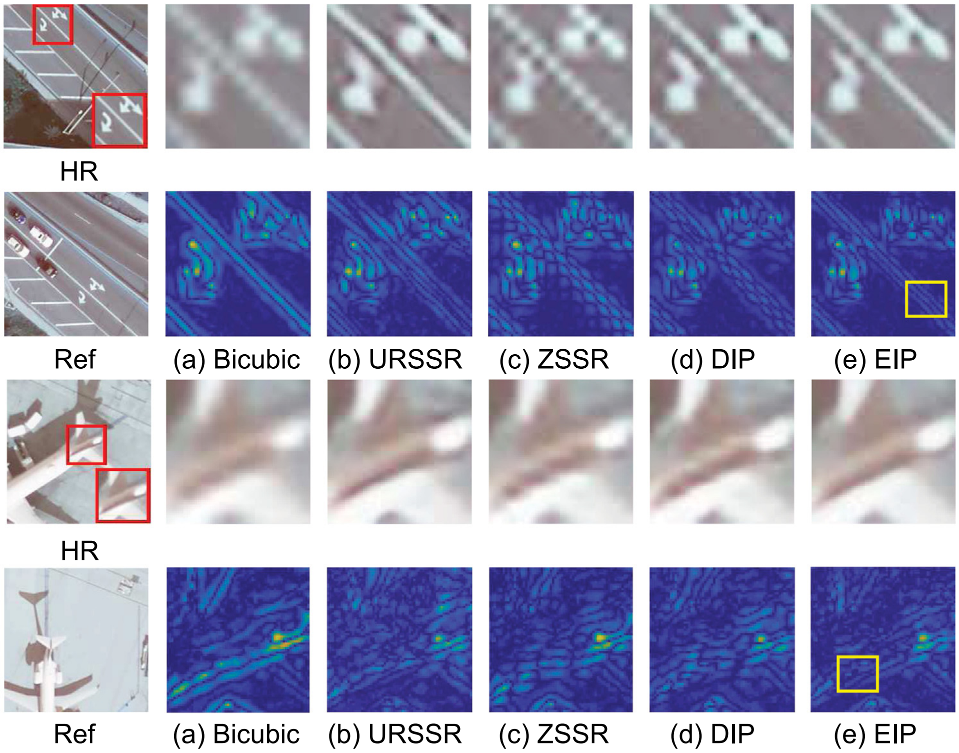

Fig. 6. Visual comparison among different SR method on draper dataset with scale factor × 4. From left to right, they are the ground truth, results of Bicubic, URSSR (Haut et al., 2018), ZSSR (Assaf Shocher, 2018), DIP (Lempitsky et al., 2018), and the proposed EIP method. The bottom images are the reconstruction error maps of the corresponding methods. The yellow box shows significant improvements.

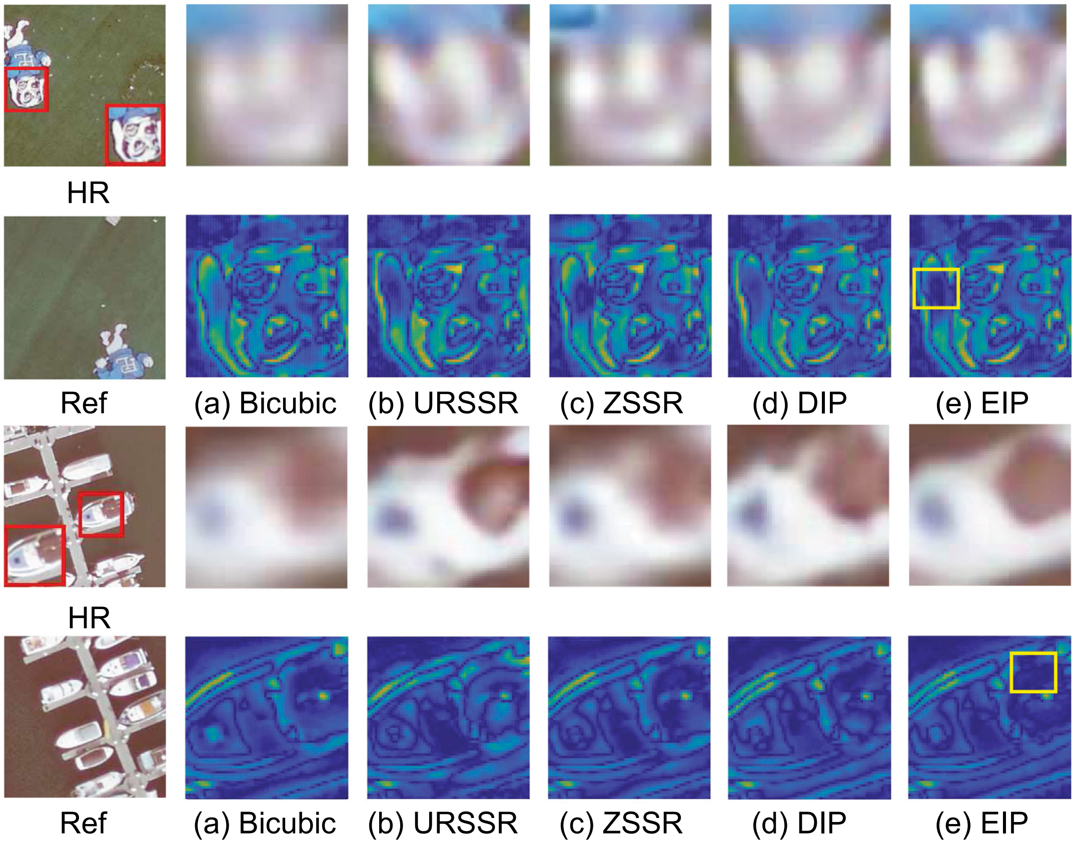

Fig. 7. Visual comparison among different SR method on Draper dataset with scale factor × 8. From left to right, they are the ground truth, results of Bicubic, URSSR (Haut et al., 2018), ZSSR (Assaf Shocher, 2018), DIP (Lempitsky et al., 2018), and the proposed EIP method. The bottom images are the reconstruction error maps of the corresponding methods. The yellow box shows significant improvements.

