



Full length article

From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution

Yi Xiao^a, Qiangqiang Yuan^{a,*}, Kui Jiang^b, Jiang He^a, Yuan Wang^a, Liangpei Zhang^c

^a School of Geodesy and Geomatics, Wuhan University, Wuhan, Hubei, China

^b Cloud BU, Huawei Technologies, Hangzhou, Zhejiang, China

^c State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, Hubei, China

ARTICLE INFO

Keywords:

Blind super-resolution
Self-supervised
Contrastive learning
Remote sensing image
Deep learning

ABSTRACT

Over the past few years, single image super-resolution (SR) has become a hotspot in the remote sensing area, and numerous methods have made remarkable progress in this fundamental task. However, they usually rely on the assumption that images suffer from a fixed known degradation process, e.g., bicubic downsampling. To save us from performance drop when real-world distribution deviates from the naive assumption, blind image super-resolution for multiple and unknown degradations has been explored. Nevertheless, the lack of a real-world dataset and the challenge of reasonable degradation estimation hinder us from moving forward. In this paper, a self-supervised degradation-guided adaptive network is proposed to mitigate the domain gap between simulation and reality. Firstly, the complicated degradations are characterized by robust representations in embedding space, which promote adaptability to the downstream SR network with degradation priors. Specifically, we incorporated contrastive learning to blind remote sensing image SR, which guides the reconstruction process by encouraging the positive representations (relevant information) while punishing the negatives. Besides, an effective dual-wise feature modulation network is proposed for feature adaptation. With the guide of degradation representations, we conduct modulation on feature and channel dimensions to transform the low-resolution features into the desired domain that is suitable for reconstructing high-resolution images. Extensive experiments on three mainstream datasets have demonstrated our superiority against state-of-the-art methods. Our source code can be found at <https://github.com/XY-boy/DRSR>

1. Introduction

With the growing demand for fine-grained remote sensing applications, high-resolution remote sensing imagery is playing an indispensable and increasingly important role in downstream tasks [1], including classification [2–4], fine-scale land-cover mapping [5–7], hyperspectral applications [8–10], etc. However, breaking through the hardware limitations to obtain high-resolution remote sensing imagery is a laborious task. Fortunately, super-resolution (SR) technology provides an effective and economical alternative, which aims to reconstruct latent high-resolution (HR) images from existent low-resolution (LR) observations [11–15].

Years of effort have brought remarkable progress in remote sensing image super-resolution [11,16], especially the success of deep-learning-based methods in many areas [17–27]. Nevertheless, state-of-the-art SR models often suffer from less generalization to real-world images, although decent results have been achieved in pre-defined synthetic

data. To mitigate the domain gap between simulation and reality, a general method is urgently required to cope with images with multiple and unknown degradation factors. In Fig. 1, we display visual comparisons of classical SR and real-world SR to better understand the domain gap.

A straightforward solution is to build an external dataset that covers general degradation. Recently, several real-world datasets [28,29] in computer vision field have been proposed. They construct paired LR-HR images through careful manipulation on advanced digital devices. However, building such a dataset is time-consuming and also fragile on aerospace sensors. For example, [30] collected Sentinel-2 and Planet images to build a paired dataset. Although images acquired at different satellite platforms are georeferenced, they undergo a problem of misalignment since data acquisition moments are not strictly consistent. Besides, the land cover between moments of paired images may already change. Further, more efforts have been paid to learning internal prior

* Corresponding author.

E-mail addresses: xiao_yi@whu.edu.cn (Y. Xiao), yqiang86@gmail.com (Q. Yuan), kuijiang_1994@163.com (K. Jiang), hej96.work@gmail.com (J. He), whuwuy0419@whu.edu.cn (Y. Wang), zlp62@whu.edu.cn (L. Zhang).

<https://doi.org/10.1016/j.inffus.2023.03.021>

Received 7 December 2022; Received in revised form 9 February 2023; Accepted 30 March 2023

Available online 6 April 2023

1566-2535/© 2023 Published by Elsevier B.V.

information before conducting super-resolution. The learned prior is required to guide the downstream SR network to learn how to adapt to various degradation distributions. [31] trained a fully-supervised kernel prediction subnetwork to express blur kernel code from an LR input. However, they explicitly rely on ground-truth blur kernels in the kernel pool, which limits the generalization since we cannot collect all the blur kernels due to complicated variations among different remote sensing scenarios.

To overcome the problems mentioned above, some semi-supervised [32,33] approaches are explored. These methods either learn prior information by self-supervised training in a degradation label-free manner or develop effective domain adaption methods to improve the generalization of SR procedure. Because of the diverse degradation factors, entangled degradations cannot be expressed precisely. Note these approaches mainly incorporated unstable training strategy with limited scopes [33,34] or require a huge memory cost [32] to tame optimization instability of cross-domain learning, which do not properly represent the degradation information in remote sensing imagery. In addition, it is also crucial to establish an effective feature adaption framework to transform LR features to some domain that is suitable for reconstructing HR images. Various studies [35–38] have demonstrated that fusing internal prior information can promote the performance of a visual task. This motivates us to learn the internal degradation prior to guiding the SR process.

Towards this end, we propose a self-supervised degradation-guided adaptive network that incorporates contrastive learning to express robust degradation representations, which synergetically help super-resolution procedures adapt to various degradation distributions. Ideally, we assume that an image contains multiple patches with similar degradations, while degradations vary among different images due to object scale and scene variations. Thus, the degradation representations extracted from the same image can be treated as positive pairs (relevant information), and those from different images constitute negative pairs. Therefore, we can introduce a contrastive learning strategy to generate discriminative representations of different degraded images in the embedding feature space to achieve the adaptive representation of degradations. Specifically, we proposed a multi-view sample augmentation strategy to extract sufficient contrastive samples in a minibatch, which saves us from demanding a memory bank to cache negative samples. In addition to this, we further encode and project these samples into embedding space to encourage the positive embeddings (i.e., similar degradation distributions) closer while pushing the negatives (i.e., dissimilar degradations) away. Finally, a dual-wise feature modulation network is developed to achieve effective feature transformation in channel and feature dimension. To sum up, our contribution is four-fold:

- (1) We proposed a simple and effective self-supervised degradation representation learning strategy which characterizes the robust and discriminative representations in the embedding space with contrastive learning.
- (2) The proposed network can be comfortably applied to arbitrary remote sensing imagery without tedious degradation labels. The learned representations can promote cross-domain super-resolution generalization.
- (3) A dual-wise feature modulation network was developed to adaptively transform low-resolution features into a desired domain which is suitable for reconstruction.
- (4) Extensive experiments on three mainstream remote sensing datasets demonstrate that our method performs favorably against existing classical and real-world super-resolution approaches on simulated and real-world images.

The remaining section of this paper is organized as follows. Section 2 reviews single image super-resolution and its related work on remote sensing imagery. In Section 3, we present the datasets involved in this paper and the key components of the proposed framework. Section 4 contains comprehensive experiments, and in Section 5 we summarize the whole paper.

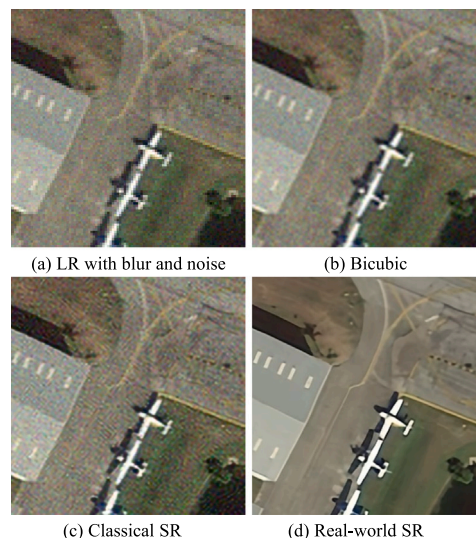


Fig. 1. Visual comparison of (c) classical SR and (d) real-world SR. There is a domain gap between their outputs, which is caused by applying a pre-trained non-blind model (bicubic downsampling) to LR input with multiple degradations (blur & noise).

2. Related work

In this section, we first briefly review the classical single-image super-resolution as it lays the foundation for blind super-resolution. Then we focus on blind super-resolution and its progress on remote sensing imagery.

2.1. Deep learning-based classical single image super-resolution

With the success of convolution neural network (CNN) [39], CNN-based SISR methods are booming. After that, researchers worked on deepening [40] and widening [41] the network design. Subsequently, attention mechanisms [42,43] have been widely deployed in SISR network to further promote their capabilities. In this period, some generative adversarial network (GAN)-based methods [44] and projection-based methods [45] were also proposed. Besides, the basic components of the network are no longer limited to the residual convolution blocks, but also contain dense blocks [46,47]. Recently, transformer-based methods [48,49] have received impressive performance because transformer is better at extracting local and global dependencies. At present, the classical SISR under bicubic-downsampling degradation is mature, and further research for multiple and unknown degradations has become a hotspot.

2.2. Deep learning-based blind image super-resolution

Once we have an external dataset with practical distributions, the complicated non-linear relationship between multi-degraded LR images and HR images can also be parameterized in the CNN to some extent. As mentioned before, it is not an easy task to establish a well-generalized dataset, which makes us focus on learning the internal degradation information in low-resolution images. Depending on whether apparent degradation estimation is performed, these kinds of methods can be subdivided into explicit and implicit degradation modeling.

Explicit degradation modeling. Early work [50] concatenated feature maps derived from explicit blur kernels into LR image features to achieve adaption. These methods allow partial feature transformation but are not sophisticated and thus have limited capacity.

Later, researchers [51,52] start introducing the maximum a posteriori (MAP) algorithm to decompose this inverse problem into two

sub-problems of deblur and super-resolution. For instance, [52] proposed a deep Unfolding Super-resolution Network (USRNet) to jointly optimize the two sub-problems by unfolding the iterative optimization process. Since these methods rely heavily on initial degradation estimates, their performance will drop dramatically due to the mismatch between the input and actual estimation. Therefore, some works try to alleviate this drawback by unifying the process of degradation estimation into a trainable network. [53] established a framework (IKC) to iteratively approach the estimated kernel to the correct kernel. In addition, they further proposed a spatial feature transformation (SFT) layer to better transform the degradation information into features, which is more effective than the direct concatenation used in SRMD [50]. Recently, the deep alternating network (DAN) [54] promoted the IKC by jointly optimizing its corrector and SR network. However, this iterative correction-based approach tends to be unstable with the number of iterations. To produce more accurate and robust degradation estimates, [32] presented a single propagation network without iterative correction, which adopts a sub-network to learn the latent degradation representation adaptively. Recently, some researchers have tried to stabilize the degradation estimation process with mathematical prior. [55] proposed a probabilistic degradation model (PDM) to learn degradation distribution by a random prior. [56] introduced least-squares constraints for degradation modeling.

Moreover, some attempts have been made to model degradation distributions from real-world data. [57] built a generalized kernel pool and then generated LR-HR training pairs through real degradation kernels on this pool. Nevertheless, they still fail to cope with arbitrary inputs that deviate from their kernel pool. Therefore, several works proposed degradation-specific learning at the image level. In 2019, [34] exploited GAN (KernelGAN) to super-resolve an LR input and then adopted a discriminator to judge whether the patch distribution in the super-resolved image is consistent with the original LR image. Another model named ZSSR [58] can realize image-specific super-resolution through zero-shot learning.

Implicit Degradation Modeling. Such methods mainly engaged in grasping domain distribution with the help of external datasets. [59] proposed a cycle-in-cycle GAN (CinCGAN) to transform arbitrary LR images into the classical bicubic-downsampling domain. Thus, they can perform non-blind SR with an optional state-of-the-art network. In addition, [60,61] put forward to realize domain adaptation from HR to LR by learning the degradation process. In degradation GAN [62], the conventional pixel-level loss function was reformed to focus on the sparse high-frequency information to ease the stress of across-domain learning.

