Meta-TR: Meta-Attention Spatial Compressive Imaging Network With Swin Transformer

Can Cui, Linhan Xu, Boyu Yang, and Jun Ke, *Member, IEEE*

***Abstract*—As a flourishing research topic in the field of remote sensing, spatial compressive imaging (SCI) can utilize prior knowledge to recover high-dimensional signals from low-resolution measurements through joint sampling and compression, thus contributing to the bandwidth reduction of information transmission. However, most of the existing SCI methods based on deep learning cannot effectively utilize prior information, and difficult to perform deep extraction of image features, so the reconstruction is not ideal in the case of low sampling ratio. To address the above difficulty, we propose an SCI network based on meta-attention (MA) and swin transformer, named Meta-TR. We adopt the swin transformer as the network backbone, through the wide application of self-attention mechanisms, to achieve deeper extraction of image features, thereby improving the reconstruction quality under low sampling ratios. In addition, we design an MA module, which adopts Squeeze-Excitation architecture to convert the metadata of SCI image degradation process to attention vectors. Then, the attention vectors are used in the channel modulation of network feature maps to guide the network training. Extensive experiments are performed on different benchmark remote sensing datasets and different sampling ratios to confirm the superiority of the proposed Meta-TR method.**

***Index Terms*—Deep learning, meta-attention (MA), remote sensing, spatial compressive imaging (SCI), swin transformer.**

I. Introduction

COMPRESSIVE sensing is an epoch-making technology in the field of signal transmission, which can recover the original signal at a lower sampling ratio than Nyquist sampling [1]. Spatial compressive imaging (SCI), as an application of compressed sensing (CS) theory in the field of image spatial compression, aims to reconstruct high-resolution (HR) images from low-resolution (LR) measurements by employing prior information [2]. With SCI algorithms, more signal information can be recovered using a low-cost hardware, which can reduce the requirement to a sensor and data transmission bandwidth. Therefore, the idea of SCI has been favored by IR imaging [3], MRI [4], radar imaging [5], and other application fields [6]–[8].

As the emergence of extensive remote sensing tasks, such as resource exploration, climate monitoring, and environmental protection in recent years, the availability of remote sensing data has also increased. However, the explosive growth of HR remote sensing data has also brought great pressure on data compression and reconstruction. Based on this, some super-resolution (SR) methods are applied in the field of remote sensing, benefit from the mapping from low-dimensional space to high-dimensional. Molini et al. [9] proposed that DeepSUM uses a self-registration method to achieve LR to HR reconstruction. Salvetti et al. [10] designed a lightweight SR method with 3-D convolution and attention mechanism. Hang et al. [11] designed an SR method using the internal correlation and projection properties of hyperspectral images. Compared with SR, SCI has some advantages in the field of image compression and reconstruction, mainly due to the application of sensing matrix in the reconstruction process, which can achieve compression and reconstruction of sparse signals at a sampling ratio far lower than the Nyquist frequency. Therefore, the SCI algorithm can effectively relieve the data transmission pressure of remote sensing systems and contribute to the development of HR earth observation applications [12], [13]. Mallat and Zhang [14] first proposed the usage of a redundant dictionary to represent sparse signals and perform reconstruction. The orthogonal matrix pursuit, by solving the sparse approximation problem on redundant dictionaries, can be used to reconstruct an object in a faster speed [15]. Besides, these scholars [16], [17] use nonconvex sparse regularization methods to calculate the global optimal solutions. In the work of [18], the rank residual minimization algorithm is used to get the original signal, by using the nonlocal self-similarity prior and the low-rank characteristics of an signal. Although high-quality reconstructions can be obtained, a main drawback of these methods is the long running time due its iterative calculations. In addition, the reconstruction quality degrades rapidly as the sampling ratio decreases, which also limits their application.

To address above issues, scholars have used deep learning methods for vision tasks [19], [20]. In [21], convolutional neural network (CNN) is used for SCI, and the reconstructions are applied for target tracking to prove that sufficient semantic information is maintained after the compression and reconstruction. Some networks [22]–[25] are specifically designed for hardware implementation friendly and low-storage requirements by jointly optimizing compression and reconstruction during training. In deep residual reconstruction network (DR2-Net) [26], the time complexity of network is greatly reduced by using multiple residual blocks, while the reconstruction quality is improved.

Manuscript received 29 May 2022; revised 10 July 2022; accepted 24 July 2022. Date of publication 29 July 2022; date of current version 10 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61675023 and in part by the Fundamental Research Funds for the Central Universities under Grant 2022CX02002. *(Corresponding author: Jun Ke.)*

The authors are with the School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China (e-mail: cc7754321cc@dingtalk.com; xlh.lh@qq.com; yangboyu1211@163.com; jke@bit.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3194949

Although the neural networks discussed above have better reconstruction quality than traditional algorithms, they are hard to interpret and rely too much on dataset while ignoring imaging process. Thus, some works, such as iterative shrinkage-thresholding algorithm network (ISTA-Net) [27], combine traditional methods with neural networks by replacing the linear or nonlinear steps in each iteration of a traditional method with designed CNN units. Cui and Sun et al. respectively use the nonlocal self-similarity prior in the measurement domain and the multi-scale feature domain to find similar vectors in the size-limited vector space, fill each other with the missing information, and reconstruct the original image [28], [29]. In the article [30], the rank residual minimization algorithm is combined with deep network units to obtain highly competitive reconstruction results.

However, the above networks still have some unique problems. First, for SCI in remote sensing field, most networks for reconstruction use basic CNN units and residual connections, which limits the deep extraction of image global features. And at low sampling ratios, the traditional CNN network cannot achieve satisfactory results in some visual tasks due to its limited representation ability. According to this, exploring a network backbone with stronger and deeper extraction capabilities is the key to the progress of SCI networks. Second, these previous networks lack the effective utilization of prior information (such as sensing matrix in SCI), leading to training process being too dependent on the dataset, resulting in problems such as overfitting and poor transferability. Therefore, how to adopt the metadata of image degradation process is also vital to the reconstruction of SCI.

To deal with the above issues, we study an end-to-end SCI network based on meta-attention (MA) and swin transformer, named Meta-TR. Compared with previous SCI networks, the proposed Meta-TR can calculate the internal autocorrelation of the input measurement frames through the self-attention mechanism [31]. In this case, the network is able to mine deeper image information for a better reconstruction, which has shown clear superiority at low sampling ratios. In addition, we design a MA module, which uses the Squeeze-Excitation network (SeNet)[32] to convert the image degradation metadata (sensing matrix in SCI) into attention vector, which are used to modulate channels in each feature extraction module of the network. In this way, Meta-TR can make full use of image degradation metadata to guide network training, and the multi-level sharing way also makes the weights of each level maintain consistent convergence. The main contributions of this study are summarized as follows.