Table 2
Average quantitative comparisons of different approaches on the Draper dataset. Bold indicates the best result. ↑ indicates that the larger the value, the better the performance, and ↓ indicates that the smaller the value, the better the performance.

Method	Scale	PSNR ↑	SSIM ↑	FSIM ↑	VIF ↑	ERGAS ↓	SAM ↓	LPIPS ↓
Bicubic	×4	28.98	0.7838	0.8135	0.4046	1.4966	0.6696	0.3772
DIP-ResNet (Lempitsky et al., 2018)		15.20	0.2361	0.5128	0.0878	6.7941	5.5102	0.8108
ZSSR (Assaf Shocher, 2018)		29.43	0.7851	0.8369	0.3508	1.3955	0.9616	0.2377
URSSR (Haut et al., 2018)		29.57	0.8055	0.8370	0.4016	1.4015	0.6657	0.3016
DIP (Lempitsky et al., 2018)		29.66	0.8129	0.8496	0.3921	1.3745	0.7489	0.2531
EIP		**30.89**	**0.8429**	**0.8689**	**0.4607**	**1.1995**	**0.6215**	**0.2240**
Bicubic	×8	26.06	0.6628	0.7109	0.2424	2.0965	1.0476	0.6119
DIP-ResNet (Lempitsky et al., 2018)		13.33	0.0980	0.4220	0.0229	8.4052	9.0296	0.8266
ZSSR (Assaf Shocher, 2018)		26.38	0.6686	0.7342	0.2114	2.0243	1.0777	0.5829
URSSR (Haut et al., 2018)		26.60	0.6737	0.7388	0.2010	1.9436	1.2973	0.4588
DIP (Lempitsky et al., 2018)		26.07	0.6732	0.7481	0.2030	2.0801	1.1024	0.4982
EIP		**27.10**	**0.7025**	**0.7534**	**0.2649**	**1.8498**	**0.9669**	**0.4455**

All the codes are provided by the authors, and the parameters are set according to models with the best performance in their papers.

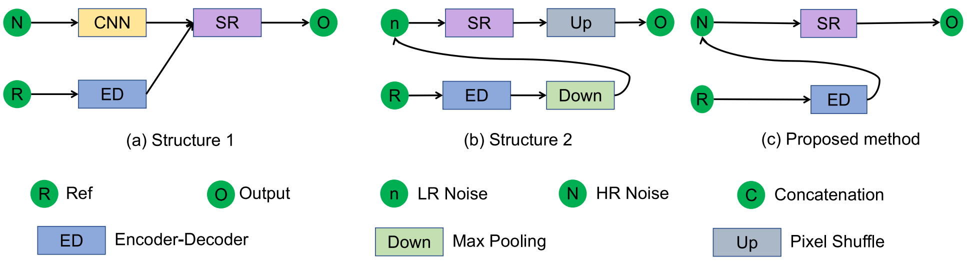

Fig. 8. Experiments for validating the advantages of the proposed structure. (a) Structure 1: fuse the reference image and the LR image in the feature domain. (b) Structure 2: input LR-size noise maps. (c) Structure 3: the proposed EIP method.

and concatenate feature maps. Following the above steps, the SR network is used to reconstruct the image.

We found that the results from Structure 1 contain a lot of undesired artifacts, as shown in Fig. 9(a) (highlighted by the blue box). The middle images are the enlarged version of the red box in the top images. The bottom images are the reconstruction error maps of the corresponding methods. Owing to the different synthetic bands in LR images and the HR reference image, the same object can present varying contrasts. We observe that the reconstructed images from Structure 1 tend to inherit the textures and other details from the reference image (the red box in the second row of Fig. 9(a)). Different from the existing supervised reference-based methods, the unsupervised framework lacks the original HR image, precluding the calculation of pixel-by-pixel errors. The weak supervising nature of this model increases the difficulty of exchanging information in high-level feature dimensions.

Structure 2 Existing SR algorithms are based on LR spaces, leading to their excellence in reducing the demand for computational resources. Therefore, we aim to explore the utility of LR inputs in this framework. In this experiment, we resize the random noise maps ($C' \times H \times W$) as the input of the network, as shown in Fig. 8(b). We use the maxpool function to down-sample reference feature maps and update the input noise maps in the latent space. Finally, we employ the pixel-shuffle layer for up-sampling, which can be formulated as,

$$\mathcal{L}_{S2}(\theta, S) = \left\| \boldsymbol{I}^{LR} - down(up(\boldsymbol{I}^{SR})) \right\|_2, \tag{8}$$

where $down(.)$ denotes the Lanczos resampling function, and $up(.)$ denotes the pixel-shuffle layer.

Experimental results and corresponding error maps in Fig. 9(b) reveal notable checkerboard phenomenon in the images, greatly affecting the visual experience. As discussed in Section 3.5, the process of network optimization, like the one in supervised super-resolution tasks, is often viewed as an ill-posed problem. The up-sampling component of the SR network improves the efficiency of the network, however, making the model prone to local optimum.

