DSMA: Reference-Based Image Super-Resolution Method Based on Dual-View Supervised Learning and Multi-Attention Mechanism

XIN LIU¹, JING LI¹, TINGTING DUAN², JIANGTAO LI², AND YE WANG³

¹College of Computer Science and Technology/College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
²State grid electronic commerce co.LTD, Beijing 100053, China
³Nanjing Nari Information Communication Technology Co., Ltd, Nanjing 210003, China

Corresponding author: Jing Li (lijing@nuaa.edu.cn)

This work was supported by the science and technology program (5700-202152169A-0-0-00) of State Grid Corporation of China

ABSTRACT Referenced image based super-resolution methods (RefSR) have made rapid and remarkable progress in the field of image super-resolution (SR) in recent years by introducing additional high-resolution (HR) images to enhance the recovery of low-resolution (LR) images. The existing RefSR methods can rely on implicit correspondence matching to transfer the HR texture from the reference image (Ref) to compensate for the information loss in the input image. However, the differences between low-resolution input images and high-resolution reference images still affect the effective utilization of Ref images, so it is an important challenge to make full use of the information in Ref images to improve the SR performance. In this paper, we propose an image super-resolution method based on dual-view supervised learning and multi-attention mechanism (DSMA). It enhances the learning of important detail features of Ref images and weakens the interference of noisy information by introducing the multi-attention mechanism, while employing dual-view supervision to motivate the network to learn more accurate feature representations. Quantitative and qualitative experiments on these benchmarks, i.e., CUFED5, Urban100 and Manga109, show that DSMA outperforms the state-of-the-art baselines with significant improvements.

INDEX TERMS Computer vision, image enhancement, image reconstruction, machine learning, super resolution.

I. INTRODUCTION Image Super-Resolution (SR) is a fundamental computer vision task that aims to recover natural high-frequency details from a given low-resolution image [1]. It is widely used in the real world in some fields that require high image quality, such as medical imaging, satellite surveying, surveillance, security, etc. [2]–[4], and it helps to improve other computer vision tasks. In general, the problem is very challenging and inherently ill-posed due to the lack of information inherent in LR images and HR images.

The study of image super-resolution is usually divided into two types: single image super-resolution (SISR) and reference-based image super-resolution (RefSR). Traditional SISR methods are mainly based on interpolation, filtering and dictionary learning [5]–[7], which use manually designed mapping or learning strategies to recover images. Since these approaches are homogeneous and very dependent on human experience, it is difficult to recover satisfactory images when faced with unknown degraded LR images. SRCNN [8] is the first deep learning method applied to the SISR problem, which uses a three-layer convolutional neural network (CNN) to learn the mapping from LR images to HR images to recover images, which provides new insights for exploring SR research. To better learn this mapping relationship from LR images to HR images, Dong et al. [9] improved SRCNN by putting the upsampling operation at the end and using smaller convolutional kernels to extract features, which greatly improved the speed of SR. Kim et al. [10] proposed a 20-layer network to extract more information using a larger perceptual field to generate higher quality images. Ledig et al. [11] introduced residual connectivity to overcome the problem of gradient propagation, which avoids gradient disappearance and gradient explosion during the training of deeper networks.
by jump connectivity and identity mapping. Kim et al. [12] proposed DRCN, which uses a recursive network structure to deepen the number of layers and expand the perceptual field of the network without increasing the network parameters. Ledig et al. [11] proposed SRGAN, which introduces generative adversarial networks to enhance the realism of the recovered images using perceptual loss and adversarial loss. ESRGAN [13] enhances SRGAN by introducing RRDB with relativistic adversarial loss. Although these deep learning models have achieved good results in the application of image super-resolution, these classical SISR methods usually cause blurring effects or visual artifacts due to the inherent lack of information between LR and HR images.

Recently, reference-based image super-resolution (RefSR) has achieved success in the field of SR, which additionally introduces HR images as Ref and provides finer details to the LR image by transferring the texture features of Ref to achieve good reconstruction performance. Traditional RefSR methods such as CrossNet [14] must build on the assumption that the Ref and LR images have similar contents or are well-aligned; otherwise, the effect of super-resolution recovery is severely degraded when Ref significantly differs from the LR image. To address this problem, SRNTT [15] uses a feature space defined by a pre-trained classification model to search and transfer texture features on Ref. This method that does not require the Ref to be well-aligned with the LR image or have similar content, which enable texture transfer. TTsr [16] and MASA [17] further introduce the idea of the transformer [18] to discover the deep feature correspondence between LR and Ref images through the attention mechanism thus transfers more accurate textures. However, some noise information that remains in such texture features will have negatively affect the subsequent fusion with low-resolution image features, network convergence and final results.