1) We adopt the swin transformer as the network backbone to extract higher-level information from LR measurement, by calculating the self-attention results of shift windows.
2) We design a novel MA module to guide the training of network, which employs dual-path pooling and SeNet to convert metadata into attention vectors.
3) The proposed Meta-TR performs better than the representative SCI methods on benchmark datasets with different bands and sampling ratios, which also shows an efficient balance among reconstruction performance, parameter size, and running time.

II. Methodology

In this section, the proposed SCI method is elucidated. For better understanding of the proposed Meta-TR, a brief review on the SCI problem formulation is given first. Then, we will introduce the structure and principle of swin transformer. Finally, we will introduce the proposed Meta-TR network architecture in detail.

A. SCI Problem Formulation

Conventionally, SCI aims to reconstruct the original high-dimensional object $x \in \mathbb{R}^n$ by inputting $m(m << n)$ random measurements $y \in \mathbb{R}^m$ [33]. Mathematically, the imaging process can be described as follows:

$$y = \Phi x \tag{1}$$

where Φ represents the sensing matrix of size $(m \times n)$, which satisfies the restricted isometry property RIP criterion [38]. However, due to $m << n$, the number of unknowns in (1) is much more than the number of equations, so there are infinite solutions in (1). Therefore, the solution condition of the underdetermined problem requires the original object x to satisfy the property of being sparse in the transform domain. Specifically expressed as follows:

$$y = \Phi x = \Phi \Psi s = \Theta s \tag{2}$$

where Ψ represents the transformation matrix, which also satisfies the RIP criterion. The parameter s represents the representation of original object x in the transform domain, which is sparse [39]. The parameter Θ represents the multiplication of Φ and Ψ. Through this transformation, the solution of the (1) can be converted into a constrained optimization problem of the l_0 norm [40], [41], as follows:

$$\min_{s \in \mathbb{R}^n} \| s \|_0, \text{s.t.} \Theta s = y \tag{3}$$

where $\| s \|_0$ represents the zero norm of s. In this way, the complexity of the calculation is greatly reduced.

Due to its strong learning ability and operational efficiency [34], [35], the neural network can use the fitting of the network parameters on the dataset to achieve the solution process of (3), i.e., to solve s from y. Compared to the traditional SCI algorithm, network-based algorithms have lower complexity and higher accuracy [36], [37]. In this article, we adopt Meta-TR to perform SCI, as shown in Fig. 1. After training on dataset, Meta-TR can reconstruct HR objects using LR measurements in an end-to-end way.

B. Swin Transformer Architecture

In this subsection, we will introduce the swin transformer architecture, which is the backbone of Meta-TR network.

Transformer was originally used in natural language processing [42], and it has also shown its superiority in remote sensing image processing in recent years [43]–[45]. However, the original transformer needs to pay attention to all pixels of image in the calculation, which leads to a sharp increase in calculation and increases the restrictions on deployment and application. Based on this, the swin transformer uses the window multihead

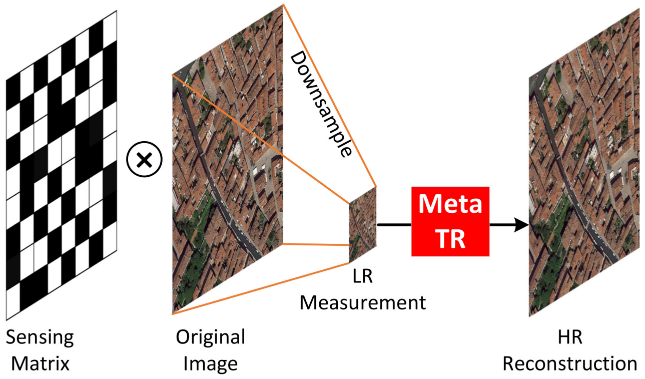


Fig. 1. Spatial compressive imaging reconstruction process using Meta-TR.

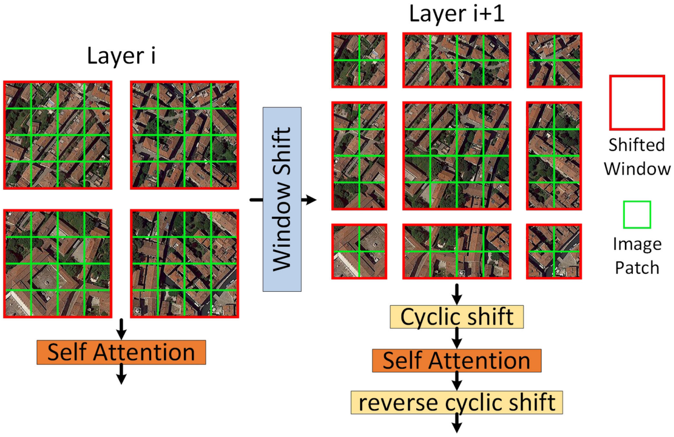


Fig. 2. Structure of W-MSA (left) and SW-MSA (right) and the window shift process.

self-attention (W-MSA) instead of the global multihead self-attention (MSA), which greatly reduces the amount of computation. In addition, in order to ensure the correlation information between windows, swin tranformer extends W-MSA to the shifting window multihead self-attention SW-MSA calculation [46]. As shown in Fig. 2, the left side represents the W-MSA, while the right side represents the SW-MSA. In SW-MSA, additional cyclic shift operations and inverse operations are used to ensure that, the window during self-attention calculation is consistent with that in W-MSA.