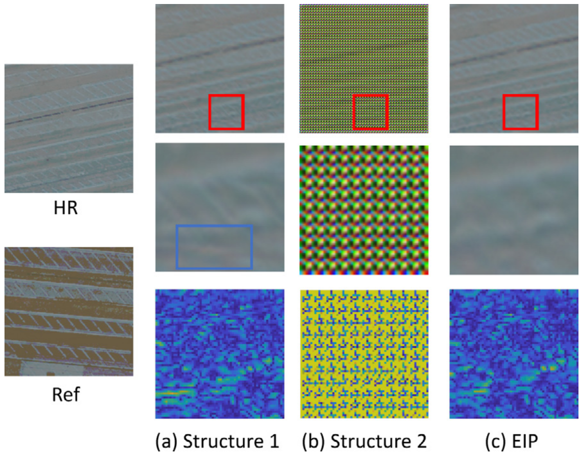

Fig. 9. The reconstructed images and the corresponding error maps for structure described in Fig. 8. The middle images are the enlarged version of the red boxes in the top images. The bottom images are the reconstruction error maps of the corresponding methods. The blue box highlights notable artifacts. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In order to further analyze the effectiveness of Structure 2, we introduce the total variation (TV) loss function (Rudin, Osher, & Fatemi, 1992) is introduced that encourages spatial smoothness in the generated image. The total loss function is the sum of the TV loss and $\mathcal{L}_{S2}$:

$$\mathcal{L}_{total} = \mathcal{L}_{S2} + \lambda \mathcal{L}_{TV}, \tag{9}$$

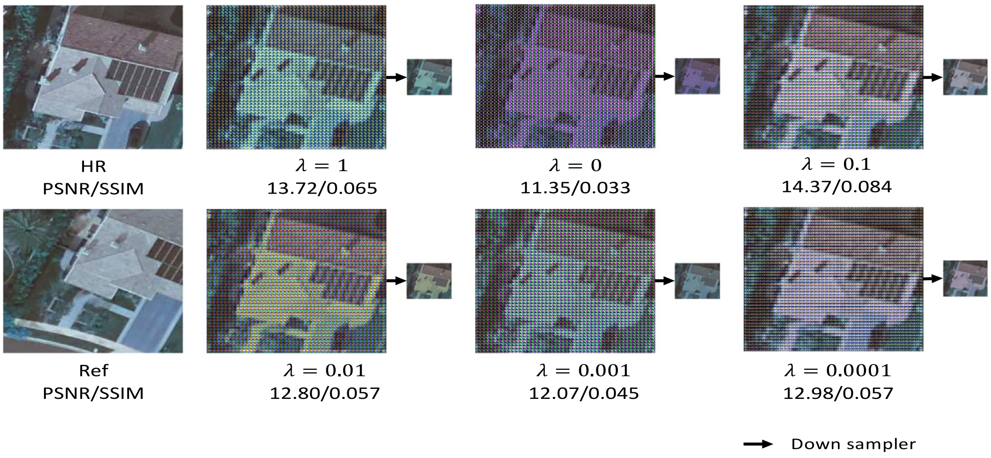

Fig. 10. Visualization of results from selected methods to demonstrate the impact of the loss function.

Table 3
Average quantitative comparisons among different approaches with scale factor ×4 on the Draper dataset. Bold indicates the best result. ↑ indicates that the larger the value, the better the performance, and ↓ indicates that the smaller the value, the better the performance.

Method	PSNR ↑	SSIM↑	FSIM ↑	VIF ↑	ERGAS ↓	SAM ↓	LPIPS ↓
DIP (Lempitsky et al., 2018)	29.66	0.8129	0.8496	0.3921	1.3745	0.7489	0.2531
Structure 1	29.86	0.8017	0.8436	0.3913	1.3368	1.0822	0.2874
Structure 2	12.16	0.0446	0.3327	0.0758	10.4490	36.1206	0.8528
EIP	**30.89**	**0.8429**	**0.8689**	**0.4607**	**1.1995**	**0.6215**	**0.2240**

where $\mathcal{L}_{TV}$ is the TV loss, and λ is used to balance the contributions of different losses. As shown in Fig. 10, we train Structure 2 with λ ranging from 1e−4 to 10. Although additional regularization can effectively improves the performance, the checkerboard phenomenon remains notable. However, there is no checkerboard phenomenon in the down-sampling version SR images. The results of this experiment suggest that a coupled up-sampling function $up(.)$ and down sampling function $down(.)$ can offset the information loss caused by the $up(.)$ function.

The proposed method (EIP) Fig. 8(c) presents the reconstructed images and corresponding error maps of the proposed EIP. We observe that the reconstructed results from EIP contain more texture information. From these error maps, the proposed method achieves the best fidelity in terms of detail recovery.

To quantitatively compare reconstructed results quantitatively, we report the average PSNR, SSIM, FSIM, VIF, ERGAS, SAM, and LPIPS of the structures described in Fig. 8 (Table 3). We observe that the proposed EIP method greatly outperforms other algorithms in all evaluation metrics. On average, the PSNR of Structure 2 is 0.2 dB higher than that of DIP, showing that, despite the artifacts in the reconstructed images, the features from the reference image were learned and transferred. The proposed method outperforms Structure 2, evidenced by the 1.03 dB improvement in the average PSNR.

4.5. Results on general images

Previous experiments are conducted on the Draper datasets with remotely sensed images. To evaluate the expendability of the proposed EIP, we conduct additional experiments on the CUFED dataset (Lempitsky et al., 2018), a widely used database that contains general images with a size of 160 × 160 pixels.

In this experiment, we randomly select 50 images from the CUFED dataset (Lempitsky et al., 2018). Table 4 reports the average performance of EIP, DIP, URSSR, ZSSR, Bicubic in four selected IQAs (i.e., PSNR, SSIM, FSIM, and VIF), where bold represents the best performance (considering the poor performance of DIP-ResNet, we do not present its results, Lempitsky et al., 2018).