To address these problems, we propose a RefSR called DSMA, which enhances the effective use of Ref information, improves the quality of feature propagation, and enables the network to learn more accurate feature representations. The design of DSMA has several advantages. First, this paper is inspired by mixed attention [19] and introduces multi-stage attention, which first achieves high-quality correspondence matching by the first stage of attention and extract similar HR texture features from Ref; then, the obtained HR texture features are reinforced by the second stage of attention to learn important features and suppress the propagation of noisy information. Second, this paper is inspired by knowledge distillation [20] and introduces a dual-view supervised learning strategy to quantify the similarity between the deep features of HR images and the fused features of LR and Ref at different scales, which is applied to the evaluation and loss functions to motivate the network to learn a more accurate feature representation.

The main contributions of this paper are as follows.

1. Based on the idea of mixed attention, we combine the transformer attention and channel attention structures to avoid the interference of noisy information during texture transfer while enhancing the learning of useful features.

2. We propose a novel supervised strategy for dual-view learning, which enables our method to combine supervised signals at both intermediate feature layer and image layer levels to motivate the network to learn more powerful feature representations.

II. RELATED WORK
A. SINGLE IMAGE SUPER-RESOLUTION

As the mainstream method of image super-resolution, SISR has been a hot topic of research in the field of image super-resolution. The SRCNN [8] proposed by Dong et al. first introduced deep learning into the field of image super-resolution; they proposed a three-layer network architecture to achieve image super-resolution by learning the mapping function of LR-HR through CNN, and achieved better results than traditional methods (e.g., interpolation, dictionary learning, etc.) [5]–[7]. Afterwards, more network structures were explored to improve the performance of SR networks. Kim et al. proposed a deeper network to generate higher-quality images. The SRRResNet [11] network applied the residual block structure to the super-resolution problem and avoided the problem of gradient disappearance and gradient explosion caused by networks that were too deep. Tong et al. [21] introduced the dense block structure combining low-level features and high-level features to improve the performance of super-resolution. Zhang et al. [22] further combined residual and dense blocks to form an RDB module to enhance the flow of information and gradients. Zhang et al. [23] proposed RCAN, which introduced a new idea of attention to enhance the performance of super-resolution reconstruction by assigning different attention weights to the feature information learned by the network. These methods use mean square error (MSE) or mean absolute error (MAE) as their objective function which ignores human perceptions. SRGAN [11] introduced generative adversarial networks, which use perceptual loss and adversarial loss to improve the realism of images. Recently, RSRGAN [24] was proposed to introduce Ranker mechanism to optimize the perceptual loss to obtain more advanced visual effects.

B. REFERENCE-BASED SUPER-RESOLUTION

Different from SISR, RefSR does not rely only on the prior knowledge obtained by the model to recover the image, it also enhances image recovery by additional high-resolution information provided by the introduced reference image. It is because of this property that RefSR overcomes the inherent information deficiency between LR images and HR images and makes the images that are recovered by the model more realistic, so RefSR has become a hot research topic in recent years.

Although reference-based image super-resolution can provide a large amount of high-frequency information, there are scale and content differences between the input low-resolution image and the high-resolution reference image due to parallax and resolution, which makes it difficult to directly
utilize high-frequency details. Therefore, how to establish the correspondence between LR and Ref images becomes the key to the success of RefSR. One branch of RefSR performs spatial alignment between LR and Ref images. Zheng et al. [14] proposed a cross-scale warping layer using optical flow alignment to perform reverse warping operation on Ref features to obtain spatially aligned features with LR features, and subsequently send LR and transformed Ref features to the decoder for fusion, however, the flow is obtained by a pre-trained network, which leads to computationally intensive and inaccurate estimation. Shim et al. [25] further proposed to overcome the shortcomings of optical flow estimation by using deformable convolution to align and extract Ref features, but the performance of this method heavily depends on the quality of alignment between LR and Ref images and is limited in finding long-distance counterparts.