In summary, the explicit degradation modeling methods represent mainstream because they are clear and easy to deploy, and the capability to appropriately express complicated degradations is crucial to their performance. As mentioned in previous work [63], the GAN-based implicit degradation modeling approaches are easy to collapse and may cause harmful artifacts in real-world images.

2.3. Remote sensing image super-resolution

On remote sensing imagery, the progress of SISR is roughly in line with computer vision. In particular, we need to make more efforts to explore and analyze the specialty of remote sensing imagery. Early work often simply transferred the models on natural images [64,65] to remote sensing imagery. Later, sophisticated methods [66–68] that integrate remote sensing properties have achieved significant results. Although remarkable progress has been made on bicubic downsampled remote sensing imagery, these non-blind methods still struggle to generalize visual-pleasing results for practical images with blur and noise. Until recently, researchers have paid more attention to blind super-resolution on remote sensing imagery.

Some researchers suggested collecting multi-temporal imagery captured across different sensors (different spatial resolutions) in the same

area to construct realistic LR-HR pairs. For instance, [30] used Sentinel-2 and Planet images of the same region for supervised training. However, this approach requires strict spatial-temporal consistency, which is challenging for alignment. Hence, an unsupervised learning strategy is introduced to ease the reliance on real-world datasets. [69] designed a recurrent CNN network to learn the downsampling and upsampling process cyclically. [70] also proposed to learn two subnetworks named degenerator and generator. With the help of these components, they can implement supervised learning by converting the optimization problem between LR and latent HR to minimize the loss function between degraded LR and the original LR. This idea has been similarly applied to hyperspectral images [71]. Although these methods partly save us from building real-world datasets through unsupervised learning, they still have two drawbacks. First, downsampling the LR images may cause further loss of structural details that are already lacking in remote sensing imagery. Second, degradations in LR images can be amplified by the degenerator. Recently, [31] developed a network to learn the implicit kernel code from original LR images and transform the code into LR features by kernel-aware layer. Unfortunately, they still cannot escape from the monotonous datasets because they need real blur kernels to enrich their degradation kernel pool. Moreover, the kernel code prediction is fully supervised. To alleviate the reliance on labeled remote sensing data, [72] designed an unsupervised multi-layer model for multi-degradation learning. [73] proposed a novel self-fusion strategy to map low-frequency remote sensing images to the high-frequency domain. However, unsupervised training in large-scale remote sensing images is easy to collapse.

To mitigate the aforementioned problems, we developed a degradation label-free network with self-supervised training. Furthermore, we can achieve effective feature adaption through the proposed dual-wise feature modulation network.

3. Methodology

3.1. Degradation formulation

Mathematically, classical super-resolution assumes that a high-resolution (HR) image Y is degraded by a single degradation factor by the following:

$$X = D_s(Y), \quad (1)$$

where D_s is a downsampler with a scale factor s , and X is the low-resolution (LR) image degraded from Y . By setting D_s to a specific operator, e.g., bicubic downsampling, an external dataset with paired LR-HR images can be established. In blind SR, the degradation process can be expanded to:

$$X = D_s(Y \otimes B) + N, \quad (2)$$

where $B \in \mathbb{R}^2$ represents a 2D blur kernel, \otimes denotes convolution operation, and N is a noise term. In this paper, we follow the setting in previous work [53]. That is, D_s is bicubic downsampling, B is set to Gaussian blur kernels, and N represents Gaussian noise.

3.2. The proposed approach

3.2.1. Overall

As illustrated in Fig. 3, our method consists of three major components. An encoder $E(\cdot)$ for degradation representation generation, a Dual-wise feature Modulation Network $DMN(\cdot)$ for representation-aware feature adaption, and an upsampler for reconstruction.

$E(\cdot)$ is trained to express degradation representations in a self-supervised manner. For an LR input $X \in \mathbb{R}^{h \times w \times c}$, the robust representation R can be denoted as:

$$R = E(X). \quad (3)$$

After feature extraction, we get an LR feature F of size $h \times w \times 64$ from X . Then, F and R are both sent into $DMN(\cdot)$. Influenced by R , F is further modulated to the target HR feature domain. The final modulated feature \tilde{F} will be established by global residual connection, that is:

$$\tilde{F} = DMN(F, R) + F. \quad (4)$$

Eventually, we adopt the pixelshuffle operation [74] to construct the final super-resolved image $X^{SR} \in \mathbb{R}^{hr \times ur \times c}$, which can be formulated as:

$$X^{SR} = Up(\tilde{F}), \quad (5)$$

where r is the upscale factor, $c = 3$ represents RGB channels, and $Up(\cdot)$ means the upsampler. Here X^{SR} should be as close as possible to the ground-truth image Y .

In the following, we will describe how to predict degradation representations with self-supervised training and the details of dual-wise feature modulation.

Algorithm 1: Self-supervised Degradation Representation Learning Algorithm.

Input: N LR images $\{X_n\}_{n=1}^N$ in a mini-batch.
1 Initialization: $N = 8$, $E(\cdot)$ is Encoder, $l^{(1)}$ and $l^{(2)}$ are two fully connected layers, $\sigma(\cdot)$ is ReLU activation.
2 foreach $\{x_n^1, x_n^2\}$ sampled from X_n **do**
3 $\tilde{x}_n^1 = f(x_n^1), \tilde{x}_n^2 = f(x_n^2);$ // Augmentation
4 $h_n^1 = E(x_n^1), h_n^2 = E(x_n^2);$ // Representations
5 $z_n^1 = P(h_n^1), z_n^2 = P(h_n^2);$ // Embeddings
6 end
7 Define $\{z_k\}_{k=1}^{2N} = \{z_1^1, z_1^2, \dots, z_N^1, z_N^2\}_{n=1}^N$;
8 forall the $i \in \{1, \dots, 2N\}$ **and** $j \in \{1, \dots, 2N\}$ **do**
9 $penalty_{i,j} = z_i^T z_j / (\|z_i\| \cdot \|z_j\|);$ // Similarity
10 end
11 Define $\ell_{i,j} = -\log \frac{\exp(penalty_{i,j}/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(penalty_{i,k}/\tau)}$;
12 $\mathcal{L}_{encoder} = \frac{1}{2N} \sum_{k=1}^{2N} [\ell_{2k-1,2k} + \ell_{2k,2k-1}];$ // Encoder Loss
13 Update $E(\cdot)$ and $P(\cdot)$ to minimize $\mathcal{L}_{encoder}$.
14 Return $E(\cdot)$ and throw away $P(\cdot)$.

3.2.2. Degradation representation

Multi-view Sample Augmentation. As shown in Fig. 2, given an LR input X_n from N images $\{X_n\}_{n=1}^N$ in a minibatch, we firstly extract two random samples $\{x_n^1, x_n^2\}$ from it. Under the assumption that any patch in an image suffers the same degradation, x_n^1 and x_n^2 should carry similar representations. Inspired by SimCLR [75], sample augmentation $f(\cdot)$ is performed in each sample to generate multi-views $\{\tilde{x}_n^1, \tilde{x}_n^2\}$. Because multiple data augmentation is crucial in yielding effective representations. Unlike SimCLR, which aims at image classification, our goal is to generate a discriminative representation of degradations in X . Therefore, $f(\cdot)$ is not allowed to destroy the spatial integrity of $\{x_n^1, x_n^2\}$. Here we adopt a simple but effective augmentation by random rotating (90°, 180°, and 270°) and flipping (horizontally or vertically) each sample. Taking x_n^1 as an example, we will get an augmentation \tilde{x}_n^1 by the following:

$$\tilde{x}_n^1 = f(x_n^1) \quad (6)$$

Encoder. Two correlated views $\{\tilde{x}_n^1, \tilde{x}_n^2\}$ are fed into an encoder to obtain two similar representations $\{h_n^1, h_n^2\}$. Note that the two encoders in Fig. 2 are parameter sharing. The encoder $E(\cdot)$ is a flexible component for generating degradation representations, which can be easily replaced by any state-of-the-art blocks. In this paper, we employed a lightweight design because we can comfortably express discriminative representation benefits from our self-supervised learning strategy.

Specifically, $E(\cdot)$ is designed to a shallow ResNet with one average pooling tail and six convolution layers followed by batch normalization (BN) and ReLU activation. The input channels of these convolution layers are $c = 3, 64, 64, 128, 128, \text{ and } 256$. The degradation representation $h_n^1 \in \mathbb{R}^{1 \times 1 \times 256}$ is obtained by:

$$h_n^1 = E(x_n^1) \quad (7)$$

Projection Head. After that, we need to further project these representations into embedding space to compute the contrastive loss. Following SimCLR, a lightweight MLP layer is employed for projection $p(\cdot)$. That is:

$$z_n^1 = p(h_n^1) = l^{(1)} \sigma(l^{(2)} h_n^1), \quad (8)$$

where $l^{(1)}$ and $l^{(2)}$ are two fully connected layers, $\sigma(\cdot)$ is a ReLU activation used to introduce nonlinearity.

So far, we have gained two positive embeddings $\{z_n^1, z_n^2\}$ which incorporate the same degradations. Different from previous work [32], which constructs a huge memory bank to save negative embeddings, we simply treat the remaining $N - 1$ images in a mini-batch that suffer different degradation from X as negative images. After the same degradation representation procedure mentioned above, we can get $2(N - 1)$ negative embeddings z_k ($k = 1, \dots, 2(N - 1)$). Next, the encoder will be trained to nicely distinguish a degradation from others with the help of contrastive learning.

Contrastive Learning. Self-supervision is carried by a contrastive loss function. It works by maximizing the similarity of positive pairs and distinguishing the negative embeddings away from positives in the embedding space. Before that, we need a penalty to measure the similarity between embeddings. For instance, the penalty between z_i and z_j , namely $penalty_{i,j}$, is determined by cosine similarity:

$$penalty_{i,j} = z_i^T z_j / (\|z_i\| \cdot \|z_j\|). \quad (9)$$

As shown in Fig. 2, if we set z_i as a positive anchor, the contrastive loss $\ell_{i,j}$ encourages pulling positive embedding z_j to z_i and pushing negative z_k away from z_i in the embedding space. Following SimCLR, the normalized temperature-scaled cross entropy (NT-Xent) loss is used to identify z_j from $\{z_k\}_{k \neq i}$ for a given z_i , which can be formulated by:

$$\ell_{i,j} = -\log \frac{\exp(penalty_{i,j}/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(penalty_{i,k}/\tau)}, \quad (10)$$

where $1_{[k \neq i]}$ means a flag evaluating to 1 if $k \neq i$ and 0 if $k = i$. Here τ is a temperature hyper-parameter. For any positive pairs, we treat the remaining $2(N - 1)$ embeddings in the same minibatch as negatives. In the training process, we define all the $2N$ embedding as $\{z_k\}_{k=1}^{2N} = \{z_1^1, z_1^2, \dots, z_N^1, z_N^2\}_{n=1}^N$. For image X_k , we calculate $\ell_{2k-1,2k}$ and the reverse one $\ell_{2k,2k-1}$. Finally, the total loss $\mathcal{L}_{encoder}$ for training encoder is written as:

$$\mathcal{L}_{encoder} = \frac{1}{2N} \sum_{k=1}^{2N} [\ell_{2k-1,2k} + \ell_{2k,2k-1}]. \quad (11)$$

In the training process, we update the parameters of the encoder and projection head and throw away the projection head in the test phase. The self-supervised degradation representation learning algorithm is summarized in Algorithm. 1.

3.2.3. Dual-wise feature modulation network

In this part, we need to perform adaptive super-resolution reconstruction with the guide of multiple degradation representations. As presented in Fig. 3, the architecture of a dual-wise modulation network (DMN) is straightforward. The basic component of DMN is the dual-wise modulation block (DMB). See Fig. 4 for more details. DMB can transform the robust degradation representations into the network from both feature and channel dimensions to achieve representation-aware feature adaption.