Now let us talk about the working mechanism of the swin transformer. For an input of size $(H \times W \times C)$, we divide it into $\frac{HW}{M^2}$ nonoverlapping windows $\mathbf{X}$ of size $(M^2 \times C)$. $\mathbf{X}$ is first processed using a layer normalization(LN), then the self-attention calculation within a window $\mathbf{X}$ is performed, specifically as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{SoftMax}(\mathbf{Q}\mathbf{K}^T/\sqrt{d} + B)\mathbf{V} \tag{4}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{M^2 \times d}$ are the $query$, key, and $value$ matrices of the preprocessed $LN(\mathbf{X})$, respectively. The parameter d represents the dimension of the key. The parameter B represents the relative position encoding. After that, the residual structure and multilayer perceptron (MLP) are applied to the self-attention result, specifically as follows:

$$\begin{aligned} \mathbf{X} &= F_{\text{wMSA}}(LN(\mathbf{X})) + \mathbf{X} \\ \mathbf{X} &= F_{\text{MLP}}(LN(\mathbf{X})) + \mathbf{X} \\ \mathbf{X} &= F_{\text{swMSA}}(LN(\mathbf{X})) + \mathbf{X} \\ \mathbf{X} &= F_{\text{MLP}}(LN(\mathbf{X})) + \mathbf{X} \end{aligned} \tag{5}$$

where F_{wMSA}, F_{MLP}, and F_{swMSA} represent W-MSA, MLP, and SW-MSA, respectively. Note that, the self-attention operation on the input of W-MSA and SW-MSA is the same, but the selection and shifting of the window are different. Details of (5) can also be found in Fig. 3(a). A gaussian error linear units (GELU) activation function is used in front of MLP. As shown in Fig. 3(a), each W-MSA is followed by an SW-MSA, and the two appear in pairs. The shifting distance of the window is $(M/2, M/2)$.

C. Meta-TR Network

In this subsection, we will first introduce the functional parts (shallow information extraction, deep information extraction, and MA) of Meta-TR, and then introduce the construction process of the loss function.

1) *Shallow Information Extraction:* In this part, we use convolutional layers to perform shallow extraction on the input image and retain most of the original information for subsequent deeper processing. After shallow information extraction, we use the layer normalization operation to prevent the gradient from disappearing and improve the network convergence speed. The formula of this part is expressed as follows:

$$I_{SIE} = F_{SIE}(I_{\text{LR}}) \tag{6}$$

where I_{LR} represents the input of network; F_{SIE} represents the shallow information extraction module and I_{SIE} represents the output of F_{SIE}.

2) *Deep Information Extraction:* After shallow information extraction, Meta-TR employs N_1 residual swin transformer block (RSTB) for deep information extraction. In RSTB, as shown in Fig. 3, short residual connections are used to aggregate features from different levels. Each RSTB contains N_2 swin transformer layer (STL), and the structure of STL is described in (5). Note that, W-MSA is used for odd-numbered STL, and SW-MSA is used for even-numbered STL. The two kinds of attention mechanism calculation methods appear alternately, in order to use the shift window to reduce the computational complexity of the network, which is also the core of the swin transformer. The formula of deep information extraction part is as follows:

$$F_{\text{RSTB}} = F_C(F_{\text{STL}_1} \ldots (F_{\text{STL}_{N_2}})) + 1 \tag{7}$$

$$F_{DIE} = F_C(F_{\text{RSTB}_1} \ldots (F_{\text{RSTB}_{N_1}})) + 1 \tag{8}$$

$$I_{DIE} = F_{DIE}(I_{SIE}) \tag{9}$$

where F_C, F_{STL}, F_{RSTB}, and F_{DIE} represent convolutional layer, STL, RSTB, and deep information extraction, respectively. I_{DIE} represent the output of F_{DIE}.

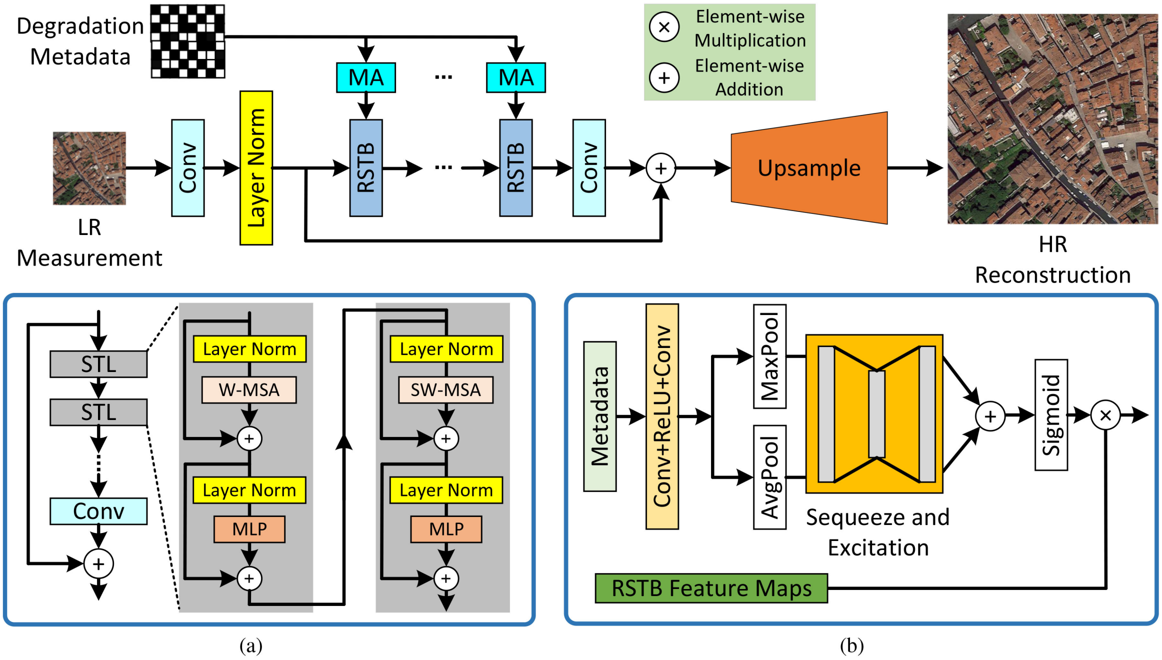


Fig. 3. Architecture of Meta-TR. The top is the overall structure of Meta-TR. (a) RSTB module in Meta-TR. (b) MA module in Meta-TR. (a) Residual Swin Transformer Block (RSTB). (b) Meta Attention (MA).

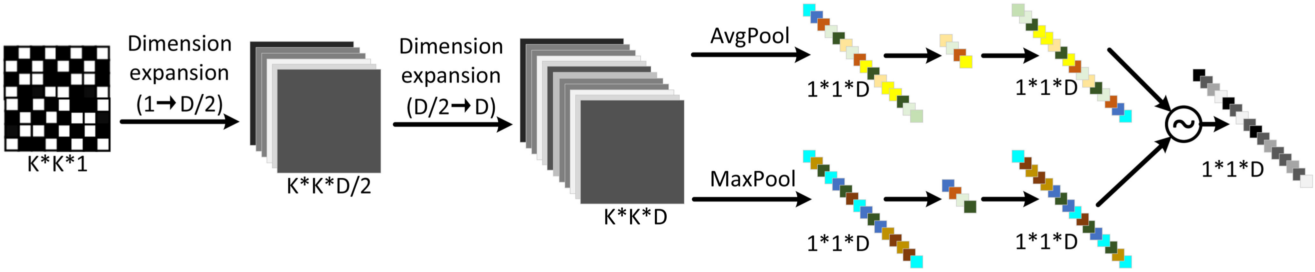


Fig. 4. Transformation process of the sensing matrix in the MA module. The subscript represents the size of the tensor.