Another branch of RefSR follows the idea of patch matching. Zhang et al. [15] proposed SRTNN, which is an adaptive patch matching approach to swap similar texture features using VGG features of LR and Ref images for patch matching, which ensures stronger robustness even when providing uncorrelated Ref images. However, it ignores the correlation between LR features and swapped features and provides all swapped features equally to the main network. Recent works TTSSR [16] and MASA [17] introduce the idea of transformer [18] to more accurately mine the deep correspondence in the LR-Ref feature space and transfer texture features in Ref. MASA also designed a coarse-to-fine matching scheme to reduce the computational effort in the matching process. However, these methods ignore the fact that the transferred features bring a large amount of noisy information while providing more useful information, which can also affect the performance of SR if they take up a large amount of computational resources. Therefore, further screening of these feature information is needed to achieve a more reasonable utilization.

### C. ATTENTIONAL MECHANISMS

In recent years, attention mechanisms have been shown to play a superior role in several computer vision tasks such as image classification [26], [27], semantic segmentation [28], target detection [29] and image super-resolution [16], [23].

Hu et al. [27] proposed a compression and excitation block, which significantly improves the performance of image classification by explicitly modeling the interdependence between channels and adaptively recalibrating the channel feature responses. Liu et al. [30] proposed a non-local operation to compute the interaction between any two positions in space to capture the long-range dependencies. Woo et al. proposed a lightweight attention module CBAM [19] that could perform attention in both channel and space dimensions with almost no increase in parameters. Yang et al. [16] proposed a texture transformation network for image super-resolution, where the spatial attention between LR and Ref images is calculated to transfer the texture information.

Although current advanced RefSR algorithms use spatial attention to enhance the ability to transfer features from Ref to compensate for LR images, the differences between low-resolution input images and high-resolution reference images still pose a great challenge to the effective utilization of Ref images. To address this problem, we propose a method called DSMA, which introduces a multi-attention mechanism to enhance the learning of important features from Ref images while using a dual-view supervised excitation network to learn a more accurate feature representation.

### III. OUR APPROACH

In this section, we introduce the proposed image super-resolution method based on two-view supervised learning and a multi-attention mechanism (DSMA). Then, we introduce the overall structure of the network in Subsection A, our proposed multi-attention mechanism in Subsection B, dual-view supervised learning strategy in Subsection C, and experimental details in Subsection D.

#### A. NETWORK STRUCTURE

As shown in Fig.1, our proposed DSMA mainly consists of three parts: a Multiscale feature extraction module, a Multi-attention module and a Decoding and reconstruction module. HR, LR, Ref, SR denote the high-resolution image, the low-resolution image, the high-resolution reference image, the 4x bicubic-downsampled reference image and the final generated super-resolution reconstructed image, respectively. Unlike the previous implicit matching-based RefSR approach, we also use the information from the HR images to supervise our model learning. The multiscale feature extraction module consists of an encoder that contains three building blocks, which can extract features for three resolution scales (1x, 2x and 4x). Here the HR and Ref images go through the encoder to obtain three different scaled features, while the LR image and Ref image only go through the first block of the encoder to produce $F_{LR}$ and $F_{Ref}$. The multi-attention module consists of two parts: the feature selection attention module (FSAM) and the feature adaptive attention module (FAAM). The feature selection attention module can establish the correspondence matching of $F_{Ref}$ and $F_{LR}$ and search for available textures features in $F_{Ref}$ through this correspondence. The feature adaptive attention module can further optimize the features searched in the previous part to obtain the features $F_{Ref}$. The decoding and reconstruction module mainly consists of a decoding network that contains some convolutional layers, residual blocks and upsampling layers, which can fuse $F_{LR}$ and $F_{Ref}$ across resolutions and scales; finally, the fused features $F_{fusion}$ are compensated to the LR image to obtain the final super-resolution image. In addition, to obtain a more accurate feature representation, our DSMA model imposes two levels of supervisory signals: the intermediate feature level, where we constrain the distance between the fused features and the fused features to expect a similar distribution as the HR image, and the final image level, where we constrain the distance between SR and HR by three losses.
We unfold both $F_{LR}$ and $F_{Reff}$ into patches, denoted as $q_i$ ($i \in [1, H_{LR} \times W_{LR}]$) and $k_j$ ($j \in [1, H_{Ref} \times W_{Ref}]$), with each position corresponding to one patch. Here $F_{Reff}$ is the first scale feature extracted by the encoder, which has the same size of length and width as the smallest scale $F_{Ref}$. Then for each patch $q_i$ in $F_{LR}$ and $k_j$ in $F_{Reff}$, we calculate the relevance $r_{i,j}$ between these two patches by:

$$r_{i,j} = \langle q_i, k_j \rangle$$

(1)

The relevance is further used to obtain the hard-attention map and soft-attention map.