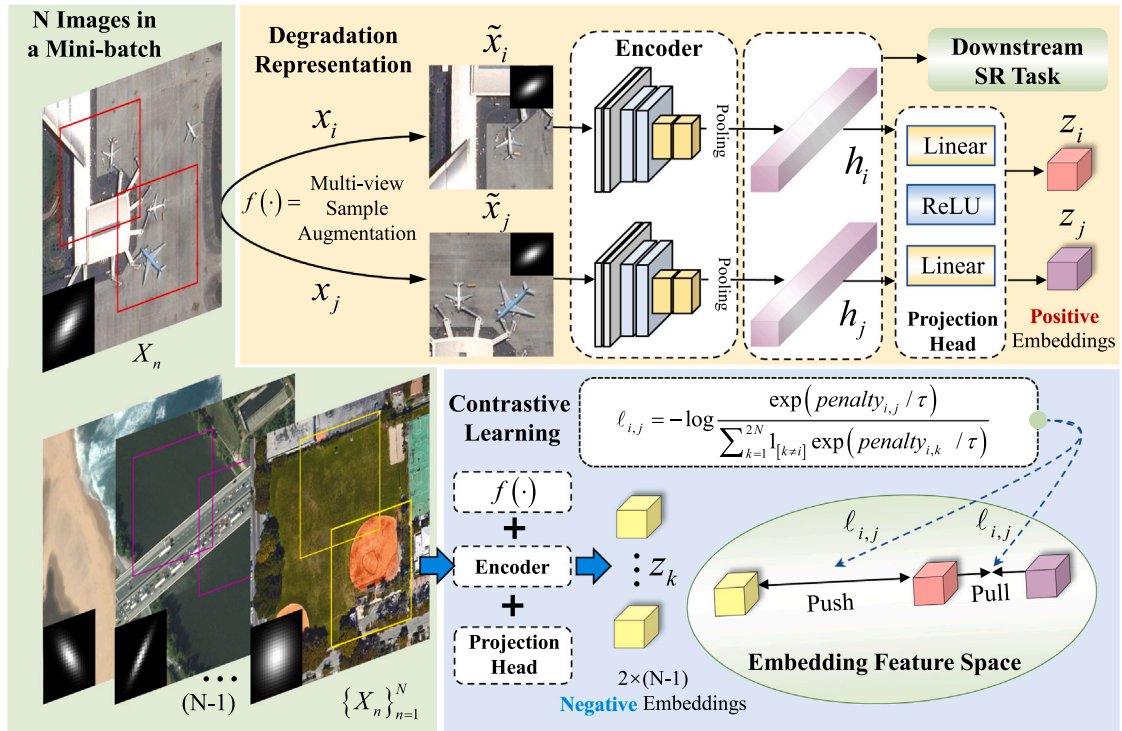


Fig. 2. The flowchart of our proposed self-supervised degradation representation approach. Given N images in a minibatch, the embeddings extracted from the same image are defined as positive pairs but the negatives from the different images. We introduce contrast learning to pull the positive embeddings closer in the feature space and push the negative embeddings away. Thus, the intrinsic degradation information is adaptively mapped for accurate blind SR reconstruction.

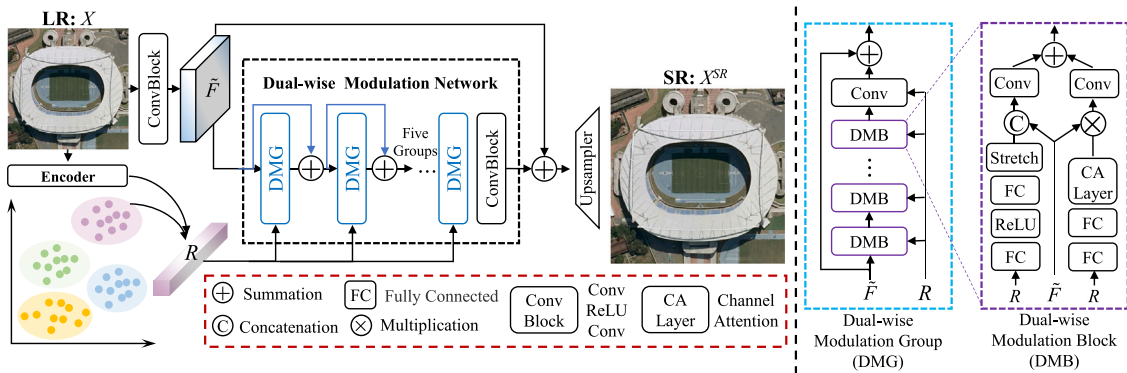


Fig. 3. The overall structure of our network. The encoder is trained by the self-supervised strategy in Fig. 2 to encode degradation representations R . Then R and low-resolution features F are sent to a dual-wise modulation network for feature adaption. The final upsampler is performed by pixel shuffle operation.

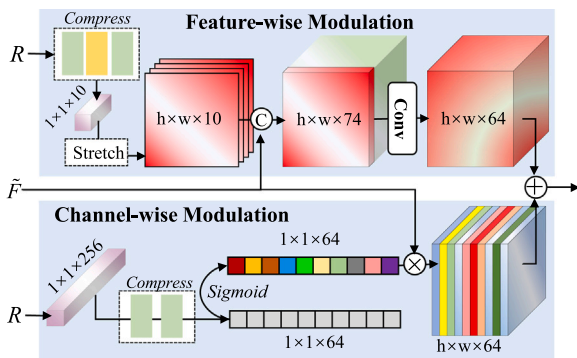


Fig. 4. The details of our dual-wise modulation block.

Feature-wise Modulation. First, the representation $h_i \in \mathbb{R}^{1 \times 1 \times 256}$ should be reshaped into the same height and width with LR feature $F \in \mathbb{R}^{h \times w \times 64}$ to realize adaption in the feature dimension. Essentially, the degradation representation R is devoted in expressing degradations, and it may not be spatially informative. Thus, we must compress the representation along the channel dimension to avoid over-influence on the LR feature. In detail, we first reduce the channel number of R to 10 by passing R through a fully connected layer, followed by ReLU activation and another fully connected layer. Subsequently, we adopt the dimensionality stretching [53] termed as *Stretching*(\cdot) to reshape R to the size of $h \times w \times 10$. Through the concatenation, the degradation representations can be integrated shallowly into F , then we set a convolution block to deeply propagate R into F . This shallow to deep modulation allows comprehensive adaption at the feature level. In short, our feature-wise modulation can be formulated as follows:

$$R^f = \text{Compress}(R), \quad (12)$$

$$F^f = \text{ConvBlock}([\text{Stretching}(R^f), F]), \quad (13)$$

where $Compress(\cdot)$ aims at reducing the channel numbers and $[-]$ means concatenation operation. R^f is the compressed representation and F^f represents the modulated features in the feature dimension.

Channel-wise Modulation. To make degradation representations transformation adaptively in channel-wise, we also need to unify the channel dimensionality of LR feature F and R . At first, we compress R to 64 channels by two fully connected (FC) layers. The internal channel number of these two FC layers are set to 256 and 64. Next, we employ the widely used channel attention mechanism for feature adaptation. In particular, the compressed representation R^c is activated by *sigmoid* function. After that, modulation in the feature dimension can be achieved by multiplying the attention map into LR features F .

$$R^c = Compress(R), \quad (14)$$

$$F^c = Sigmoid(R^c) \otimes F, \quad (15)$$

where \otimes denotes the channel-wise multiplication, and F^c represents the modulated features in the channel dimension.

Note that the output of each DMG will be sent as input for the next DMG with a local residual connection. After the dual-wise modulation network, we set a global residual connection to stabilize the gradient optimization. The LR feature F is adaptively modulated to the target domain \tilde{F} , which is ready for the reconstruction of the desired HR image. In the final upsampler layer, we conduct two $\times 2$ pixel shuffle upscale operations to reconstruct the super-resolved image X^{SR} , that is:

$$X^{SR} = Up(\tilde{F}). \quad (16)$$

4. Experiment

4.1. Dataset setting

Firstly, three remote sensing datasets involved in this paper are introduced in detail. Next, we present the degradation setting used for synthesizing LR images.

(1) AID Dataset. AID (Aerial Image Dataset) [76] is a widely used benchmark dataset for aerial scene classification, which can provide over 10,000 images with fixed 600×600 pixels of 30 scene classifications. Here, we utilize AID for training and testing. Specifically, 5000 images were randomly picked from AID to build the training set. In the remaining part, we further selected ten random images in each scene classification for testing. Finally, the test set contains a total of 300 images, and we denoted it as AID-T.

(2) DOTA Dataset. DOTA (Dataset of Object deTecton in Aerial images) [77] is a large-scale benchmark for object detection in aerial images. In DOTA, 11268 aerial images of 18 object categories are available, and the image size varies from 800×800 to 20000×20000 . The authors collected these images from various sensors, including Google Earth, the Gaofen-2 Satellite, Jilin-1 Satellite, and airborne platforms. We only use DOTA for testing and do not involve it in the training process to validate the generalization of our model on diverse datasets. To save memory cost, we simply choose images below 2000×2000 pixels to build the test set (DOTA-T). After a random selection, we ended up with 250 images for testing.

(3) Video Satellite Imagery Dataset. The third dataset was derived from Jilin-1 satellite videos. Different from AID-T and DOTA-T, which were collected from static images, satellite video has a higher temporal resolution for dynamic observations. Also, Jilin-T was only established for testing. In detail, ten source videos with a frame size of 4096×2160 were cropped to 200 video clips with a size of 640×640 . In each clip, we randomly sampled one frame from 100 consecutive frames for testing. Hence, we can build a test set (Jilin-T) that covers 200 video satellite images. For more details about the satellite video data source, please refer to our previous work [67].

Some typical samples in the aforementioned three datasets can be found in Fig. 6. Thanks to the extensive sensor types and remote sensing scenes in these datasets, we can comprehensively validate our model.

Degradation Setting. Following previous work [32], we degrade the original images to construct LR-HR training pairs. In Eq. (2), the size of the Gaussian blur kernel is fixed to 21×21 . We first build LR images with noise-free and isotropic Gaussian blur, and the kernel width σ is randomly chosen from $[0.2, 4]$. Then we develop a more general case by degrading images with noise injection and anisotropic Gaussian blur kernel. Here the width of the anisotropic Gaussian blur kernel is determined by a Gaussian probability density function, which follows a normal distribution $N(0, \Sigma)$. Σ is covariance and it is determined by two random eigenvalues $\lambda_1, \lambda_2 \sim U(0.2, 4)$ and a random rotation angle $\theta \sim U(0, \pi)$. The noise level ranges from $[0, 25]$.

4.2. Implementation details

In this paper, we only implement 4 \times super-resolution, i.e., $r = 4$. The number of dual-wise modulation groups (DMG) and dual-wise modulation blocks (DMB) is set to 5 in the final model. The temperature parameter τ in Eq. (10) is 0.1. More details about the network design are discussed in the ablation study section. During model training, the minibatch is set to 8. Benefiting from our multi-view sample augmentation strategy, our model does not require a large batchsize to cover lavish negative images. In each batch, we randomly crop and rotate 8 HR images and then degrade them by eight random Gaussian blurs and noises. To stable the training process, we adopt a two-stage training strategy. First, we only update the parameters of the encoder and projection head, where the learning rate is 1×10^{-3} and decays by a factor of 10 at half of the total 100 epochs. In the next stage, we train the entire network for 600 epochs with an initial learning rate of 1×10^{-4} , and it decays to 1/2 of the previous one at every 100 epochs. Adam Optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is used for optimization. The \mathcal{L}_1 function is used to measure the pixel-level difference between the super-resolved image X^{SR} and the ground-truth Y by the following:

$$\mathcal{L}_{SR} = \|Y - X^{SR}\|_1, \quad (17)$$

Finally, the overall loss \mathcal{L} can be written as:

$$\mathcal{L} = \mathcal{L}_{encoder} + \mathcal{L}_{SR}. \quad (18)$$

We conducted all the experiments on a single NVIDIA RTX 2080Ti GPU. It takes nearly 48 h to train our model.