Table 4
Quantitative comparisons of different approaches from CUFED dataset with four IQAs. Bold indicates the best result. ↑ indicates that the larger the value, the better the performance.

Method	Scale	PSNR ↑	SSIM ↑	FSIM ↑	VIF ↑
Bicubic	×4	25.36	0.6882	0.7952	0.3548
ZSSR (Assaf Shocher, 2018)		25.80	0.7156	0.8270	0.3630
URSSR (Haut et al., 2018)		25.82	0.7087	0.8404	0.3322
DIP (Lempitsky et al., 2018)		25.59	0.7162	0.8358	0.3482
EIP		**26.39**	**0.7357**	**0.8436**	**0.3931**
Bicubic	×8	23.10	0.5745	0.6957	0.2262
ZSSR (Assaf Shocher, 2018)		23.36	0.5857	0.7262	0.2084
URSSR (Haut et al., 2018)		22.82	0.5473	0.7418	0.1599
DIP (Lempitsky et al., 2018)		23.16	0.5981	0.7353	0.2202
EIP		**23.82**	**0.6107**	**0.7365**	**0.2526**

All the codes are provided by their authors, and parameters are set according to models with the best performance in their papers.

Table 4 shows that the proposed method significantly outperforms other methods with ×4 and ×8. The PSNR of the proposed EIP is 0.59 dB higher than the second-best method with upsampling factor ×4. With a scale factor ×8, the average PSNR in EIP is 23.82 dB.

Figs. 11 and 12 show the qualitative comparison among the proposed EIP, URSSR (Haut et al., 2018), ZSSR (Assaf Shocher, 2018), and DIP (Lempitsky et al., 2018) with upsampling factors ×4 and ×8. We observe that results from URSSR (Haut et al., 2018) and DIP (Lempitsky et al., 2018) present noticeable artifacts. In addition, there exists a notable problem of local optimization from the results of DIP (Lempitsky et al., 2018) with upsampling factor ×8. ZSSR (Assaf Shocher, 2018) is able to recover sharp contours, however, with dirty effects (*e.g.*., the deformed wall lamp). In comparison, results from EIP contain sharp edges and clear image content (please refer to the regions marked with yellow boxes), suggesting the expendability of the proposed method on general images.

Table 5
Quantitative comparisons among structures described in Fig. 8. ↑ indicates that the larger the value, the better the performance, and ↓ indicates that the smaller the value, the better the performance.

Method	PSNR ↑	SSIM↑	FSIM ↑	VIF ↑	ERGAS ↓	SAM ↓	LPIPS ↓
Structure 1-L2	29.81	0.7989	0.8418	0.3874	1.3466	1.0923	0.2876
Structure 1-L3	29.76	0.7989	0.8407	0.3855	1.3551	1.1044	0.2915
Structure 1-L4	29.77	0.7985	0.8406	0.3841	1.3519	1.0680	0.2948
Structure 1-L1	29.86	0.8017	0.8436	0.3913	1.3368	1.0822	0.2874
EIP-L2	30.88	0.8435	0.8692	0.4612	1.1984	0.6226	0.2243
EIP-L3	30.88	0.8433	0.8692	0.4609	1.1992	0.6223	0.2247
EIP-L4	30.88	0.8431	0.8688	0.4598	1.1988	0.6241	0.2244
EIP	30.89	0.8429	0.8689	0.4607	1.1995	0.6215	0.2240

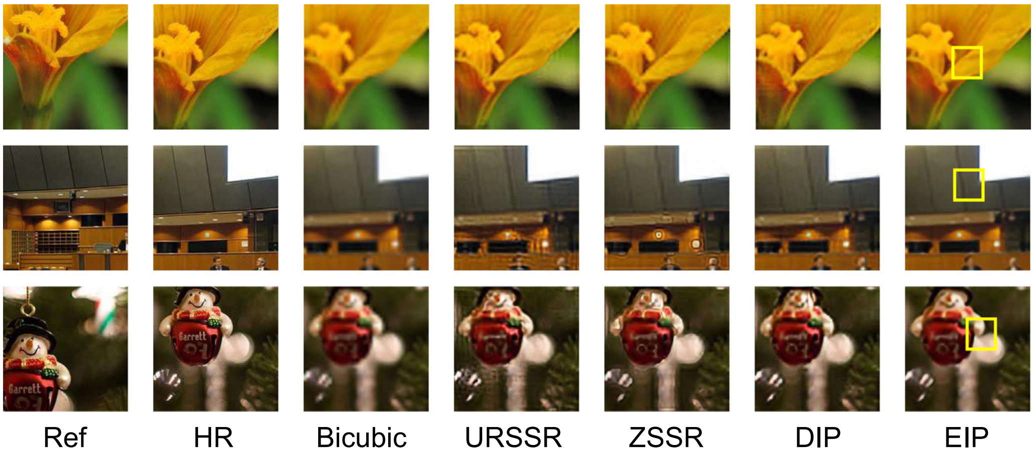

Fig. 11. Reconstruction results on UCFED dataset with the scale factor of 4. The yellow box shows significant improvements.

4.6. Results on real satellite images

To test the capability of EIP in handling images in a real-world scenario, we conduct experiments on the SuperView-1 satellite images[5] (0.5 m resolution) to further illustrate the applicability of the proposed algorithm. SuperView-1 is a satellite constellation network that includes four satellites with an orbital altitude of 530 km. The reference images were uploaded on November 20, 2017, and the LR video images were taken on January 16, 2018. These images mainly cover the region of Malaysia. We crop the center region of the LR image to derive a sub-image with 255 × 255 pixels while the corresponding reference Ref-HR image is cropped to 765 × 765 pixels.