### Stage 2: Hard Attention

Hard attention is used to select features to be transferred from Ref by the relevance between LR and Ref images. Specifically, we first calculate a hard attention graph $H$ where the $i$-th element $h_i$ ($i \in [1, H_{LR} \times W_{LR}]$) is calculated from the relevance $r_{i,j}$:

$$h_i = \arg \max_j r_{i,j}$$

(2)

where the value of $h_i$ can be considered as an index, which represents the most relevant position in the Ref to the $i$-th position in the LR image. Then we apply an index selection operation to the unfolded patches of $F_{Ref}$ using the obtained hard-attention map as the index:

$$t_i = v_{h_i}$$

(3)

where $t_i$ denotes the value of $T$ in the $i$-th position, which is selected from the $h_i$-th position of $F_{Ref}$. Finally, we aggregate all $t_i$ ($i \in [1, H_{LR} \times W_{LR}]$) to get the HR feature $T$.

### Stage 3: Soft Attention

To avoid treating the information at each position in $T$ equally, we further apply different attention weights to different positions of feature $T$ by using the relevance $r_{i,j}$ of $q_i$ and $k_j$. Specifically, we first calculate a soft attention graph $S$ where the $i$-th element $s_i$ ($i \in [1, H_{LR} \times W_{LR}]$) is calculated from $r_{i,j}$:

$$s_i = \max_j r_{i,j}$$

(4)
second layer of this shared network. First, the two channel descriptors reduces the number of channels by reduction ratio $r$. After being activated by ReLU, it is then increased with ratio $r$ by a channel-upscaling layer. Finally, the channel weights of each layer are mapped between 0,1 by the Sigmoid function. Using this method to downscale and subsequently upscale the channel dimensions, the number of parameters of the network is reduced, while the nonlinear relationship between channel descriptors and channel weights can be better learned by the two activation functions over ReLU and Sigmoid. The obtained weights of each channel further act on the input feature mapping $x$ to achieve a different focus on each channel:

$$y_c = s_c \cdot x_c$$  \hspace{1cm} (9)$$

where $s_c$ and $x_c$ are the weight and feature map in the $c$-th channel.

Finally, we use the residual block to apply the channel attention to the feature $F$ to obtain the final transferred features $\bar{F}_{Ref}$:

$$\bar{F}_{Ref} = F + CAM (W_2 \delta (W_1 F))$$  \hspace{1cm} (10)$$

where $W_1$ and $W_2$ are the weights of the two convolutional layers in the residual block, $CAM$ denotes the channel attention operation.

**C. DUAL-VIEW SUPERVISED LEARNING STRATEGY**

Since the RefSR model has a more complex network structure than the general SISR model, using only the supervised signal of the final image layer for optimal learning leads to difficult gradient propagation, which poses some difficulties for the training of the network. To address the above challenges, we propose a dual-view supervised learning strategy to obtain more accurate feature representations. Specifically, in the intermediate feature layer, we employ an additional triple supervision loss to minimize the features of HR images and the fused features of LR and Ref images as follows:

$$\mathcal{L}_{sup} = \frac{1}{3} \sum_{n=1}^{3} \left\| \Phi_n^{encoder} (I_{HR}) - F_{fusion} \right\|_1$$  \hspace{1cm} (11)$$

where $I_{HR}$ denotes the real image, $\Phi_n^{encoder} (\cdot)$ denotes the $n$-th layer of the encoder, $F_{fusion}$ denotes the fusion feature of LR features and transferred features.

In the final image layer, we use the following three losses for supervision:

**Reconstruction loss.** The first loss is the reconstruction loss, which is used to measure the difference in pixel space between the HR image and the SR image:

$$\mathcal{L}_{rec} = \|I_{HR} - I_{SR}\|_1$$  \hspace{1cm} (12)$$

where $I_{HR}$ and $I_{SR}$ denote the ground truth image and the network output.

**Perceptual loss.** The second loss is the perceptual loss, which has been shown to help improve the visual quality and has been applied in [11], [31]. Its core idea is to enhance the
similarity in feature space between the predicted and target images:

\[ \mathcal{L}_{pec} = \| \phi^{vgg}_i(I_{HR}) - \phi^{vgg}_i(I_{SR}) \|_2 \]  

where \( \phi^{vgg}_i \) is the feature map of the \( i \)-th layer of the VGG network, and Here we use the \textit{conv5}_4 layer of the network, which is the last convolutional layer in the VGG network.