4.3. Experiments on noise-free and isotropic blur

First, let us consider a simple case where an image is only blurred by an isotropic kernel. We carefully evaluate our method on three test sets and compare it with various models, including Bicubic, HAN [43], ACT [78], ZSSR [58], IKC [53], AdaTarget [79], and CMDSR [33]. HAN is a state-of-the-art non-blind super-resolution method designed for bicubic degradation. Here we retrained HAN on our synthesized LR-HR training pairs with multiple degradations for a fair comparison. ACT is a powerful transformer-based image restoration model. We compare our CNN-based model with ACT to demonstrate our superiority. ZSSR is an unsupervised method based on zero-shot learning, which can tackle isotropic and anisotropic blurs. IKC requires explicit blur kernel estimation, but it is only designed for isotropic blur degradation. AdaTarget is an adaptive method without explicit blur kernel estimation, and it can handle both isotropic and anisotropic blurs. CMDSR also employs a modulation idea. By comparing with CMDSR, we can prove the effectiveness of our dual-wise modulation. All these methods are retrained in our training set following the official settings. In quantitative results, we use peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) for evaluation. For qualitative comparison,

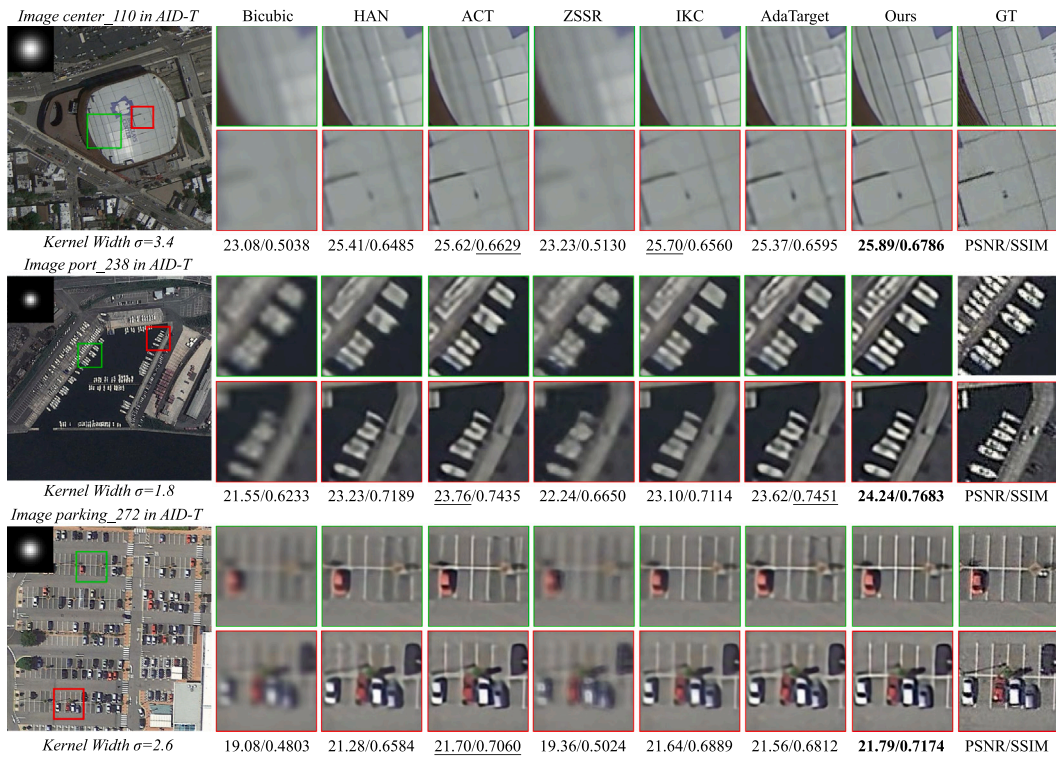


Fig. 5. Visual comparisons of experiments on noise-free and isotropic Gaussian blur. All these restored images are selected from AID-T. The best PSNR/SSIM is shown in **bold** below the image, and the second place is marked by underline.

we zoom and display the visual details to validate the fidelity of the reconstruction images.

Quantitative results. Table 1 summarizes the PSNR/SSIM performance of each method on AID-T, and we further categorize the results by 30 scene classifications. Our method achieves the best metrics in all scenes, which demonstrates our considerable performance can be well-generalized to various remote sensing scenes. In the “playground” category, we can surpass the second place IKC by a significant improvement of 0.45 dB. Furthermore, we find that the non-blind method HAN can lead to ZSSR and CMDSR. On the one hand, this illustrates that deblur can be parameterized in CNN to some extent. On the other hand, it indicates the purely unsupervised method ZSSR, which relies on internal statistics, is struggling to handle complex degradations. As the authors stated in their paper [58], ZSSR can only maintain its stability in limited cases. CMDSR put forward a multi-task meta-learning strategy to predict conditional features. However, this complicated multi-task learning may not precisely express the degradation. Thanks to our straightforward contrastive learning, we can comfortably generate discriminative representations for feature adaption. The blind methods ACT and AdaTarget produce closing performance, which suggests the favorable deblur potential of the transformer-based method. Besides, it reflects the performance bottleneck of AdaTarget caused by missing degradation estimation. With the help of kernel estimation, IKC ranks second. Unlike IKC, our method throws away the explicit blur kernel estimation and only predicts an abstract representation, which alleviates the performance drop caused by kernel mismatch. In conclusion, our method can achieve the best performance benefit from robust and accurate degradation representation.

Table 2 gives the average PSNR/SSIM results of each model on DOTA-T and Jilin-T. In Fig. 8, we display the reconstruction performance of the model with different blur kernel widths (0.2, 1.0, 1.8, 1.6, 3.4). It can be seen that our method exceeds all the comparison methods in all widths of blur kernels. It is worth noting that we only train each model on AID and directly test them on DOTA-T and Jilin-T, which reveals the good generalizability of our method across different datasets.



Fig. 6. Some typical samples of (a) AID, (b) DOTA-T, and (c) Jilin-T. AID has 30 scene classifications, DOTA-T is collected from multiple sensors, and Jilin-T is derived from satellite video.

Qualitative results. In Fig. 5, we exhibit the reconstruction results of three scenes in AID-T. For better comparison, the PSNR/SSIM metrics are also listed under each image. In image “center_110”, Bicubic and ZSSR can barely recover fine-grained details, and the non-blind method HAN produces distinct blurring. In the green box, unfortunately, none of the methods can restore the complete lines. Besides, AdaTarget creates unpleasing distortions. As mentioned before, methods without degenerate estimation are not yet robust enough. IKC and ACT are slightly satisfactory compared to the other methods. However, the texture is still not very clear and blurred. It is straightforward to find

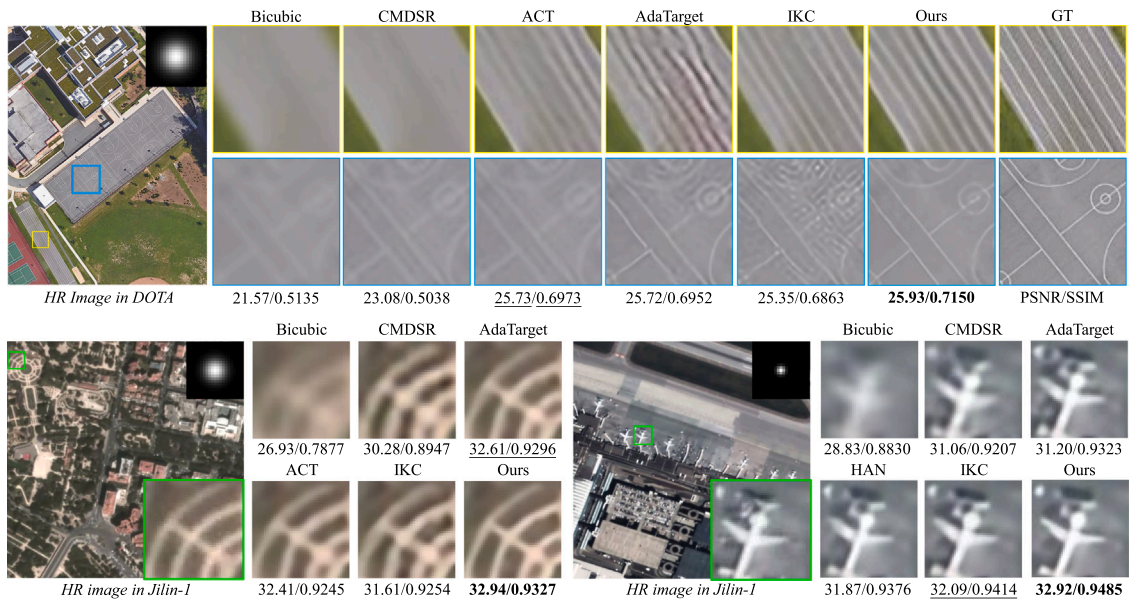


Fig. 7. Visual comparisons of experiments on noise-free and isotropic Gaussian blur. These restored images are selected from DOTA-T and Jilin-T. The best PSNR/SSIM is shown in bold below the image, and the second place is marked by underline.

Table 1

The PSNR/SSIM results of experiments on noise-free and isotropic Gaussian blur. We test these models on AID-T with five kernel widths $\sigma = 0.2, 1.0, 1.8, 2.6,$ and 3.4 . The results are then divided into 30 classes according to scene categories. The best result is shown in bold, and the second place is marked by underline.