3) *Metaattention:* In this subsection, we introduce the framework of MA. MA aims to utilize the metadata of image degradation to guide the overall training of the network. In the design of MA, we mainly adopt maxpooling, avgpooling, and SeNet structure. The two pooling structures are to extract the maximum and mean information of the sensing matrix, and reduce the 3-D tensor to a 1-D vector. After the pooling, each element has a global receptive field, and global features are obtained. Then, the SeNet is adopted to use the global features to obtain the nonlinear relationship between channels, and finally obtains a series of modulation factors between (0, 1) to guide the training of each RSTB module. In SCI, the most critical factor of image degradation is the sensing matrix. The following is a detailed description of the transformation process of the sensing matrix in MA.

As shown in the Fig. 4, for a sensing matrix of size $(K \times K \times 1)$, pass through the dimension expansion module consisting of convolutional layers, and the output tensor size is $(K \times K \times D)$, where D is equal to the number of channels in each RSTB. After that, MA utilizes average pooling and max pooling for core information compression, and outputs two vectors of size $(1 \times 1 \times D)$. And then, the SeNet is used to modulate the vector to achieve the extraction of core features with a small amount of parameters. Finally, the modulated vectors are added and activated using the sigmoid function, resulting in a final attention vector of size $(1 \times 1 \times D)$, named meta attention output (MA-OUT). At this point, MA-OUT represents the core information of the sensing matrix, and then we use it to channel-modulate the output of each RSTB. In this way, the network reconstruction quality can be improved by making the network pay more attention to feature maps with more important information.

4) *Loss Function:* Finally, after the upsample module, Meta-TR outputs a reconstruction with the same size of the original object. In this article, we utilize the maximum *a posteriori* (MAP) to construct the loss function, as follows:

$$\widehat{x} = \arg\min_{x} \frac{1}{2\sigma^2}||\Phi x - y||^2 + R(x) \tag{10}$$

TABLE I
QUANTITATIVE COMPARISON (AVERAGE PSNR/SSIM) WITH STATE-OF-THE-ART SCI METHODS FOR RED AND NIR IMAGES ON PROBA-V DATASET

Sampling ratio	Datasets	TVAL-3		ReconNet+res		MSRResNet		RAMS		Joinput-CiNet		Meta-CiNet		RCAN		**Meta-TR**	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
1/4	NIR	33.56	0.7868	20.01	0.3831	38.27	0.9443	38.17	0.9331	38.01	0.9421	38.09	0.9445	39.01	0.9487	**39.21**	**0.9516**
	RED	37.24	0.8224	22.31	0.4123	41.46	0.9771	41.01	0.9551	40.96	0.9668	41.05	0.9723	41.97	0.9764	**42.11**	**0.9814**
	ALL	35.62	0.8017	21.25	0.3916	40.12	0.9604	39.82	0.9432	39.38	0.9513	39.47	0.9521	40.35	0.9625	**40.80**	**0.9661**
1/16	NIR	27.65	0.6779	19.61	0.3611	33.15	0.8310	33.84	0.8265	33.06	0.8291	33.08	0.8298	33.79	0.8388	**34.01**	**0.8416**
	RED	31.25	0.7116	21.21	0.3822	35.98	0.8601	35.77	0.8556	35.81	0.8552	35.85	0.8561	36.12	0.9691	**36.45**	**0.8716**
	ALL	29.63	0.6998	20.60	0.3718	34.78	0.8452	34.85	0.8437	34.50	0.8402	34.52	0.8414	35.09	0.8487	**35.21**	**0.8556**
1/36	NIR	26.61	0.5842	19.56	0.3765	31.56	0.7694	31.66	0.7611	31.08	0.7648	31.21	0.7654	31.79	0.7694	**31.95**	**0.7721**
	RED	30.35	0.6442	20.15	0.3895	34.97	0.8023	33.99	0.8001	34.85	0.8012	34.91	0.8016	34.99	0.8091	**35.22**	**0.8114**
	ALL	28.16	0.6119	19.96	0.3824	33.01	0.7899	32.81	0.7812	32.86	0.7879	32.84	0.7882	33.13	0.7991	**33.26**	**0.7991**

Best performance are in bold.

TABLE II
QUANTITATIVE COMPARISON (AVERAGE PSNR/SSIM) WITH STATE-OF-THE-ART SCI METHODS FOR RGB IMAGES ON SATELLITE DATASET I (GLOBAL CITIES)

Sampling ratio	TVAL-3		ReconNet+res		MSRResNet		Joinput-CiNet		Meta-CiNet		**Meta-TR**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
1/4	30.12	0.8421	31.79	0.8716	32.45	0.8913	32.33	0.8907	32.29	0.8814	**32.61**	**0.8956**
1/16	24.25	0.6358	25.91	0.6796	26.46	0.7111	26.41	0.7025	26.39	0.7015	**26.62**	**0.7239**

Best performance are in bold.

where $\widehat{x}$ represents the output of Meta-TR, σ is the noise level, $R(x)$ is a regularization term. We can rewrite (10) as a function of parameters y, Φ, σ, and Θ, where Θ represents the parameters of MAP inference, specifically as follows:

$$\widehat{x} = M(y, \Phi, \sigma, \Theta). \tag{11}$$

Based on this, we design the loss function for Meta-TR as follows:

$$L(\Theta) = \frac{1}{2i_n} \sum_{i=1}^{i_n} ||M(y_i, \Phi, \sigma, \Theta) - x_i||^2 \tag{12}$$

where i_n represents the batch size of a sample in training.

III. EXPERIMENTS

In this section, we compare the proposed Meta-TR with the state-of-the-art SCI methods on remote sensing datasets with multiple bands and sampling ratios. First, the datasets and training details are introduced. Then, the comparison between our method and other SCI methods on visual effects and evaluation metrics is presented. After that, ablation experiments of the MA module and internal structure are performed to confirm its effectiveness. Finally, the parameter size and running time of the network is discussed.

A. Datasets and Training Details

Datasets: In this article, we train and test on two datasets detailed below: 1) Project for on-board autonomy vegetation (PROBA-V) [47]; 2) Satellite dataset I (global cities) [48].