Since no ground-truth images are available in this experiment, we introduce two non-reference IQA metrics: the spatial–spectral entropy-based quality (Liu, Liu, Huang, & Bovik, 2014) (SSEQ, 0 as the best value) and the naturalness image quality evaluator (Mittal, Soundararajan, & Bovik, 2012) (NIQE, 0 as the best value), to calculate the sharpness of the SR images. In order to verify the superiority of the proposed EIP framework, a supervised single image SR algorithm (RCAN) (Zhang et al., 2018) trained on the Draper dataset is included as a comparison.

Fig. 13 shows the reconstructed images and the performance of two IQAs. From the reconstructed images, our proposed method achieves a better performance in both SSEQ and NIQE. The reconstructed results from RCAN (Zhang et al., 2018) reveal over-smoothed details. The results from EIP suggest that it successfully restores the dome of the building with fewer jagged lines and few ringing artifacts. The great performance of EIP on satellite images with large time spans between the input and reference images shows the great applicability of our method in handling unknown degradation scenarios. On the other hand, the results prove that the performance of the supervised SR algorithm is limited for images that are captured in an unknown acquisition process, coinciding with studies by Levin et al. (2009), Shocher et al. (2018) and Soh et al. (2020).

5. Ablation experiments

5.1. Ablation study on reference similarity

The above results indicate that reference images play an important role in the performance of the proposed method. In this section, we explore the influence of reference images on algorithm performance. Zhang et al. (2019) concluded that the performance of the feature transformation based method is positively correlated with reference similarity. To investigate the gap between the feature transformation based method and the proposed method, we conduct an experiment by replacing the reference image. Meanwhile, we compare the proposed method with the "Structure 1" described in Section 4.4, which can be seen as a representative of the feature transformation-based method in an unsupervised framework.

Table 5 lists the average performance of "Structure 1" and the proposed method with reference images at four different similarity levels (see Section 4.1 for details) under ×4 as the scale factor. Metrics for evaluation include PSNR, SSIM, FSIM, VIF, ERGAS, SAM, and LPIPS. It can be observed that the performance of the feature transformation-based method is closely related to the reference similarity level, even in an unsupervised framework. With the increase of the similarity, the PSNR/SSIM of "Structure 1" increases accordingly. In comparison, similarity levels play a weak role in the performance of the proposed EIP framework, demonstrating the strong adaptiveness of EIP, as it realizes the implicit information transmission instead of simple texture migration.

[5] http://www.spacewillinfo.com/Satellite/Satellite/superview/#main

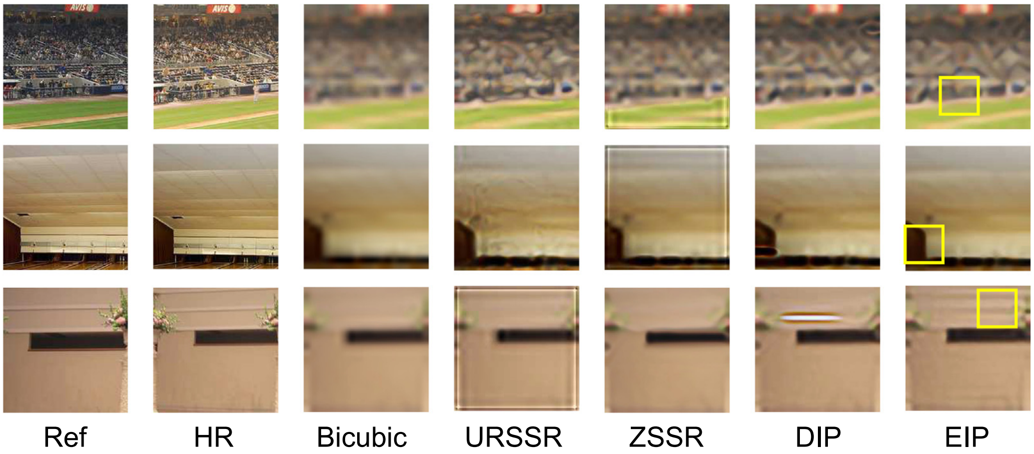

Fig. 12. Reconstruction results on UCFED dataset with the scale factor of 8. The yellow box shows significant improvements.

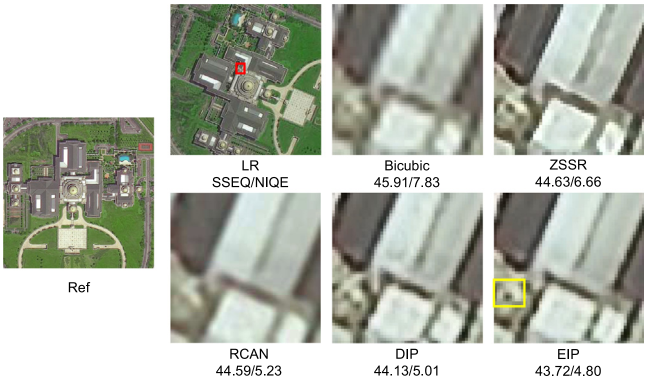

Fig. 13. Comparison of reconstructed HR images obtained via various methods with a scale factor of 3 on video satellite images. The yellow box shows significant improvements.

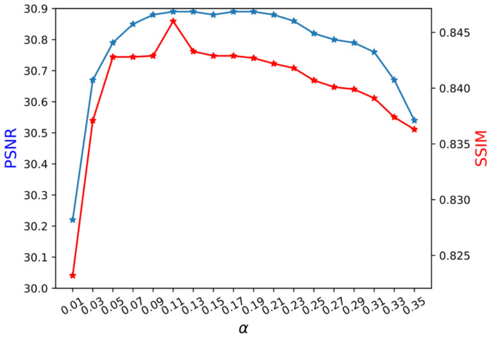

Fig. 14. Effect of hyperparameter α on model performance. When $\alpha = 0.11$, the proposed method achieves the best performance.