**Adversarial loss.** The adversarial loss [32] \( \mathcal{L}_{adv} \) is effective in generating visually pleasing images with natural details. It’s core idea is to use the mutual game learning of generative model G and discriminative model D to produce better outputs:

\[ \mathcal{L}_D = -E_{I_{HR}} \log (D(I_{HR}, I_{SR})) \]
\[ = -E_{I_{SR}} \log (1 - D(I_{SR}, I_{HR})) \]
\[ \mathcal{L}_D = -E_{I_{HR}} \log (1 - D(I_{HR}, I_{SR})) \]
\[ = -E_{I_{SR}} \log (D(I_{SR}, I_{HR})) \]  

Finally, we combine supervised signals at both intermediate feature layer and image layer levels to learn more powerful feature representations:

\[ \mathcal{L} = \lambda_{sup} \mathcal{L}_{sup} + \lambda_{rec} \mathcal{L}_{rec} + \lambda_{per} \mathcal{L}_{pec} + \lambda_{adv} \mathcal{L}_{adv} \]  

The model training algorithm for image super-resolution based on dual-view supervised learning and multi-attention mechanism is shown in Algorithm 1.

**Algorithm 1** Image Super-Resolution Method Based on Dual-View Supervised Learning and Multi-Attention Mechanism

**Input:** \( I_{LR} - LR \) image, \( I_{Ref} - \) Ref image, \( I_{HR} - HR \) image, \( numsteps, batchsize = 9, epochs \), initial learning rate = \( 10^{-4} \)

**Output:** \( \tilde{X} - SR \) image

1. Initialize \( \theta_w \) - parameters of network
2. for \( t = 1 \) to epochs do
3. for \( n = 1 \) to numsteps do
4. Extract the features of \( I_{LR}, I_{HR}, I_{Ref}, I_{Ref} \), through the encoder network noted as \( F_{LR}, F_{HR}, F_{Ref}, F_{Ref} \)
5. Mine the relationship between \( F_{LR} \) and \( F_{Ref} \) through attention module and use this relationship to extract the feature \( F_{Ref} \) from \( F_{Ref} \)
6. Concat \( F_{Ref} \) and \( F_{LR} \) as the new input, generate the fusion feature \( F_{fusion} \) through the decoder network and calculate the gradient \( \nabla_g f(I_{HR}, I_{SR}) \)
7. Compensate \( F_{fusion} \) for \( I_{LR} \) to generate the SR image \( I_{SR} \), calculate the gradient \( \nabla_g f(I_{HR}, I_{SR}) \)
8. Update \( \theta_w \) by \( \nabla_g f(I_{HR}, I_{SR}) \) and \( \nabla_g f(I_{HR}, I_{SR}) \) with Adam algorithm
9. end for
10. end for
11. return Parameters of the network

**D. IMPLEMENTATION DETAILS**

The encoder network contains three building blocks, each consisting of one convolutional layer and four ResBlocks, to output texture features in three different scales. In the feature selection attention module, to reduce the consumption of both time and GPU memory, the relevance embedding is only applied to the smallest scale and further propagated to other scales. The feature adaptation attention module uses three networks with the same network structure and different parameters to perform adaptive scaling of the information at each scale in parallel. The decoder network contains multiple convolutional layers, deconvolutional layers, ResBlocks [22], and feature fusion modules [17], where the fusion is first applied to the smallest scale and propagated upward to other scales one by one.

**IV. EXPERIMENTS**

**A. DATASETS AND SETTINGS**

In this paper, we mainly use CUFED5 [15] as the training set in our experiments, which is the benchmark dataset proposed in SRTNN and widely used in subsequent studies of RefSR. CUFED5 consists of 11871 training image pairs, each of which contains an original HR image and a reference image with 160x160 resolution. To compare with other SR models, we choose a bicubic downsampling operation with a scaling factor of 4x on the original HR image to obtain the LR image following the previous works, and we augment the training images by randomly horizontally and vertically flipping followed by randomly rotating 90°, 180° and 270° to enhance the robustness of our model during training. To investigate the generalization capacity of our model, we conduct experiments on three widely used benchmarks: CUFED5 testing set, Urban100 [33] and Manga109 [34]. There are 126 test images in the CUFED5 test set, and each image corresponds to 4 reference images with different similarities. Urban100 contains 100 building images without references, and each image takes its LR image as a reference so that the network can explore the self-similarity of the input images, and for Manga109, which also lacks reference images, the HR images in the randomly sampled dataset are used as reference images because its dataset is all composed of simple lines and flat colored areas. All results of PSNR and SSIM are evaluated on the Y channel of YCbCr color space.