Land cover	Bicubic	HAN	ACT	ZSSR	IKC	AdaTarget	CMDSR	Ours
Airport	27.16/0.7173	29.82/0.8109	29.90/0.8162	27.79/0.7415	<u>30.20/0.8223</u>	29.88/0.8167	28.76/0.7876	30.30/0.8249
Bare land	29.97/0.6684	31.72/0.7296	31.39/0.7320	30.47/0.6917	<u>31.87/0.7380</u>	31.79/0.7380	30.77/0.7202	32.23/0.7452
Baseball field	27.88/0.7286	30.56/0.8167	30.73/0.8247	28.48/0.7508	<u>30.95/0.8268</u>	30.75/0.8250	29.47/0.7934	31.09/0.8308
Beach	32.36/0.8390	<u>34.69/0.8778</u>	34.29/0.8807	32.88/0.8501	<u>34.47/0.8810</u>	34.49/0.8804	31.97/0.8663	35.14/0.8859
Bridge	28.44/0.7836	31.31/0.8523	<u>31.58/0.8575</u>	29.02/0.7998	<u>31.43/0.8608</u>	31.25/0.8559	29.75/0.8322	32.01/0.8640
Center	25.53/0.6382	28.21/0.7635	<u>28.57/0.7804</u>	26.18/0.6718	28.08/0.7737	28.26/0.7777	26.68/0.7207	29.27/0.7929
Church	22.69/0.5478	25.08/0.6821	<u>25.42/0.7045</u>	23.32/0.5840	<u>25.50/0.7012</u>	25.32/0.7012	24.04/0.6310	25.65/0.7146
Commercial	24.96/0.6561	27.79/0.7901	27.96/0.7966	25.64/0.6903	<u>28.13/0.8000</u>	27.80/0.7955	26.69/0.7592	28.22/0.8039
D-Residential	21.90/0.4761	23.95/0.6370	24.14/0.6493	22.36/0.5135	<u>24.33/0.6546</u>	24.06/0.6508	23.23/0.5905	24.35/0.6609
Desert	38.09/0.9075	39.09/0.9209	36.96/0.9223	37.97/0.9095	39.36/0.9225	<u>39.36/0.9226</u>	38.76/0.9176	39.81/0.9242
Farmland	31.22/0.7898	33.99/0.8552	34.24/0.8621	31.78/0.8062	<u>34.67/0.8674</u>	34.08/0.8603	32.85/0.8403	34.68/0.8687
Forest	26.71/0.5252	28.27/0.6380	28.40/0.6475	27.17/0.5650	<u>28.47/0.6495</u>	28.39/0.6548	27.86/0.6203	28.58/0.6646
Industrial	24.92/0.6045	27.50/0.7334	27.67/0.7429	25.54/0.6369	<u>27.90/0.7527</u>	27.58/0.7447	26.44/0.7051	28.03/0.7561
Meadow	34.94/0.8035	<u>36.40/0.8306</u>	35.96/0.8333	35.21/0.8135	35.97/0.8342	<u>36.30/0.8342</u>	35.36/0.8249	36.82/0.8385
M-Residential	26.13/0.5978	28.30/0.7096	28.58/0.7246	26.61/0.6259	28.22/0.7213	<u>28.26/0.7215</u>	26.90/0.6693	29.15/0.7348
Mountain	29.05/0.6821	31.12/0.7708	31.06/0.7731	29.60/0.7103	<u>31.25/0.7756</u>	31.07/0.7748	30.23/0.7563	31.36/0.7808
Park	26.40/0.6179	28.63/0.7297	28.69/0.7341	26.95/0.6497	<u>28.89/0.7408</u>	28.66/0.7368	27.83/0.7099	28.93/0.7433
Parking	22.33/0.6421	26.26/0.8005	26.84/0.8187	23.07/0.6750	<u>27.01/0.8188</u>	26.56/0.8132	24.63/0.7461	27.31/0.8306
Playground	28.42/0.7929	32.37/0.8860	32.47/0.8909	29.35/0.8166	<u>32.71/0.8958</u>	32.45/0.8909	30.36/0.8576	33.26/0.9005
Pond	27.82/0.7279	29.55/0.7883	29.64/0.7936	28.19/0.7433	<u>29.87/0.7996</u>	29.64/0.7963	28.82/0.7742	29.95/0.8023
Port	27.01/0.7904	29.93/0.8675	30.20/0.8736	27.65/0.8091	<u>30.21/0.8762</u>	29.98/0.8737	28.46/0.8467	30.51/0.8810
Railway station	25.87/0.6242	28.74/0.7619	28.86/0.7671	26.66/0.6645	<u>29.04/0.7741</u>	28.79/0.7699	27.48/0.7276	29.20/0.7828
Resort	25.78/0.6697	28.62/0.7752	28.82/0.7828	26.41/0.6953	<u>28.87/0.7842</u>	28.59/0.7810	27.00/0.7426	29.18/0.7918
River	29.12/0.6780	30.86/0.7570	30.87/0.7625	29.53/0.7014	<u>31.07/0.7664</u>	30.86/0.7651	30.23/0.7447	31.16/0.7703
School	25.77/0.6526	28.67/0.7835	<u>28.96/0.7913</u>	26.40/0.6827	28.79/0.7963	28.62/0.7888	27.29/0.7487	29.26/0.8002
S-Residential	25.39/0.4721	26.63/0.5743	26.76/0.5873	25.73/0.5064	<u>26.83/0.5909</u>	26.74/0.5912	26.24/0.5494	26.85/0.5977
Square	24.86/0.6481	27.81/0.7723	<u>28.23/0.7857</u>	25.53/0.6781	27.99/0.7858	27.94/0.7833	26.29/0.7361	28.65/0.7965
Stadium	24.89/0.6703	27.65/0.7826	27.84/0.7918	25.48/0.6947	<u>28.06/0.7988</u>	27.71/0.7908	26.52/0.7505	28.15/0.8012
Storage tanks	24.62/0.6262	26.76/0.7305	26.92/0.7391	25.13/0.6519	<u>27.07/0.7415</u>	26.89/0.7416	25.95/0.7029	27.14/0.7492
Viaduct	26.06/0.6286	28.73/0.7582	28.83/0.7636	26.60/0.6599	<u>29.17/0.7752</u>	28.77/0.7656	27.63/0.7250	29.24/0.7792
Overall	27.21/0.6736	29.63/0.7729	29.69/0.7810	27.76/0.6996	<u>29.88/0.7842</u>	29.70/0.7814	28.48/0.7466	30.18/0.7906

that our method not only recovers the maximum number of lines but also successfully removes the blur and yields sharper edges. In the red box, only our method can restore complete details closest to the ground truth. Image “port_238” contains a dense distribution of multi-scale ships. Our method can better distinguish the ships, while the other models produce blending drawbacks between multi-scale objects. In image “parking_272”, paying attention to the distribution of markings on the ground and the boundary of the cars, our method yields the sharpest edge distribution with minor distortion. This proves

that our representations can characterize precise degradations even in multi-scale objects.

Fig. 7 exhibits the visual comparison of the DOTA-T and Jilin-T. In the yellow box of an image in DOTA-T, CMDSR almost lost all the markings on the ground. ACT and IKC give some pleasing results, but still suffer from severe artifacts and cannot restore all the details. In the basketball court below, IKC has a dramatic performance drop. Because when real distribution deviates from the estimated kernel, kernel misestimation may provide interference, which badly damages












Table 2

The average PSNR/SSIM and model efficiency results of experiments on noise-free and isotropic Gaussian blur. We test these models on AID-T, DOTA-T, and Jilin-T with five kernel widths $\sigma = 0.2, 1.0, 1.8, 2.6,$ and 3.4 . The PSNR/SSIM results under each kernel width are then averaged. The best result is shown in **bold**, and the second place is marked by underline. FLOPs of different models are calculated on an LR image of size 150×150 . The test time is the total time to complete the test on three test sets.

Method	Parameter	FLOPs	Test time (s)	AID-T	DOTA-T	JiLin-T
Bicubic	–	–	–	27.21/0.6736	24.80/0.6432	29.61/0.8468
HAN	64.19 M	1475.06 G	242.46	29.50/0.7720	28.65/0.7803	32.23/0.9215
ACT	46.18 M	230.84 G	51.53	29.69/0.7810	29.06/0.7942	33.26/0.9328
IKC	9.45 M	130.58 G	85.30	<u>29.88/0.7842</u>	<u>29.17/0.7970</u>	33.11/ <u>0.9334</u>
AdaTarget	16.70 M	403.38 G	196.03	29.70/0.7814	28.91/0.7914	31.80/0.9155
CMDSR	1.50 M	2.86 G	31.47	28.48/0.7466	27.52/0.7426	31.20/0.9047
Ours	5.74 M	115.23 G	48.43	30.18/0.7906	29.36/0.8033	33.34/0.9350

Table 3

The PSNR/SSIM results of experiments on noises and anisotropic Gaussian blur. We test these models on AID-T with 11 typical kernel widths and noise levels from 0 to 10. The best result is shown in **bold**, and the second place is marked by underline.

Model	Noise	Anisotropic gaussian blur kernels										Average	
													
Bicubic	0	27.67	27.56	26.93	26.27	26.09	26.81	27.02	26.20	26.62	26.08	25.83	26.64
ACT		<u>29.54</u>	29.46	29.17	28.35	28.53	29.12	29.28	28.74	28.78	28.55	28.49	28.91
IKC		29.33	<u>29.85</u>	29.09	27.43	26.90	27.18	29.05	27.73	27.62	27.98	29.18	28.30
AdaTarget		29.32	29.46	29.10	28.59	28.54	28.69	29.02	28.55	28.67	28.40	28.12	28.77
DASR		29.50	29.52	<u>29.31</u>	<u>28.89</u>	<u>28.71</u>	<u>29.22</u>	<u>29.34</u>	<u>28.79</u>	<u>28.97</u>	<u>28.66</u>	28.68	<u>29.05</u>
CMDSR		28.76	28.90	28.60	27.55	27.78	28.19	28.55	27.72	27.65	27.91	28.29	28.17
Ours		30.01	30.01	29.86	29.39	29.23	29.76	29.87	29.41	29.56	29.22	29.14	29.59
Bicubic	5	27.26	27.16	26.59	25.98	25.80	26.48	26.66	25.91	26.30	25.80	25.56	26.32
ACT		28.59	28.49	28.00	27.31	<u>27.23</u>	<u>27.93</u>	<u>28.12</u>	<u>27.40</u>	<u>27.73</u>	<u>27.27</u>	<u>27.08</u>	<u>27.74</u>
IKC		27.85	27.76	26.87	26.03	25.77	26.67	26.96	25.92	26.45	25.76	25.45	26.50
AdaTarget		28.27	28.26	27.59	26.87	26.69	27.41	27.66	26.86	27.33	26.65	26.28	27.26
DASR		<u>28.60</u>	<u>28.53</u>	<u>28.05</u>	<u>27.46</u>	<u>27.23</u>	27.90	28.07	27.24	27.67	27.14	26.99	27.72
CMDSR		28.14	28.15	27.74	27.01	26.97	27.50	27.75	27.05	27.20	27.01	26.94	27.41
Ours		28.90	28.83	28.35	27.73	27.51	28.19	28.40	27.65	28.08	27.50	27.29	28.04
Bicubic	10	26.43	26.34	25.86	25.33	25.18	25.76	25.92	25.27	25.61	25.17	24.97	25.62
ACT		27.80	27.70	27.21	26.64	26.55	27.16	<u>27.32</u>	<u>26.67</u>	<u>26.98</u>	<u>26.57</u>	<u>26.40</u>	<u>27.00</u>
IKC		26.20	26.11	25.51	24.92	24.73	25.38	25.58	24.83	25.24	24.72	24.49	25.25
AdaTarget		27.36	27.28	26.61	26.00	25.81	26.54	26.74	25.98	26.46	25.78	25.52	26.37
DASR		27.79	<u>27.73</u>	<u>27.25</u>	<u>26.72</u>	26.53	27.12	27.28	26.54	26.90	26.46	26.31	26.96
CMDSR		27.45	27.42	27.05	26.49	26.40	26.89	27.07	26.47	26.69	26.42	26.30	26.79
Ours		28.02	27.94	27.46	26.92	26.74	27.34	27.51	26.85	27.23	26.74	26.55	27.21

the fidelity. We can observe a similar issue in the Jilin-T images. The reconstruction is easily distorted and blurred at the high-frequency part, such as edges and boundaries. In this challenging case, our method is visually satisfactory with sharper information and less blur (see Table 5).

4.4. Experiments on noises and anisotropic blur

In this experiment, the degradation setting is more general. That is, an image contains noise and anisotropic blur with Gaussian probability distribution. Now, it is challenging for the isotropic-oriented method IKC to tackle this issue. Therefore, we additionally add a model named DASR [32] for a fair comparison. DASR is designed for isotropic/anisotropic blur and noise. Also, ACT, IKC, AdaTarget, DASR, and CMDSR were retrained with the general degradation setting.

Qualitative Results. In Table 3, we visualize 11 representative anisotropic blur kernels for evaluation. The noise level is set to 0, 5, and 10. Even in noise-free conditions, IKC can no longer cope with anisotropic blur and cannot be compared with ACT, AdaTarget, and DASR. Our method can lead the second place DASR by 0.54 dB, which proves our superiority under more complicated degradations. As noise level increases, all methods have performance drop. At a noise of 5, DASR and ACT have comparable performance. Different from the experiment on isotropic blur, CMDSR has shown its strength in deblurring and denoising. CMDSR can surpass AdaTarget and IKC. Since degradation has become more severe, we can still ahead of ACT by 0.3 dB at the noise of 5 and 0.21 dB at the noise of 10. This demonstrates the robustness of our representations under severe noise.

Table 4 shows a more comprehensive comparison of DOTA-T and Jilin-T. AdaTarget and DASR achieved second place, respectively. Our approach achieves the best results in all test sets, illustrating our good generalization.