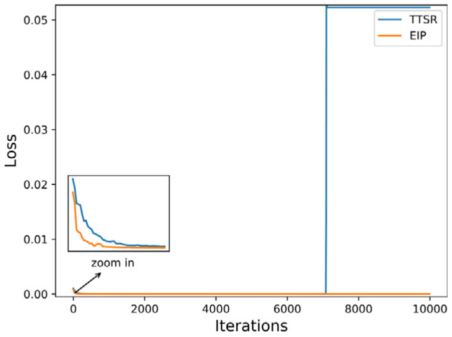

Fig. 15. Curves of the training procedure between proposed EIP and TTSR (Yang et al., 2020), a state-of-the-art RefSR based SR algorithm. These curves show that it is difficult for TTSR (Yang et al., 2020) to converge, while the proposed EIP can converge in a stable manner.

5.2. Ablation study on hyperparameter

The framework of the proposed method implies that hyperparameter α potentially determines the model performance as it plays a balancing role in the image prior enhancing. To study the effectiveness of hyperparameter α and to derive α that leads to the best performance of the model, we train EIP with α that increases from between 0.01 to 0.35 with 0.02 as an interval. Other configurations remain unchanged.

The quantitative results by PSNR/SSIM with respect to different α are shown in Fig. 14. We notice that the performance of EIP increases when α increases from 0.01 to 0.11, and it becomes stable when α increase from 0.11 to 0.21. As α further increases (from 0.21 to 0.35), the performance reduces. We conclude that the proposed method achieves the best performance when $\alpha = 0.11$. Therefore, we choose $\alpha = 0.11$ as our final setting.

5.3. Comparison with the RefSR method

To further verify the superiority of the proposed model, curves of loss values per each iteration on the Draper dataset are presented in Fig. 15. We implement TTSR (Yang et al., 2020), a state-of-the-art RefSR based SR algorithm,[6] with the loss function Eq. (7) in an unsupervised framework. These curves show that it is difficult (Yang et al., 2020) for TTSR to converge, while the proposed EIP can converge in a stable manner. The fast convergence in the early stage of TTSR (Yang et al., 2020) can be explained by the fact that its takes low-resolution (LR) images as input. The original version of TTSR (Yang et al., 2020) benefits from its capability in texture alignment and transferring under strongly supervised learning. However, as demonstrated in the experiments of Structure 2 (see Section 4.4), LR-sized input results in the checkerboard phenomenon, which inevitably leads to the increase of the loss. Therefore, the loss of TTSR follows a two-stage pattern. Thus, it is difficult to establish a stable unsupervised TTSR model (Yang et al., 2020).

[6] https://github.com/researchmm/TTSR

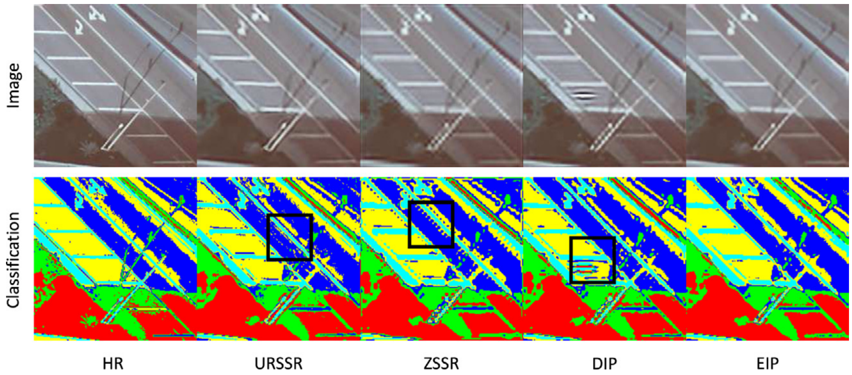

Fig. 16. Reconstructed SR images their corresponding classification results using the ISODATA classification method. Black boxes emphasize areas with notable artifacts or blurring.

5.4. Effectiveness for post-processing

We employ the iterative self-organizing data analysis techniques algorithm (ISODATA), a classical unsupervised semantic segmentation for satellite images, to evaluate the results from different SR methods. The number of the classes is set to five. We set the maximum iteration to five times as Chang, Yan, Fang, Zhong, and Liao (2018). Fig. 16 shows the results of reconstructed SR images and their corresponding classification results. The black boxes emphasize the areas with notable artifacts and blurring. Specifically, results from URSSR (Haut et al., 2018), an hourglass model, show notable artifacts and missing information, presumably due to the repeated procedure of downsampling and upsampling. As for ZSSR (Assaf Shocher, 2018), its simple feature maps and shallow convolution blocks bring limited performance limitation. While DIP (Lempitsky et al., 2018) constructs high resolution images using fixed noise as input, the initialized noise is noticeable on the resultant images. In comparison, the classification result of the proposed EIP is visually the closest to the high resolution image, suggesting the superiority of the proposed method.

6. Conclusion

In this paper, we propose **a new unsupervised learning framework, Enhanced Image Prior (EIP), which achieves high performance in SR tasks without low/high resolution image pairs**. First, we feed random noise maps into a designed GAN for satellite image SR reconstruction. Then, **we convert the reference image to latent space as the enhanced image prior**. Finally, we update the input noises via a **recurrent updating strategy and further transfer the texture and structured information from the reference image**. Our experiments using both simulated and real-world data reveal the competitive performance of the proposed approach when super-resolving satellite images. Future efforts can focus on exploring an LR-input-based framework, serving as a potential replacement of the proposed architecture for efficient reconstruction. We believe that the proposed method can be generalized to other remote sensing image enhancement tasks, such as cloud removal and denoising.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahn, N., Yoo, J., & Sohn, K. -A. (2020). SimUSR: A simple but strong baseline for unsupervised image super-resolution. In *Computer vision and pattern recognition workshops* (pp. 474–475).