PROBA-V is an earth observation satellite used to map global land and vegetation cover. This dataset has been released by the Advanced Concepts team of the European Space Agency. The PROBA-V dataset includes LR images of size (128 × 128) and HR images of (384 × 384). All images in the dataset are 14-b depth and single-channel. Additionally, this dataset contains 1160 scenes, 566 from the near infrared (NIR) band and 594 from the visible red (RED) band.

Satellite dataset I (global cities) is a subset of the wuhan university (WHU) building dataset, which is collected from remote sensing resources around the world and is mainly constructed with urban building clusters. This dataset includes 204 red green blue (RGB) images of size (512×512). In addition to satellite sensor differences, factors such as atmospheric conditions and seasonal changes make this dataset more informative and suitable for neural network training.

Training Details: In the experiment, Meta-TR is trained on the PROBA-V and Satellite datasets. There are 1160 images in PROBA-V dataset, 1000 as training set, and 160 as test set. There are 204 images in Satellite dataset, 153 as training set, and 51 as test set. As described in Section II-A, this article proposes an SCI method, and the LR input in the SCI process is obtained by HR through sensing matrix modulation and downsampling. The core component (sensing matrix) in the compression process of SCI can be regarded as the metadata of image degradation. Therefore, in the PROBA-V dataset and Satellite dataset I (global cities), we only need the HR dataset, and the LR dataset to be manually generated, by using the sensing matrix. During dataset preparation, for sampling ratios of 1/4, 1/16, and 1/36, we use sensing matrices of size (2 × 2), (4 × 4), and (6 × 6), respectively, to generate LR datasets by sliding and dot producting on HR datasets.

The LR patch size is set to (24 × 24), and the batch size is set to 16. It can be deduced that when the sampling rates are 1/4, 1/16, and 1/36, the corresponding HR image sizes are (48 × 48), (96 × 96), and (144 × 144). Data augmentation is performed

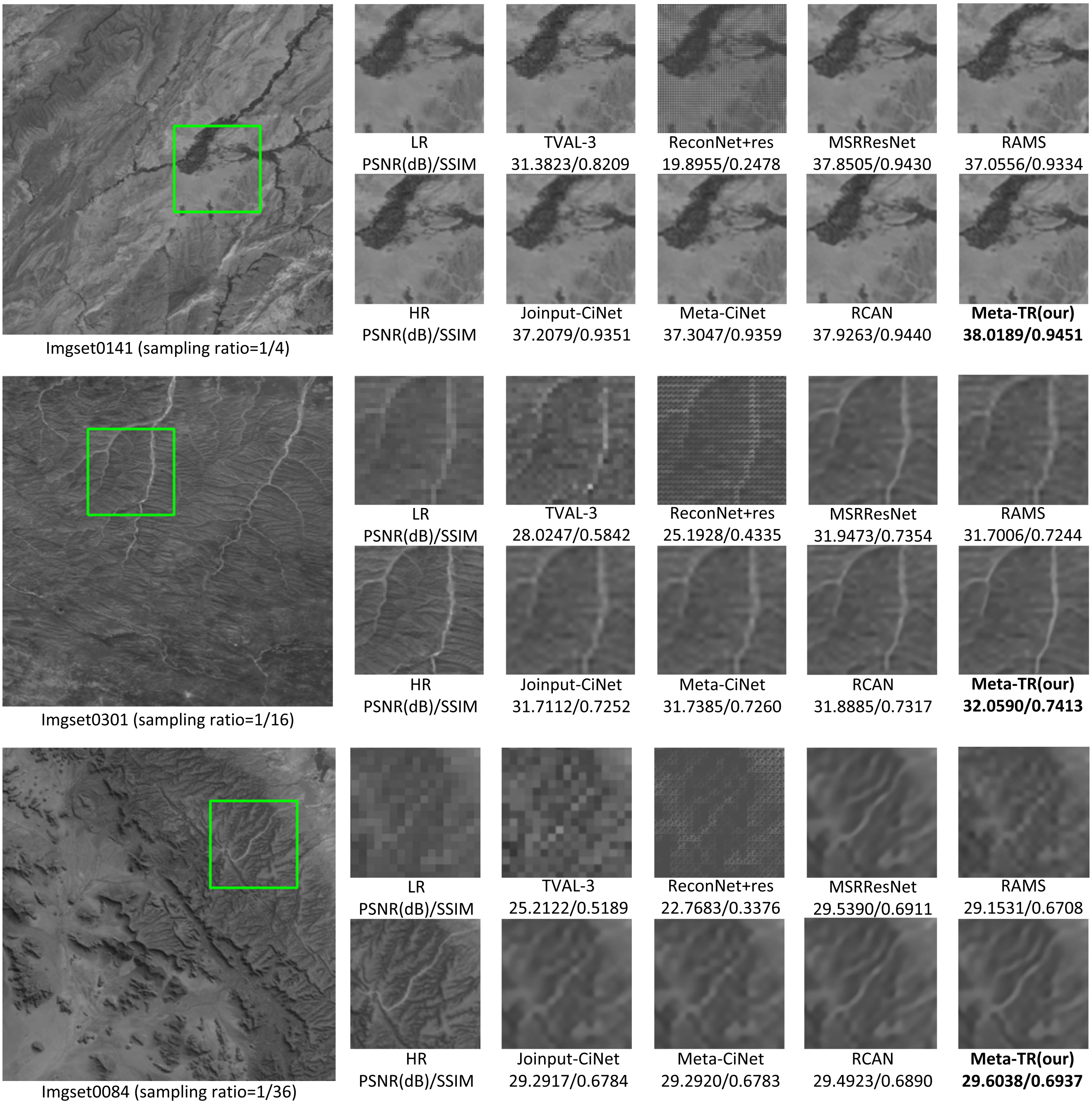


Fig. 5. Comparison of visual effects of different SCI methods on the PROBA-V dataset (RED band). Sampling ratio is set to 1/4, 1/16, and 1/36, respectively.

with rotation and cropping during training. The evaluation indicators of network reconstruction performance are peak signal to noise ratio (PSNR) and structural similarity (SSIM)[49]. The Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ is adopted to train the Meta-TR [50]. The initial learning rate is set to 1e$-$4. We train Meta-TR on an Nvidia GTX 3090 GPU for approximately two days to achieve the optimal results.

B. Comparing SCI Methods

In this subsection, we compare Meta-TR with representative SCI methods in recent years, including total variation augmented lagrangian alternating Direction algorithm (TVAL-3), ReconNet+res, modified super resolution residual network (MSRResNet), residual attention multi-image super-resolution (RAMS), Joinput-CiNet, Meta-CiNet, and residual channel attention network (RCAN). For a fair comparison, all methods use the same sensing matrix and dataset, and the networks are trained to convergence. These methods are described in detail below.