Quantitative Results. In the “parking_137” images shown in Fig. 9. As the noise level increases, the reconstructed visual effects all decrease. Among them, IKC contains obvious noise due to the inherent design that IKC can only handle blur. The restored car in CMDSR is difficult to distinguish. ACT is slightly better, but the boundary is badly mixed with the ground. AdaTarget and DASR are more sharpened than ACT, but AdaTarget seems a little irregular in the shape of cars. Our method is nicely immune to all types of noise and visually satisfying. In the DOTA-T image of Fig. 10, our approach is undoubtedly the best as well.

4.5. Experiments on real degradations

In the real-world experiment, we will directly super-resolve the ground truth images to predict the latent high-resolution counterparts. The degradation contained in the real image is unknown. Here we adopt a reference-free indicator NIQE [80] to evaluate these methods objectively. The smaller the NIQE, the more the image matches the human eye perception. Thanks to our self-supervised degradation representation mechanism, we can first learn the potential degradations on real-world images without ground-truth degradation labels. In particular, we train our encoder in the synthesized dataset with anisotropic blur and noise for the first 50 epochs. After that, we train the encoder

Table 4

The average PSNR/SSIM results of experiments on noises and anisotropic Gaussian blur. We test these models on DOTA-T and Jilin-T with 11 typical kernel widths and noise levels from 0 to 10. The results under each degradation are then averaged. The best result is shown in **bold**, and the second place is marked by underline.

Dataset	Method	Noise level			Average
		0	5	10	
DOTA-T	ACT	27.65/0.7529	26.68/ <u>0.7026</u>	<u>26.02/0.6743</u>	26.78/0.7099
	IKC	27.27/0.7506	25.36/0.6271	24.34/0.5409	25.66/0.6395
	AdaTarget	<u>27.93/0.7638</u>	<u>26.69/0.7020</u>	25.88/0.6647	<u>26.84/0.7102</u>
	DASR	27.61/0.7515	26.58/0.6992	25.99/0.6735	26.73/0.7081
	CMDSR	26.82/0.7227	26.20/0.6836	25.69/0.6598	26.24/0.6887
	Ours	28.59/0.7784	27.23/0.7195	26.42/0.6881	27.41/0.7287
JILIN-T	ACT	31.82/0.9152	30.11/0.8660	29.48/0.8380	30.47/0.8730
	IKC	30.23/0.8912	28.52/0.7895	26.75/0.6950	28.50/0.7919
	AdaTarget	30.50/0.8938	28.79/0.8321	27.45/0.7878	28.91/0.8379
	DASR	<u>32.12/0.9235</u>	30.81/0.8726	29.24/0.8354	<u>30.72/0.8771</u>
	CMDSR	30.32/0.8832	29.58/0.8562	28.58/0.8272	29.49/0.8555
	Ours	32.92/0.9250	<u>30.74/0.8743</u>	<u>29.28/0.8401</u>	30.98/0.8798

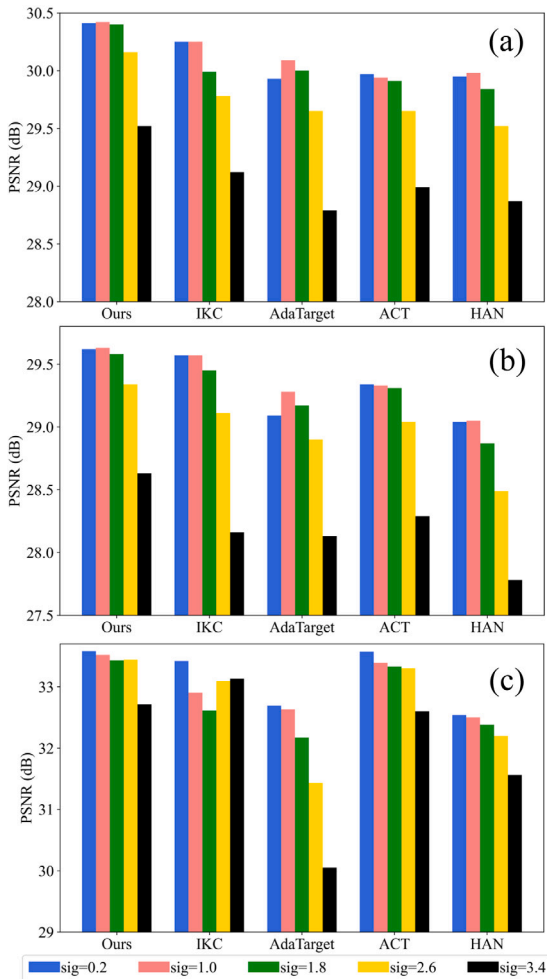


Fig. 8. The PSNR performance on AID-T (a), DOTA-T (b), and Jilin-T (c). The kernel width σ is set from 0.2 to 3.4.

with real data for another 50 epochs to enhance real generalizability. Note that we do not need real degradation labels to train our encoder. And finally, we train the whole model with the synthesized training set for 600 epochs. The visual results are shown in Fig. 11. We can see that IKC produces much noisy artifacts since it can not implement denoising. Our results are relatively cleaner and have less blur. NIQE metrics also demonstrate we can generate objectively visual-pleasing results.

Table 5

Ablation experiments on model designs. The PSNR is calculated on isotropic Gaussian blur with kernel width of 3.4. The best PSNR is shown in **bold** and the second place is marked by underline.

Method	Contrastive learning	Dual-wise modulation	ℓ_2 Norm	PSNR/dB
Model-1	✗	✓	✓	25.76
Model-2 (Ours)	✓	✓	✓	28.11
Model-3 (w SFT Layer)	✓	✗	✓	27.92
Model-4 (w DACov)	✓	✗	✓	<u>27.96</u>
Model-5	✓	✓	✗	Fail

4.6. Ablation study

In this section, we discuss the hyperparameters setting and the effectiveness of the key components in our model. Note that all the experiments are conducted on noise-free and isotropic Gaussian blur degradation settings.

(1) **The number of DMB and DMG.** As shown in Fig. 13(e) and (f), the reconstruction performance improves as the number of DMB N_{DMB} increases but reaches its bottleneck at $N_{DMB} > 3$. The same goes for the number of DMG N_{DMG} , where there is a tiny improvement in PSNR after $N_{DMG} > 3$. To save computational costs while maintaining the capability of CNN, we cascade five DMBs and DMGs in our final model.

(2) **The number of batchsize.** Many studies have confirmed that an adequate amount of negative samples is crucial for contrastive learning. In this paper, the negative images are located in a minibatch. In Fig. 13(b), we find that the PSNR even decreases when the batchsize is larger than 8. On the one hand, this demonstrates that our multi-view sample augmentation strategy can enrich enough negative samples. On the other hand, it illustrates that excessive negative samples may increase the pressure of contrastive learning and thus reduce performance.

(3) **Temperature hyper-parameter.** Following the setting in SimCLR [75], we discuss different temperature parameters ($\tau = 0.05, 0.1, 0.5, 1$) used in the contrastive loss. The PSNR results can be found in Table 6. Our encoder can better express the degradations at $\tau = 0.1$ and lead to the best performance.

(4) **Self-supervised Degradation Representation.** To explore the effectiveness of our self-supervised degenerate representations, we set up a model (Model-1) which throws away the contrastive learning loss $\mathcal{L}_{encoder}$ during model training. In this case, the degradation representations are adaptively learned by the encoder in an end-to-end manner. The training process is shown in Fig. 13(a) and the quantitative results are listed in Table 5. Without contrastive learning, Model-1 can hardly distinguish various degradations, thus dropping 2.37 dB in PSNR compared with our final model (25.76 dB v.s. 28.11 dB). This

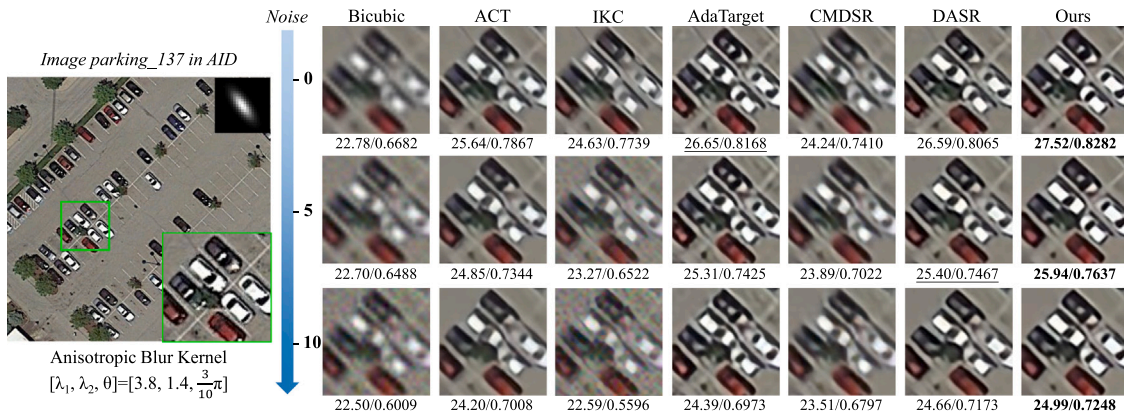


Fig. 9. The visual comparison of experiments on noises and anisotropic Gaussian blur. This image “parking_137” is picked from AID-T. The best PSNR/SSIM is shown in **bold** below the image, and the second place is marked by underline.



Fig. 10. The visual comparison of experiments on noises and anisotropic Gaussian blur. This image is picked from DOTA-T and suffers from anisotropic blur kernel of $[\lambda_1, \lambda_2, \theta] = [3.8, 1.8, \frac{7}{10}\pi]$ and noise level of 5. The best PSNR/SSIM is shown in **bold** below the image, and the second place is marked by underline.

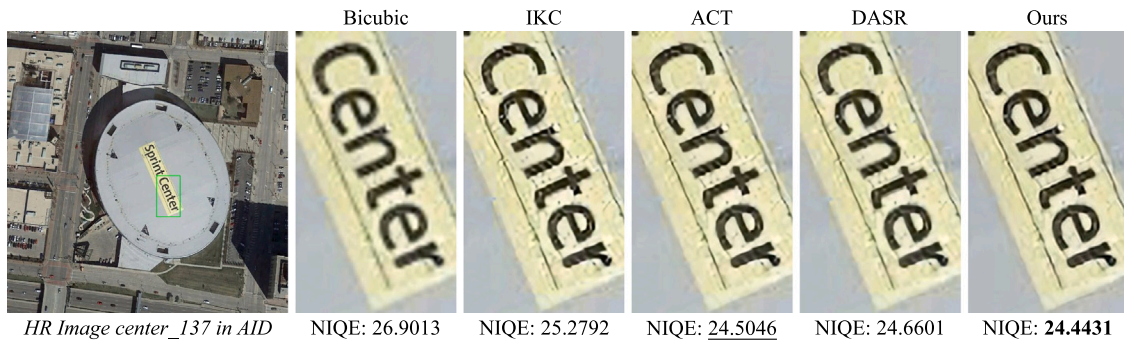


Fig. 11. The visual comparison of experiments on real degradations. The best NIQE is shown in **bold**, and the second place is marked by underline.

fully illustrates that our degradation representations can predict precise degradation priors and synergistically help super-resolution procedures adapt to various degradation distributions.

In Fig. 12, we visualize the degraded representations in AID-T with t-SNE [81] algorithm. In detail, we degraded AID-T with different Gaussian blur kernels and noises. Then we encode them with our

Table 6

Ablation experiments on hyperparameters. The PSNR is calculated on isotropic Gaussian blur with a kernel width of 3.4. The best PSNR is shown in **bold**.