Assaf Shocher, M. I. (2018). "Zero-Shot" super-resolution using deep internal learning. In *IEEE conference on computer vision and pattern recognition*.

Cao, F., Yao, K., & Liang, J. (2020). Deconvolutional neural network for image super-resolution. *Neural Networks*, *132*, 394–404.

Cao, S., et al. (2018). Main processes for OVS-1A and OVS-1B: From manufacturer to user. *Journal of Computer and Communications*, *6*(11), 126.

Chang, Y., Yan, L., Fang, H., Zhong, S., & Liao, W. (2018). HSI-DeNet: Hyperspectral image restoration via convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(2), 667–682.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Neural information processing systems* (pp. 2172–2180).

Dai, T., Cai, J., Zhang, Y., Xia, S. -T., & Zhang, L. (2019). Second-Order attention network for single image super-resolution. In *IEEE conference on computer vision and pattern recognition* (pp. 11065–11074).

Dong, C., Loy, C. C., He, K., & Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 295–307.

Dong, R., Zhang, L., & Fu, H. (2021). RRSGAN: Reference-based super-resolution for remote sensing image. *IEEE Transactions on Geoscience and Remote Sensing*.

Freedman, G., & Fattal, R. (2011). Image and video upscaling from local self-examples. *ACM Transactions on Graphics*, *30*(2), 1–11.

Geng, T., Liu, X. -Y., Wang, X., & Sun, G. (2021). Deep shearlet residual learning network for single image super-resolution. *IEEE Transactions on Image Processing*, *30*, 4129–4142.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Neural information processing systems* (pp. 2672–2680).

Gu, Y., Wang, T., Jin, X., & Gao, G. (2020). Detection of event of interest for satellite video understanding. *IEEE Transactions on Geoscience and Remote Sensing*, 1–12.

Guo, Y., Chen, J., Wang, J., Chen, Q., Cao, J., Deng, Z., et al. (2020). Closed-Loop matters: Dual regression networks for single image super-resolution. In *IEEE conference on computer vision and pattern recognition* (pp. 5407–5416).

Haut, J. M., Fernandez-Beltran, R., Paoletti, M. E., Plaza, J., Plaza, A., & Pla, F. (2018). A new deep generative network for unsupervised remote sensing single-image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, *56*(11), 6792–6810.

He, Z., He, D., Mei, X., & Hu, S. (2019). Wetland classification based on a new efficient generative adversarial network and Jilin-1 satellite image. *Remote Sensing*, *11*(20), 2455.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Huang, J. -B., Singh, A., & Ahuja, N. (2015). Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5197–5206).

Huynh-Thu, Q., & Ghanbari, M. (2008). Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*, *44*(13), 800–801.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448–456).

Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. (2015). Spatial transformer networks. In *Neural information processing systems* (pp. 2017–2025).

Jiang, K., Wang, Z., Yi, P., Jiang, J., Xiao, J., & Yao, Y. (2018). Deep distillation recursive network for remote sensing imagery super-resolution. *Remote Sensing*, *10*(11), 1700.

Jiang, K., Wang, Z., Yi, P., Wang, G., Lu, T., & Jiang, J. (2019). Edge-enhanced GAN for remote sensing image superresolution. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(8), 5799–5812.

Kim, J., Lee, J. K., & Lee, K. M. (2016). Accurate image super-resolution using very deep convolutional networks. In *IEEE conference on computer vision and pattern recognition* (pp. 1646–1654).

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations*.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., et al. (2017). Photo-Realistic single image super-resolution using a generative adversarial network. In *IEEE conference on computer vision and pattern recognition* (pp. 105–114).

Lei, S., Shi, Z., & Zou, Z. (2019). Coupled adversarial training for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(5), 3633–3643.

Lempitsky, V., Vedaldi, A., & Ulyanov, D. (2018). Deep image prior. In *IEEE conference on computer vision and pattern recognition* (pp. 9446–9454).

Levin, A., Weiss, Y., Durand, F., & Freeman, W. T. (2009). Understanding and evaluating blind deconvolution algorithms. In *IEEE conference on computer vision and pattern recognition* (pp. 1964–1971). IEEE.

Li, C., Shao, Z., Zhang, L., Huang, X., & Zhang, M. (2021). A comparative analysis of index-based methods for impervious surface extraction using multi-seasonal sentinel-2 satellite data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Lim, B., Son, S., Kim, H., Nah, S., & Lee, K. M. (2017). Enhanced deep residual networks for single image super-resolution. In *IEEE conference on computer vision and pattern recognition workshops* (pp. 1132–1140).

Liu, H., & Cao, F. (2020). Improved dual-scale residual network for image super-resolution. *Neural Networks*, *132*, 84–95.

Liu, L., Liu, B., Huang, H., & Bovik, A. C. (2014). No-reference image quality assessment based on spatial and spectral entropies. *Signal Processing: Image Communication*, *29*(8), 856–863.

Liu, C., & Sun, D. (2011). A Bayesian approach to adaptive video super resolution. In *IEEE conference on computer vision and pattern recognition* (pp. 209–216).

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *IEEE international conference on computer vision*: *Vol. 2*, (pp. 1150–1157). IEEE.