TVAL-3 [51]: This is a classic traditional algorithm, which adopts an augmented Lagrangian based total variational regularization model to achieve iterative SCI reconstruction.

ReconNet+res [21]: ReconNet is a block-to-block SCI network. For remote sensing datasets, a residual structure is

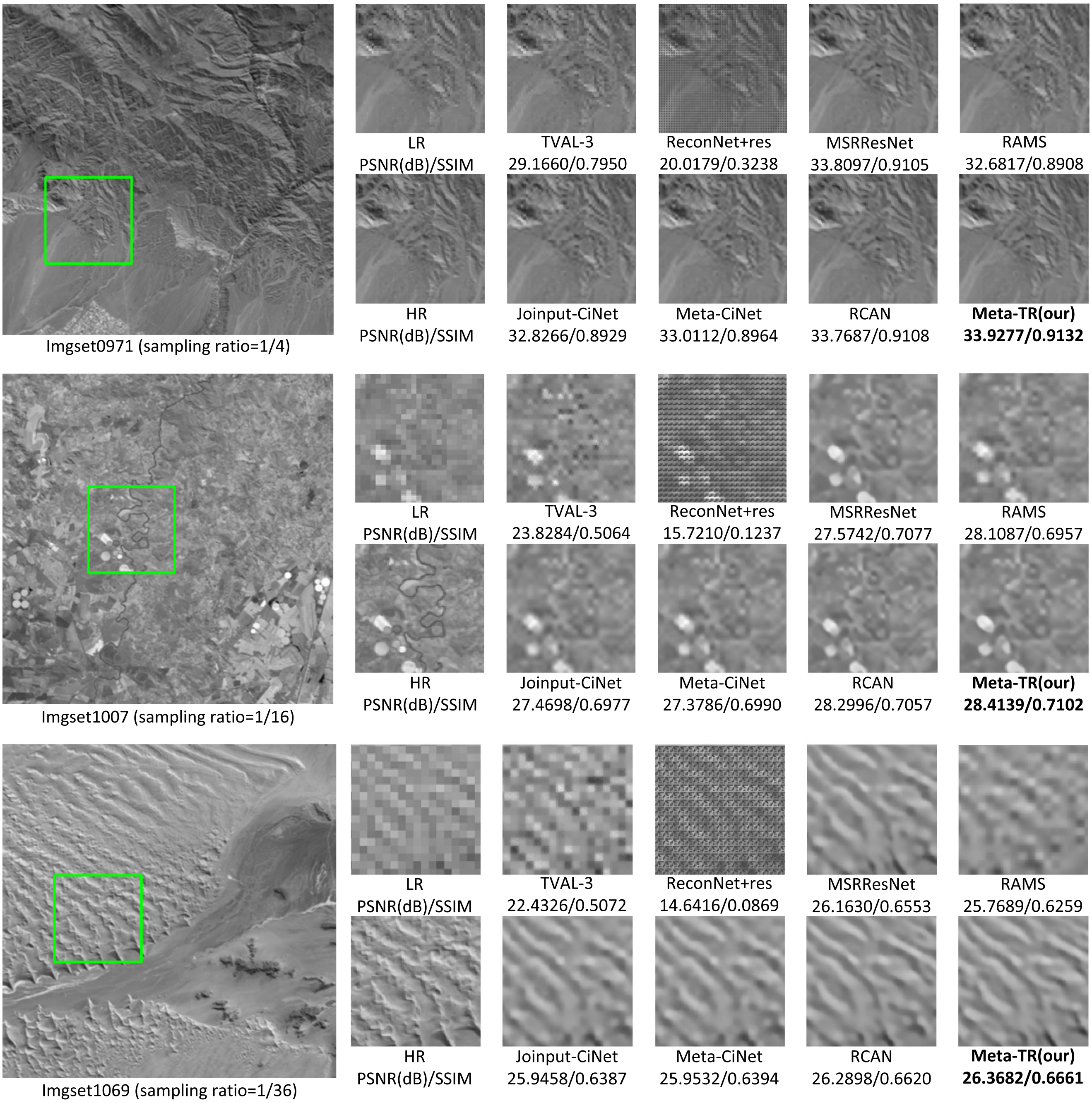


Fig. 6. Comparison of visual effects of different SCI methods on the PROBA-V dataset (NIR band). Sampling ratio is set to 1/4, 1/16, and 1/36, respectively.

added to enhance the reconstruction performance in this article.

MSRResNet [52]: The original MSRResNet is a modified version of the super-resolution reconstruction residual network. This article trains it to perform SCI reconstruction.

RAMS [10]: This is a representative lightweight network for remote sensing images reconstruction, which builds feature and temporal attention mechanism modules through 3-D convolution, and achieves excellent results on the PROBA-V dataset.

Joinput-CiNet [2]: It is a SCI network with joint input of degradation maps and LR measurements, which uses principal component analysis to extract sensing matrix information to guide reconstruction.

MetaCiNet [54], [55]: It is an improved version of Joinput-CiNet, which extracts more dimensional information of the sensing matrix than the former.

RCAN [53]:This is one of the most representative CNN SR networks, which uses residual-in-residual and channel attention mechanism to build a very deep network to achieve high-quality reconstruction.

C. Results on PROBA-V Dataset

In this subsection, we train and test all methods on the PROBA-V dataset. Experiments are carried out at sampling ratios 1/4, 1/16, and 1/36. In Table I, we summarize reconstruction PSNR and SSIM values using TVAL-3, ReconNet+res,