The discriminator uses the same structure as [17]. In the training process, we choose Adam [35] as the algorithm for training optimization, and the parameters \( \beta_1 \) and \( \beta_2 \) are set to 0.9 and 0.999, respectively, and set the initial learning rate to \( 10^{-4} \), and use the cosine annealing strategy to vary the learning rate, which decreases to \( 10^{-6} \) every 500 epoch. The weight coefficients \( \lambda_{sup}, \lambda_{rec}, \lambda_{per} \) and \( \lambda_{adv} \) are set to 0, 5, 1, 1, and 0.005, respectively. The DSMA proposed in this paper is implemented based on the Pytorch framework and trained on an NVIDIA 2080Ti GPU.
To evaluate the effectiveness of DSMA, we compare our model with other state-of-the-art SISR and RefSR methods. The SISR methods include SRCNN [8], MDSR [36], RDN [22], RCAN [23], SRGAN [11], ENet [37], ESRGAN [13], RSRGAN [24]. The RefSR methods include SRNTT [15], TTSR [16], MASA [17]. All the models are trained on the CUFED5 training set, and tested on the CUFED5 testing set of Urban100 [33] and Manga109 [34]. All experiments are performed with a scaling factor of 4× between LR and HR images.

**Quantitative evaluations.** For a fair comparison with other MSE minimization-based methods on PSNR and SSIM, we train only another version of DSMA by minimizing the reconstruction loss, denoted as DSMA-rec. Table 1 shows a quantitative comparison of PSNR and SSIM, where the best and suboptimal results are indicated in red and blue, respectively. As shown in Table 1, our model outperforms the state-of-the-art method on these three benchmark datasets.

| Method       | CUFED5  | Urban100 | Manga109 |
|--------------|---------|----------|----------|
| SRCNN [8]    | 25.33/0.745 | 24.41/0.738 | 27.12/0.850 |
| MDSR [36]    | 25.93/0.777 | 25.51/0.783 | 28.93/0.891 |
| RDN [22]     | 25.95/0.769 | 25.38/0.768 | 29.24/0.894 |
| RCAN [23]    | 26.06/0.769 | 25.42/0.768 | 29.38/0.895 |
| SRGAN [11]   | 24.40/0.702 | 24.07/0.729 | 25.12/0.802 |
| ENet [37]    | 24.24/0.695 | 23.63/0.711 | 25.25/0.802 |
| ESRGAN [13]  | 21.90/0.633 | 20.91/0.620 | 23.53/0.797 |
| RSRGAN [24]  | 22.31/0.635 | 21.47/0.624 | 25.04/0.803 |
| SRNTT [15]   | 25.61/0.764 | 25.09/0.774 | 27.54/0.862 |
| SRNTT-rec [15]| 26.24/0.784 | 25.00/0.783 | 28.95/0.885 |
| TTSR [16]    | 25.53/0.765 | 24.62/0.747 | 28.70/0.886 |
| TTSR-rec [16]| 27.09/0.804 | 25.87/0.784 | 30.09/0.907 |
| MASA [17]    | 24.92/0.729 | 23.78/0.712 | 27.23/0.845 |
| MASA-rec [17]| 27.54/0.814 | 26.09/0.786 | 30.18/0.908 |
| DSMA         | 25.61/0.758 | 24.55/0.733 | 28.25/0.859 |
| DSMA-rec     | 27.63/0.820 | 26.20/0.790 | 30.55/0.912 |

**V. ABLATION STUDY**

In this section, we perform several ablation experiments to analyze our proposed method. We validate the effectiveness of our proposed attention module and the dual-view supervised strategy, and also analyze the effect of different similarity levels of reference images.

**Attention module.** The role of the feature adaptation attention module (FAAM) is to weaken the effect of noisy information in the Ref features and enhance the effect of important information. As shown in Fig.6, we compare the training of the model after adding attention and our model achieves faster convergence and higher PSNR level after convergence.

The effect of the FAAM is further illustrated in Fig.7. As shown in the left side of Fig.7, the baseline pays more attention to texture details on channel 10 and less details on channel 11, while our DSMA shows more attention to channel 11 and less attention to channel 10. This shows that our model can adaptively scale the channels according to the importance of the channel information. The right side of the figure shows the final recovery effect of the image. The DSMA with channel attention has an advantage in PSNR/SSIM values, and our DSMA can recover more realistic leaves in terms of visual perception, while the baseline shows a more blurred effect.