Lu, T., Wang, J., Zhang, Y., Wang, Z., & Jiang, J. (2019). Satellite image super-resolution via multi-scale residual deep neural network. *Remote Sensing*, *11*(13), 1588.

Luo, Y., Zhou, L., Wang, S., & Wang, Z. (2017). Video satellite imagery super resolution via convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, *14*(12), 2398–2402.

Lv, X., Ming, D., Lu, T., Zhou, K., Wang, M., & Bao, H. (2018). A new method for region-based majority voting CNNs for very high resolution image classification. *Remote Sensing*, *10*(12), 1946.

Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T. S., & Shi, H. (2020). Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Computer vision and pattern recognition* (pp. 5690–5699).

Merino, M., & Nunez, J. (2007). Super-resolution of remotely sensed images with variable-pixel linear reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(5), 1446–1457.

Michaeli, T., & Irani, M. (2013). Nonparametric blind super-resolution. In *Proceedings of the IEEE international conference on computer vision* (pp. 945–952).

Mittal, A., Soundararajan, R., & Bovik, A. C. (2012). Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, *20*(3), 209–212.

Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., et al. (2020). Single image super-resolution via a holistic attention network. In *European conference on computer vision* (pp. 191–207). Springer.

Park, S. C., Park, M. K., & Kang, M. G. (2003). Super-resolution image reconstruction: A technical overview. *IEEE Signal Processing Magazine*, *20*(3), 21–36.

Prajapati, K., Chudasama, V., Patel, H., Upla, K., Ramachandra, R., Raja, K., et al. (2020). Unsupervised single image super-resolution network (usisresnet) for real-world data using generative adversarial network. In *Computer vision and pattern recognition workshops* (pp. 464–465).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D. Nonlinear Phenomena*, *60*(1–4), 259–268.

Shao, Z., Fu, H., Li, D., Altan, O., & Cheng, T. (2019). Remote sensing monitoring of multi-scale watersheds impermeability for urban hydrological evaluation. *Remote Sensing of Environment*, *232*, Article 111338.

Sheikh, H. R., & Bovik, A. C. (2006). Image information and visual quality. *IEEE Transactions on Image Processing*, *15*(2), 430–444.

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., et al. (2016). Real-Time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE conference on computer vision and pattern recognition* (pp. 1874–1883).

Shim, G., Park, J., & Kweon, I. S. (2020). Robust reference-based super-resolution with similarity-aware deformable convolution. In *IEEE conference on computer vision and pattern recognition* (pp. 8425–8434).

Shocher, A., Cohen, N., & Irani, M. (2018). "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3118–3126).

Soh, J. W., Cho, S., & Cho, N. I. (2020). Meta-Transfer learning for zero-shot super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3516–3525).

Turkowski, K. (1990). Filters for common resampling tasks. In *Graphics gems* (pp. 147–165). Academic Press Professional, Inc..

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2018). Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9446–9454).

Veganzones, M. A., Simoes, M., Licciardi, G., Yokoya, N., Bioucas-Dias, J. M., & Chanussot, J. (2015). Hyperspectral super-resolution of locally low rank images from complementary multisource data. *IEEE Transactions on Image Processing*, *25*(1), 274–288.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wang, Y., Lin, Z., Shen, X., Mech, R., Miller, G., & Cottrell, G. W. (2016). Event-specific image importance. In *IEEE conference on computer vision and pattern recognition* (pp. 4810–4819).

Wang, Y., Liu, Y., Heidrich, W., & Dai, Q. (2017). 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), 2357–2364.

Wang, J., Shao, Z., Huang, X., Lu, T., Zhang, R., & Lv, X. (2021). Spatial-temporal pooling for action recognition in videos. *Neurocomputing*.

Xu, H., Ma, J., Jiang, J., Guo, X., & Ling, H. (2020). U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, F., Yang, H., Fu, J., Lu, H., & Guo, B. (2020). Learning texture transformer network for image super-resolution. In *IEEE conference on computer vision and pattern recognition* (pp. 5791–5800).

Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., & Lin, L. (2018). Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 701–710).

Yue, H., Sun, X., Yang, J., & Wu, F. (2013). Landmark image super-resolution by retrieving web images. *IEEE Transactions on Image Processing*, *22*(12), 4865–4878.

Yue, H., Zhou, T., Jiang, Z., Yang, J., & Hou, C. (2021). Reference guided image super-resolution via efficient dense warping and adaptive fusion. *Signal Processing: Image Communication*, *92*, Article 116062.

Yuhas, R. H., Goetz, A. F., & Boardman, J. W. (1992). *Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm*.

Zhang, L., Chen, D., Ma, J., & Zhang, J. (2020). Remote-sensing image superresolution based on visual saliency analysis and unequal reconstruction networks. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(6), 4099–4115.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE conference on computer vision and pattern recognition* (pp. 586–595).

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. In *European conference on computer vision* (pp. 294–310).

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2018). Residual dense network for image super-resolution. In *IEEE conference on computer vision and pattern recognition* (pp. 2472–2481).

Zhang, Z., Wang, Z., Lin, Z., & Qi, H. (2019). Image super-resolution by neural texture transfer. In *IEEE conference on computer vision and pattern recognition* (pp. 7982–7991).

Zhang, L., Zhang, L., Mou, X., & Zhang, D. (2011). FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, *20*(8), 2378–2386.

Zheng, H., Ji, M., Wang, H., Liu, Y., & Fang, L. (2018). CrossNet: An end-to-end reference-based super resolution network using cross-scale warping. In *European conference on computer vision* (pp. 87–104).

Zhu, Y., Zhang, Y., & Yuille, A. L. (2014). Single image super-resolution using deformable patches. In *IEEE conference on computer vision and pattern recognition*: *Vol. 2014*, (pp. 2917–2924).