Mini-Batch	PSNR	Temperature (τ)	PSNR	N_{DMB}	PSNR	N_{DMG}	PSNR
4	28.10	0.05	27.98	1	27.73	1	27.66
8	28.11	0.1	28.11	2	27.96	5	28.11
16	28.06	0.5	28.04	3	28.09	10	28.12
32	28.02	1	27.85	5	28.11	15	28.09

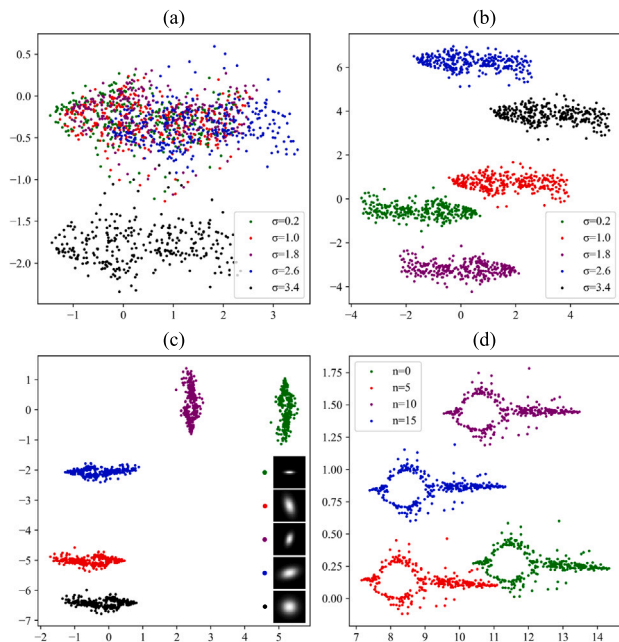


Fig. 12. The t-SNE visualization of representations for various degradations. (a) illustrates representations generated from five isotropic Gaussian blur kernels without degradation representation learning (Model-1). (b) illustrates representations generated from five isotropic Gaussian blur kernels by our method (Model-2). (c) presents representations generated from five isotropic Gaussian blur kernels, and (d) displays representations derived from four noises.

encoder to see whether we can distinguish various degradations in the embedding space. The representations in (a) are mixed and indistinguishable, while our method (2) can cluster them well with contrast learning. Furthermore, the visualization results on anisotropic blur kernels and noises are also displayed in (3) and (4). With the guide of discriminative representations, our dual-wise modulation network can learn how to adapt to diverse degradation distributions, thus leading a good generalization on real-world scenes with multiple degradations.

(5) ℓ_2 normalization of embeddings. The embeddings $\{z_k\}_{k=1}^{2N}$ need to be normalized by ℓ_2 function. Otherwise, we will fail in model training because of gradient explosion. Besides, we plot the correlation matrices of five embeddings. They are derived from the same image but contain five isotropic Gaussian kernels. In Fig. 14, there is a low correlation between embeddings of different degradations, which indicates we can properly identify different blurs in the embedding space.

(6) **Dual-wise Feature Modulation** We compare the proposed dual-wise modulation network (DMN) with two state-of-the-art feature adaption approaches. Firstly, we replace our DMN with the spatial-feature transformation layer (SFT Layer) network used in IKC. It can integrate the blur kernel estimation into feature space by an affine transformation. Secondly, we adopt the degradation-aware convolution (DACov) network proposed in DASR to realize feature adaption. Their training processes are shown in Fig. 13(c). Our dual-wise modulation network can gather higher PSNR by 0.15 dB than DACov, which demonstrates the proposed DMN is valid for feature adaption to the

desired domain. In a word, we achieve a shallow-to-deep modulation on feature dimension to mitigate the over-influence of LR features. And the channel-wise modulation indeed helps us realize effective feature adaption on the channel dimension.

(7) **Model efficiency.** In Table 2, we calculate the number of parameters, the floating-point operations (FLOPs), and the test time for each method. Here, the FLOPs is calculated with an LR input of size 150×150 . The test time is the total time taken to complete the test on three test sets. Compared to the second-place IKC in AID-T, the proposed model has a 39.3% reduction in the parameters (5.74M v.s. 9.45M), a 15.35G drop in the FLOPs (115.23G v.s. 130.58G), but outperforms it by 0.3 dB on average. Thanks to the equipment of shallow ResNet and an efficient dual-wise modulation network, we reach a favorable tradeoff between performance and efficiency.

5. Conclusion

In this paper, we investigate the blind SISR problem for real-world remote sensing imagery with multiple degradations. Inspired by self-supervised learning, we developed a degradation-guided adaptive network to learn how to realize feature adaption to a target domain. Specifically, we use a shallow ResNet to encode and project multi-view samples into embedding space for contrastive learning. Thanks to the robust representations derived from self-supervised learning, we can receive a precise degradation estimation in a label-free manner to collaboratively guide the downstream SR network. Besides, an effective dual-wise modulation network was proposed to realize transformation on feature and channel dimensions. Extensive experiments on three mainstream remote sensing datasets demonstrate that our framework can handle diverse degradation distributions with a single forward propagation.

In future work, we will further establish a sophisticated degradation learning approach to consider the specialties of remote sensing images, such as degradation variations among multi-scale objects in remote sensing images.

CRedit authorship contribution statement

Yi Xiao: Methodology, Data processing, Analysis of experiment, Writing – original draft. **Qiangqiang Yuan:** Supervision. **Kui Jiang:** Data processing. **Jiang He:** Data processing. **Yuan Wang:** Data processing. **Liangpei Zhang:** Supervision.

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.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61971319,42230108) and the Hubei Science Foundation for Distinguished Young Scholars (2020CFA051).

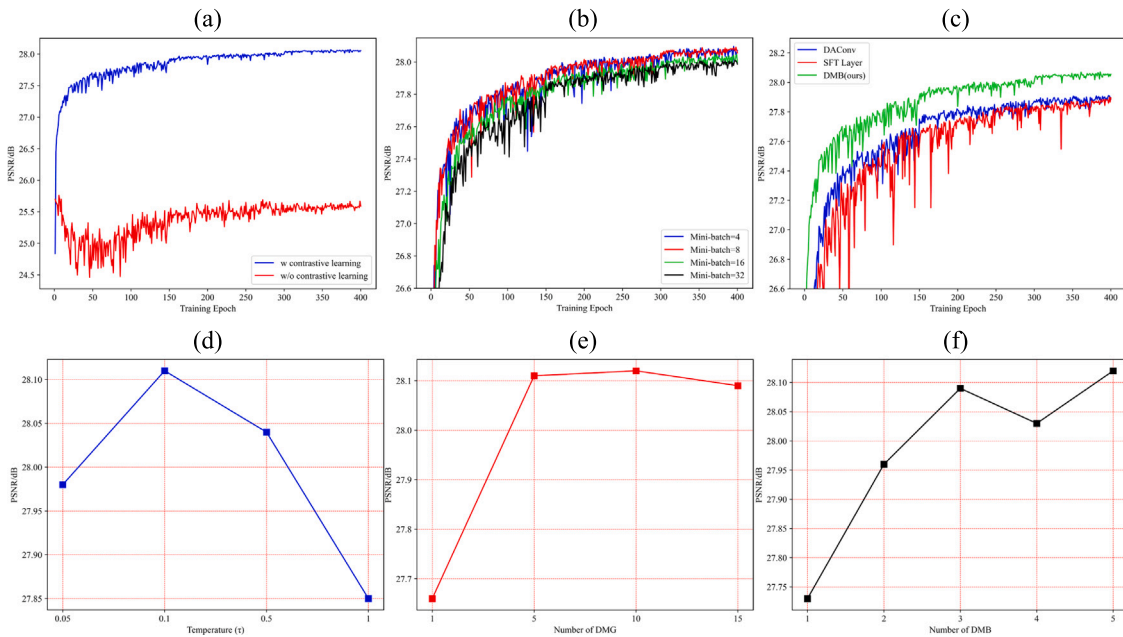


Fig. 13. The training process of ablation experiments. (a) is about degenerate representational learning. (b) is about the number of batchsize. (c) is involved in different modulation approaches. (d) is temperature parameter. (e) and (f) represent the number of DMG and DMB, respectively.

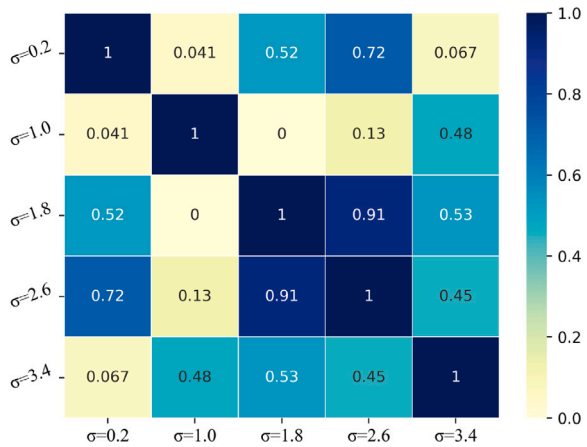


Fig. 14. The correlation map of embeddings under five isotropic Gaussian blur kernels.

References

[1] J. He, Q. Yuan, J. Li, L. Zhang, PoNet: A universal physical optimization-based spectral super-resolution network for arbitrary multispectral images, *Inf. Fusion* 80 (2022) 205–225.

[2] G. Vivone, Multispectral and hyperspectral image fusion in remote sensing: A survey, *Inf. Fusion* 89 (2023) 405–417.

[3] X. Gu, C. Zhang, Q. Shen, J. Han, P.P. Angelov, P.M. Atkinson, A self-training hierarchical prototype-based ensemble framework for remote sensing scene classification, *Inf. Fusion* 80 (2022) 179–204.

[4] B. Rasti, P. Ghamisi, Remote sensing image classification using subspace sensor fusion, *Inf. Fusion* 64 (2020) 121–130.

[5] L. Huang, W. Zhao, A.W.-C. Liew, Y. You, An evidential combination method with multi-color spaces for remote sensing image scene classification, *Inf. Fusion* (2023).

[6] Z. Zheng, Y. Zhong, J. Wang, A. Ma, L. Zhang, Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters, *Remote Sens. Environ.* 265 (2021) 112636.

[7] J. Wang, A. Ma, Y. Zhong, Z. Zheng, L. Zhang, Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: From airborne to spaceborne imagery, *Remote Sens. Environ.* 277 (2022) 113058.

[8] D. Liu, J. Li, Q. Yuan, L. Zheng, J. He, S. Zhao, Y. Xiao, An efficient unfolding network with disentangled spatial-spectral representation for hyperspectral image super-resolution, *Inf. Fusion* (2023).

[9] J. Wang, C. Tang, Z. Li, X. Liu, W. Zhang, E. Zhu, L. Wang, Hyperspectral band selection via region-aware latent features fusion based clustering, *Inf. Fusion* 79 (2022) 162–173.

[10] Q. Zhang, Q. Yuan, M. Song, H. Yu, L. Zhang, Cooperated spectral low-rankness prior and deep spatial prior for HSI unsupervised denoising, *IEEE Trans. Image Process.* 31 (2022) 6356–6368.

[11] X. Wang, J. Ma, P. Yi, X. Tian, J. Jiang, X.-P. Zhang, Learning an epipolar shift compensation for light field image super-resolution, *Inf. Fusion* 79 (2022) 188–199.

[12] S. Aymaz, C. Köse, A novel image decomposition-based hybrid technique with super-resolution method for multi-focus image fusion, *Inf. Fusion* 45 (2019) 113–127.

[13] J. Li, W. Guan, Adaptive lq-norm constrained general nonlocal self-similarity regularizer based sparse representation for single image super-resolution, *Inf. Fusion* 53 (2020) 88–102.

[14] P. Yi, Z. Wang, K. Jiang, Z. Shao, J. Ma, Multi-temporal ultra dense memory network for video super-resolution, *IEEE Trans. Circuits Syst. Video Technol.* 30 (8) (2019) 2503–2516.

[15] M. Hu, K. Jiang, L. Liao, J. Xiao, J. Jiang, Z. Wang, Spatial-temporal space hand-in-hand: Spatial-temporal video super-resolution via cycle-projected mutual learning, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3574–3583.

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

[17] Y. Wang, J. Zhao, DIC-net: Upgrade the performance of traditional DIC with Hermite dataset and convolution neural network, *Opt. Lasers Eng.* 160 (2023) 107278.