We also compared FAAM with other attention methods (including RCAN [23] and SENet [27]), as shown in Table 2. FAAM improves PSNR and SSIM by 0.09 dB, 0.11 dB, and 0.37 dB on the three public datasets, respectively, compared to the baseline (without introducing channel attention), only introducing 0.22M more parameters. All three methods obtained performance improvements on the CUFED5 dataset compared to the baseline, which suggests that the introduction of channel attention may have enhanced the model fitting ability, while SENet performed poorly on the other two datasets, even less than baseline, and RCAN and FAAM achieved some improvement on all three datasets, which suggests that the addition of residual structure may have enhanced the generalization ability of the model.

**Supervision loss.** The role of supervision loss is to restrict
FIGURE 5. Visual comparison among different SR methods on CUFED5 testing set (top five examples), Urban100 (the last example).

FIGURE 6. Convergence analysis with (blue) and without (orange) channel attention on CUFED5.

TABLE 2. Ablation study on the mixed attention module.

|                | baseline | SENet    | RCAN     | FAAM(ours) |
|----------------|----------|----------|----------|------------|
| CUFED5         | 27.34/0.816 | 27.66/0.823 | 27.62/0.819 | 27.63/0.820 |
| Urban100       | 26.09/0.786 | 25.95/0.782 | 26.13/0.787 | 26.20/0.790 |
| Manga109       | 30.18/0.908 | 30.02/0.901 | 30.23/0.910 | 30.55/0.912 |

param 4.03M 4.03M 4.25M 4.25M

FIGURE 7. The left side of the figure shows the performance of our DSMA model with baseline in the middle feature, and the right side of the figure shows their performance in the final image layer.

The fusion features and HR features to be similar. As shown in Fig.8, using this loss can motivate the network to learn more accurate feature representations and result in pleasing visual results. In addition, this loss improves the quantitative metrics PSNR and SSIM of DSMA from 25.43/0.750 to 25.61/0.758.

Influence of different reference similarity. In order to investigate the effect of correlation between LR and Ref images on DSMA, we conducted experiments on the CUFED5 test set, which has four different correlation levels of reference images. L1 and L4 represent the reference images provided by the CUFED5 test set, L1 is the most relevant level while...
L4 is the least relevant level. LR means using the input image itself as the reference image. As shown in Table 4, DSMA achieves the best results regardless of the relevance level of the reference image, with L1 as the reference image having the best performance.

**User Study.** We conducted a user study to qualitatively demonstrate the superiority of our method. In total, 10 users were asked to compare the visual quality of our method with the state-of-the-art techniques on the CUFED5 dataset, including ESRGAN, RSRGAN, TTSSR, and MASA. We showed them two images, one of which was the result of our method, and asked the users to choose the one with better visual quality. As shown in Fig.9, more than 80% of the users think that our method provided better results than the existing methods.

### VI. CONCLUSION

In this paper, we propose an image super-resolution method based on dual-view supervised learning and multi-attention mechanism (DSMA), which uses the attention mechanism to search for HR features on the Ref image to compensate for the LR recovery, weaken the noisy information in the features that affect recovery, and enhance the important information that helps recovery. In addition, we also introduce a dual-view supervised approach to integrate the supervised signals at both feature and image layers to optimize the network to learn a more robust feature representation. In the future, we will further investigate new attention structures based on reducing GPU memory usage and explore the use of easier LR-HR matching to guide LR-Ref matching to achieve better performance.

### REFERENCES

[1] Michal Irani and Shmuel Peleg. Improving resolution by image registration. CVGIP: Graphical models and image processing, 53(3):231–239, 1991.

[2] Y Li, Bruno Sicou, and F Peyrin. A review of the deep learning methods for medical images super resolution problems. IRBM, 42(2):120–133, 2021.

[3] Heltin Genitha Cyril Amala Dhason, Indhumathi Muthaia, Shanmuga Priyaa Sakhthivel, and Sanjeevi Shanmugam. Super-resolution mapping of hyperspectral satellite images using hybrid genetic algorithm. IET Image Processing, 14(7):1281–1290, 2020.

[4] Pejman Rasti, Tonis Uiboupin, Sergio Escalera, and Gholamreza Anbarjafari. Convolutional neural network super resolution for face recognition in surveillance monitoring. In International conference on articulated motion and deformable objects, pages 175–184. Springer, 2016.

[5] Lei Zhang and Xiaolin Wu. An edge-guided image interpolation algorithm via directional filtering and data fusion. IEEE transactions on Image Processing, 15(8):2226–2238, 2006.