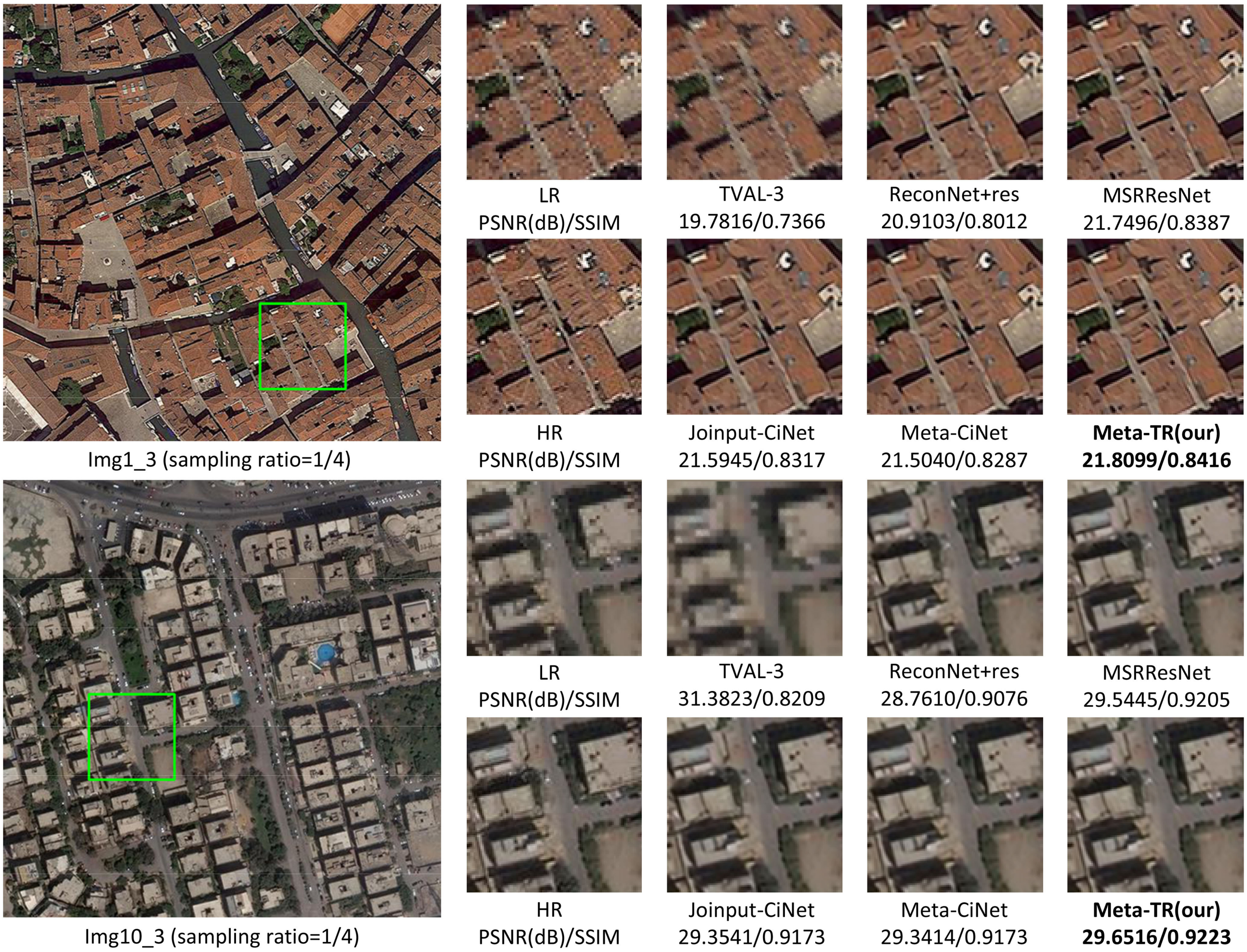


Fig. 7. Comparison of visual effects of different SCI methods on the Satellite dataset I (global cities). Sampling ratio is set to 1/4.

MSRResNet, RAMS, Joinput-CiNet, Meta-CiNet, RCAN, and Meta-TR. In the table, NIR, RED, and ALL represent the reconstruction results of each method in the infrared band, visible light band, and all bands, respectively. It can be seen that, the images in the RED band show better results than the NIR images at each sampling ratio, because the RED images have lower average brightness compared to the NIR images. In conclusion, our Meta-TR achieves the best PSNR/SSIM values on all sampling ratios and datasets. At sampling ratios of 1/4, 1/16, and 1/36, Meta-TR can achieve average improvements of 0.68 dB/0.0057, 0.43 dB/0.0104, and 0.25 dB/0.0092 compared to the classic MSRResNet method. It is worth mentioning that, the parameter amount of Meta-TR is about 1/16 of that of RCAN, but the reconstruction quality still exceeds that of RCAN under different datasets and sampling ratios. Extensive quantitative data demonstrate the superiority of the proposed Meta-TR on SCI.

Figs. 5 and 6 show the reconstruction visual results of different SCI methods in the RED and NIR bands, respectively. In each band, we can find that, as the sampling ratio decreases, the reconstruction results of all methods also decrease. Compared with other methods, the reconstruction results of the proposed Meta-TR have more detailed information (rivers, mountains, etc.), which is beneficial to the subsequent identification and analysis of remote sensing images. In addition, our method shows superiority in both RED and NIR bands, confirming that it can work well in different wavelengths.

D. Results on Satellite Dataset I (Global Cities)

Similar to the above subsection, we train and test all methods on the Satellite dataset I (global cities). The dataset consists of RGB images, which are located in the visible light band, and mainly reflect the information of urban building groups. Table II shows the PSNR/SSIM values of the reconstruction results of different methods at sampling ratios 1/4 and 1/16. It can also be found that, Meta-TR achieves the best indicators under different sampling ratios. Compared to the second best method, Meta-TR can achieve 0.16 dB/0.0043, 0.16 dB/0.0128 improvements in PSNR/SSIM at 1/4, 1/16 sampling ratios, respectively. Figs. 7 and 8 show the reconstruction visual results of different methods at sampling ratios 1/4 and 1/16, respectively. From the figure, we can find that, compared with other methods, the reconstructions of Meta-TR have more detailed information of buildings and roads, which is beneficial to the follow-up target monitoring and terrain mapping of remote sensing data. Extensive