[18] Y. Wang, Q. Yuan, S. Zhou, L. Zhang, Global spatiotemporal completion of daily high-resolution TCCO from TROPOMI over land using a swath-based local ensemble learning method, *ISPRS J. Photogramm. Remote Sens.* 194 (2022) 167–180.

[19] Y. Wang, Q. Yuan, T. Li, L. Zhu, Global spatiotemporal estimation of daily high-resolution surface carbon monoxide concentrations using deep forest, *J. Clean. Prod.* 350 (2022) 131500.

[20] Y. Wang, Q. Yuan, L. Zhu, L. Zhang, Spatiotemporal estimation of hourly 2-km ground-level ozone over China based on himawari-8 using a self-adaptive geospatially local model, *Geosci. Front.* 13 (1) (2022) 101286.

- [21] Q. Zhang, Q. Yuan, T. Jin, M. Song, F. Sun, SGD-SM 2.0: an improved seamless global daily soil moisture long-term dataset from 2002 to 2022, *Earth Syst. Sci. Data* 14 (10) (2022) 4473–4488.
- [22] Q. Zhang, Q. Yuan, Z. Li, F. Sun, L. Zhang, Combined deep prior with low-rank tensor SVD for thick cloud removal in multitemporal images, *ISPRS J. Photogramm. Remote Sens.* 177 (2021) 161–173.
- [23] Y. Xiao, Y. Wang, Q. Yuan, J. He, L. Zhang, Generating a long-term (2003–2020) hourly 0.25° global PM_{2.5} dataset via spatiotemporal downscaling of CAMS with deep learning (DeepCAMS), *Sci. Total Environ.* 848 (2022) 157747.
- [24] H. Wang, Q. Yuan, H. Zhao, H. Xu, In-situ and triple-collocation based assessments of CYGNSS-R soil moisture compared with satellite and merged estimates quasi-globally, *J. Hydrol.* 615 (2022) 128716.
- [25] Q. Yang, Q. Yuan, L. Yue, T. Li, H. Shen, L. Zhang, Mapping PM_{2.5} concentration at a sub-km level resolution: A dual-scale retrieval approach, *ISPRS J. Photogramm. Remote Sens.* 165 (2020) 140–151.
- [26] Q. Yang, Q. Yuan, M. Gao, T. Li, A new perspective to satellite-based retrieval of ground-level air pollution: Simultaneous estimation of multiple pollutants based on physics-informed multi-task learning, *Sci. Total Environ.* 857 (2023) 159542.
- [27] M. Li, Q. Yang, Q. Yuan, L. Zhu, Estimation of high spatial resolution ground-level ozone concentrations based on Landsat 8 TIR bands with deep forest model, *Chemosphere* 301 (2022) 134817.
- [28] C. Chen, Z. Xiong, X. Tian, Z.-J. Zha, F. Wu, Camera lens super-resolution, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1652–1660.
- [29] J. Cai, H. Zeng, H. Yong, Z. Cao, L. Zhang, Toward real-world single image super-resolution: A new benchmark and a new model, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3086–3095.
- [30] M. Galar, R. Sesma, C. Ayala, L. Albizua, C. Aranda, Super-resolution of sentinel-2 images using convolutional neural networks and real ground truth data, *Remote Sens.* 12 (18) (2020) 2941.
- [31] R. Dong, L. Mou, L. Zhang, H. Fu, X.X. Zhu, Real-world remote sensing image super-resolution via a practical degradation model and a kernel-aware network, *ISPRS J. Photogramm. Remote Sens.* 191 (2022) 155–170.
- [32] L. Wang, Y. Wang, X. Dong, Q. Xu, J. Yang, W. An, Y. Guo, Unsupervised degradation representation learning for blind super-resolution, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10581–10590.
- [33] G. Yin, W. Wang, Z. Yuan, W. Ji, D. Yu, S. Sun, T.-S. Chua, C. Wang, Conditional hyper-network for blind super-resolution with multiple degradations, *IEEE Trans. Image Process.* 31 (2022) 3949–3960.
- [34] S. Bell-Kligler, A. Shocher, M. Irani, Blind super-resolution kernel estimation using an internal-gan, *Adv. Neural Inf. Process. Syst.* 32 (2019).
- [35] X. Luo, J. Peng, K. Xian, Z. Wu, Z. Cao, Defocus to focus: Photo-realistic bokeh rendering by fusing defocus and radiance priors, *Inf. Fusion* 89 (2023) 320–335.
- [36] I. Bakkouri, K. Afdel, Computer-aided diagnosis (CAD) system based on multi-layer feature fusion network for skin lesion recognition in dermoscopy images, *Multimedia Tools Appl.* 79 (29–30) (2020) 20483–20518.
- [37] I. Bakkouri, K. Afdel, Multi-scale CNN based on region proposals for efficient breast abnormality recognition, *Multimedia Tools Appl.* 78 (2019) 12939–12960.
- [38] I. Bakkouri, K. Afdel, MLCA2F: Multi-level context attentional feature fusion for COVID-19 lesion segmentation from CT scans, *Signal Image Video Process.* (2022) 1–8.
- [39] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, *IEEE Trans. Pattern Anal. Mach. Intell.* 38 (2) (2015) 295–307.
- [40] J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.
- [41] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, T. Huang, Wide activation for efficient and accurate image super-resolution, 2018, arXiv preprint arXiv:1808.08718.
- [42] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: *Proceedings of the European Conference on Computer Vision, ECCV*, 2018, pp. 286–301.
- [43] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, H. Shen, Single image super-resolution via a holistic attention network, in: *European Conference on Computer Vision*, Springer, 2020, pp. 191–207.
- [44] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, Esrgan: Enhanced super-resolution generative adversarial networks, in: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.
- [45] Z.-S. Liu, L.-W. Wang, C.-T. Li, W.-C. Siu, Y.-L. Chan, Image super-resolution via attention based back projection networks, in: *2019 IEEE/CVF International Conference on Computer Vision Workshop, ICCVW, IEEE*, 2019, pp. 3517–3525.
- [46] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472–2481.
- [47] S. Anwar, N. Barnes, Densely residual Laplacian super-resolution, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (3) (2022) 1192–1204, <http://dx.doi.org/10.1109/TPAMI.2020.3021088>.
- [48] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, Swinir: Image restoration using swin transformer, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.
- [49] X. Chen, X. Wang, J. Zhou, C. Dong, Activating more pixels in image super-resolution transformer, 2022, arXiv preprint arXiv:2205.04437.
- [50] K. Zhang, W. Zuo, L. Zhang, Learning a single convolutional super-resolution network for multiple degradations, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3262–3271.
- [51] K. Zhang, W. Zuo, L. Zhang, Deep plug-and-play super-resolution for arbitrary blur kernels, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1671–1681.
- [52] K. Zhang, L.V. Gool, R. Timofte, Deep unfolding network for image super-resolution, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3217–3226.
- [53] J. Gu, H. Lu, W. Zuo, C. Dong, Blind super-resolution with iterative kernel correction, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1604–1613.
- [54] Y. Huang, S. Li, L. Wang, T. Tan, et al., Unfolding the alternating optimization for blind super resolution, *Adv. Neural Inf. Process. Syst.* 33 (2020) 5632–5643.
- [55] Z. Luo, Y. Huang, S. Li, L. Wang, T. Tan, Learning the degradation distribution for blind image super-resolution, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6063–6072.
- [56] Z. Luo, H. Huang, L. Yu, Y. Li, H. Fan, S. Liu, Deep constrained least squares for blind image super-resolution, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17642–17652.
- [57] R. Zhou, S. Susstrunk, Kernel modeling super-resolution on real low-resolution images, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2433–2443.
- [58] 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*, 2018, pp. 3118–3126.
- [59] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, L. Lin, Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 701–710.
- [60] A. Bulat, J. Yang, G. Tzimiropoulos, To learn image super-resolution, use a gan to learn how to do image degradation first, in: *Proceedings of the European Conference on Computer Vision, ECCV*, 2018, pp. 185–200.
- [61] Y. Zhou, W. Deng, T. Tong, Q. Gao, Guided frequency separation network for real-world super-resolution, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 428–429.
- [62] M. Fritsche, S. Gu, R. Timofte, Frequency separation for real-world super-resolution, in: *2019 IEEE/CVF International Conference on Computer Vision Workshop, ICCVW, IEEE*, 2019, pp. 3599–3608.
- [63] H. Wu, N. Ni, S. Wang, L. Zhang, Blind super-resolution for remote sensing images via conditional stochastic normalizing flows, 2022, arXiv preprint arXiv:2210.07751.
- [64] Z. Shi, C. Chen, Z. Xiong, D. Liu, Z.-J. Zha, F. Wu, Deep residual attention network for spectral image super-resolution, in: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.
- [65] J.M. Haut, R. Fernandez-Beltran, M.E. Paoletti, J. Plaza, A. Plaza, Remote sensing image superresolution using deep residual channel attention, *IEEE Trans. Geosci. Remote Sens.* 57 (11) (2019) 9277–9289.
- [66] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, J. Jiang, Edge-enhanced GAN for remote sensing image superresolution, *IEEE Trans. Geosci. Remote Sens.* 57 (8) (2019) 5799–5812.
- [67] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, L. Zhang, Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection, *IEEE Trans. Geosci. Remote Sens.* 60 (2022) 1–19, <http://dx.doi.org/10.1109/TGRS.2021.3107352>.
- [68] Y. Xiao, Q. Yuan, J. He, Q. Zhang, J. Sun, X. Su, J. Wu, L. Zhang, Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial-temporal transformer, *Int. J. Appl. Earth Obs. Geoinf.* 108 (2022) 102731.
- [69] P. Wang, H. Zhang, F. Zhou, Z. Jiang, Unsupervised remote sensing image super-resolution using cycle CNN, in: *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, IEEE*, 2019, pp. 3117–3120.
- [70] N. Zhang, Y. Wang, X. Zhang, D. Xu, X. Wang, G. Ben, Z. Zhao, Z. Li, A multi-degradation aided method for unsupervised remote sensing image super resolution with convolution neural networks, *IEEE Trans. Geosci. Remote Sens.* 60 (2020) 1–14.
- [71] L. Zhang, J. Nie, W. Wei, Y. Li, Y. Zhang, Deep blind hyperspectral image super-resolution, *IEEE Trans. Neural Netw. Learn. Syst.* 32 (6) (2021) 2388–2400, <http://dx.doi.org/10.1109/TNNLS.2020.3005234>.
- [72] X. Kang, J. Li, P. Duan, F. Ma, S. Li, Multilayer degradation representation-guided blind super-resolution for remote sensing images, *IEEE Trans. Geosci. Remote Sens.* 60 (2022) 1–12.
- [73] D. Mishra, O. Hadar, Self-FuseNet: Data free unsupervised remote sensing image super-resolution, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* (2023).

- [74] W. Shi, J. Caballero, F. Huszár, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1874–1883.
- [75] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.
- [76] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, X. Lu, AID: A benchmark data set for performance evaluation of aerial scene classification, *IEEE Trans. Geosci. Remote Sens.* 55 (7) (2017) 3965–3981.
- [77] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, Object detection in aerial images: A large-scale benchmark and challenges, *IEEE Trans. Pattern Anal. Mach. Intell.* (2021) 1, <http://dx.doi.org/10.1109/TPAMI.2021.3117983>.
- [78] J. Yoo, T. Kim, S. Lee, S.H. Kim, H. Lee, T.H. Kim, Enriched CNN-transformer feature aggregation networks for super-resolution, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2023.
- [79] Y. Jo, S.W. Oh, P. Vajda, S.J. Kim, Tackling the ill-posedness of super-resolution through adaptive target generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16236–16245.
- [80] A. Mittal, R. Soundararajan, A.C. Bovik, Making a “completely blind” image quality analyzer, *IEEE Signal Process. Lett.* 20 (3) (2012) 209–212.
- [81] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* 9 (11) (2008).