[6] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In Asian conference on computer vision, pages 111–126. Springer, 2014.

[7] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang. Coupled dictionary training for image super-resolution. IEEE transactions on image processing, 21(8):3467–3478, 2012.

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence, 38(2):295–307, 2015.

[9] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In European conference on computer vision, pages 391–407. Springer, 2016.

[10] Jiwon Kim, Jung Kwon Lee, and Kyong Mu Lee. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1646–1654, 2016.

[11] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4681–4690, 2017.

[12] Jiwon Kim, Jung Kwon Lee, and Kyong Mu Lee. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1637–1645, 2016.

[13] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European conference on computer vision (ECCV) workshops, pages 0–0, 2018.

[14] Haitian Zheng, Mengji Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In Proceedings of the European conference on computer vision (ECCV), pages 88–104, 2018.
[15] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7962–7971, 2019.

[16] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5791–5800, 2020.

[17] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6368–6377, 2021.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[19] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018.

[20] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219, 2017.

[21] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In Proceedings of the IEEE international conference on computer vision, pages 4799–4807, 2017.

[22] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2472–2481, 2018.

[23] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European conference on computer vision (ECCV), pages 286–301, 2018.

[24] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3096–3105, 2019.

[25] Gyumin Shim, Jinsun Park, and In So Kweon. Robust reference-based super-resolution with similarity-aware deformable convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8433–8443, 2020.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[27] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132–7141, 2018.

[28] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3146–3154, 2019.

[29] Zhengwei Wang, Qi She, and Alsjoa Smolic. Action-net: Mipathp excitation for action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13214–13223, 2021.

[30] Meng Liu, Zhengyang Wang, and Shuiwang Ji. Non-local graph neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[31] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In European conference on computer vision, pages 694–711. Springer, 2016.

[32] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

[33] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5197–5206, 2015.

[34] Yusuke Matsui, Koto Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications, 76(20):21811–21838, 2017.

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[36] Bee Lim, Sanghyun Son, Heewon Kim, Seungjin Nah, and Kyoun Mu Lee. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 136–144, 2017.

[37] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirschl. Enhancenet: Single image super-resolution through automated texture synthesis. In Proceedings of the IEEE international conference on computer vision, pages 4491–4500, 2017.

XIN LIU was born in 1997. He received the bachelor’s degree in engineering from Anhui University of Technology, in 2019. He is currently pursuing the M.S. degree in computer science and technology with the Nanjing University of Aeronautics and Astronautics. His research interests include deep learning, computer vision, and image processing.

JING LI was born in 1976. She received the B.S. and M.S. degrees from the School of Computer Science and Engineering, Changhun University of Technology, in 1998 and 2001, respectively, and the Ph.D. degree in computer science and technology from Nanjing University, China, in 2004. She is currently an Associate Professor with the Nanjing University of Aeronautics and Astronautics, China.

TINGTING DUAN was born in 1992. She received the M.S. degree in Electronic Information from Beijing University of Posts and Telecommunications, in 2017. She is mainly engaged in information technology, blockchain technology and technology project management. She has participated in the specific industrial internet projects of MIIT such as "Industrial Internet Public Service Platform for Blockchain Innovation Application", and major scientific research projects of State Grid Corporation such as "Key Technology Research on Electricity Blockchain Smart Contracts for Grid Application Scenarios" and "Research on Business Application Transformation on Cloud and Full Link Operation Analysis Technology", and has won one second prize of Electricity Construction Science and Technology Progress Award.

JIANGTAO LI was born in 1994. He received the M.S. degree in Computer Technology from Beijing Technology and Business University, in 2019. He is mainly engaged in blockchain technology research, project management and standards preparation. He has participated in the preparation of State Grid Corporation project standards such as "Electricity Blockchain: Part 5 Cross-Chain Implementation Guide" and other important scientific research projects of State Grid Corporation such as "Design of Electricity Market Institutional Mechanism and Key Technology Research under the New Quota System Policy" and "Key Technology Research Direction of Energy and Electricity Blockchain", and has received awards such as the First Prize of State Grid Corporation Staff Innovation and Gold Prize of State Grid Corporation Youth Innovation Competition.
YE WANG was born in 1984. He received the M.S. degree in Electrical Engineering and its Automation from the State Grid Institute of Electric Power Science, in 2009. He is mainly engaged in technical research, project management and standards preparation for information systems and network security, and has participated in the preparation of many corporate standards of the State Grid Corporation, such as “Network Security Risk Monitoring and Warning Platform Security Monitoring Data Specification”.

***