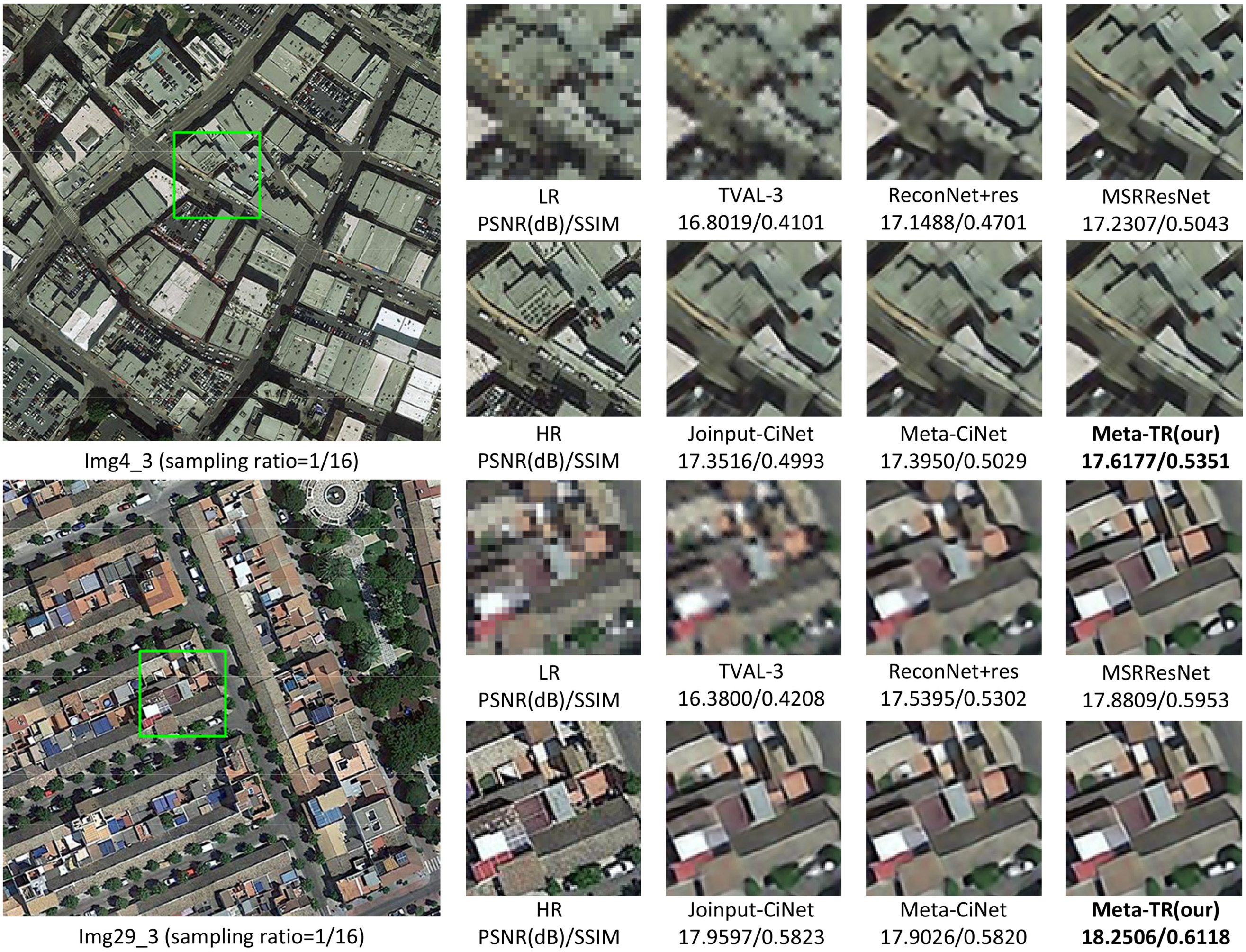


Fig. 8. Comparison of visual effects of different SCI methods on the Satellite dataset I (global cities). Sampling ratio is set to 1/16.

reconstruction indicators and visual results demonstrate the superiority and universality of Meta-TR.

E. Ablation Experiments of MA Module

In this subsection, the validity of the MA module and its internal structure are verified. As shown in Fig. 3(b), the core part of MA consists of Maxpool, Avgpool, and SeNet. Therefore, Meta-TR trains the following four versions on the PROBA-V dataset with a sampling ratio of 1/16:

1) Baseline (without MA);
2) Avgpool (MA that only contains Avgpool);
3) Maxpool (MA that only contains Maxpool);
4) Avgpool+Maxpool (with complete MA).

The training results are tested in the NIR and RED bands.

As shown in Table III, Meta-TR (Avgpool+Maxpool) has about 0.11 dB and 0.0034 improvement in PSNR and SSIM than Meta-TR (Baseline). This proves that MA module can boost the reconstruction indicator of Meta-TR. Furthermore, Meta-TR (Avgpool+Maxpool) surpasses the models using Avgpool or Maxpool alone, which confirms the rationality of the MA internal structure in this article.

TABLE III
Ablation Experiments of MA Structure

Algorithm	NIR		RED		ALL	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Baseline	33.12	0.8317	36.80	0.8661	35.10	0.8534
Avgpool	33.17	0.8334	36.85	0.8674	35.13	0.8545
Maxpool	33.18	0.8341	36.89	0.8681	35.15	0.8550
Avgpool+Maxpool	**33.27**	**0.8354**	**36.93**	**0.8701**	**35.21**	**0.8568**

The bold entities represents the method proposed in this paper.

Fig. 9 shows the visual reconstructions of the two network versions. It can be found that, Meta-TR (w/ MA) has more advantages in detail reconstruction, and has improvement in both bands. Through the ablation experiments in this section, it is confirmed that the MA module can make full use of the degradation information to guide the network training, and also prove its effectiveness for SCI.

F. Comparison of Parameters and Running Time

In this subsection, the parameter quantities and running time of different SCI networks are compared and discussed. Table IV

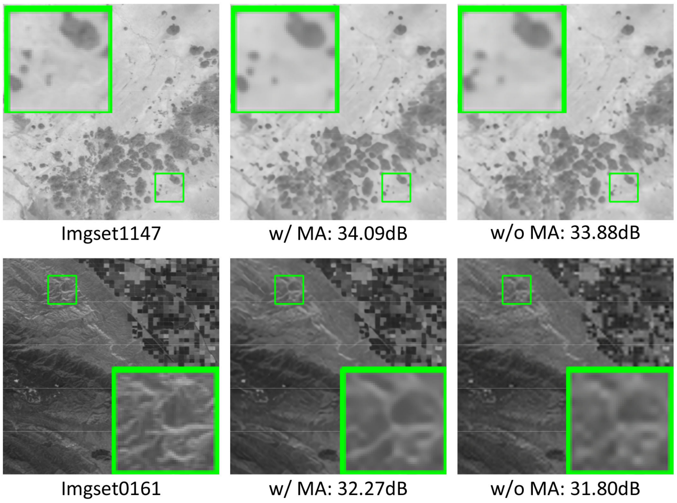


Fig. 9. Comparison of reconstruction results of Meta-TR with and without MA module at sampling ratio 1/16. The value in parentheses is PSNR (dB).

TABLE IV
COMPARISON OF PARAMETERS OF DIFFERENT SCI NETWORKS AT MULTIPLE SAMPLING RATIOS

Algorithm	1/4	1/16	1/36
	Parameters	Parameters	Parameters
ReconNet+res	229K	231K	233K
MSRResNet	1296K	1331K	1331K
RAMS	304K	321K	380K
Joinput-CiNet	1483K	1498K	2811K
Meta-CiNet	1491K	1504K	2819K
RCAN	15480K	15629K	15814K
Meta-TR	**913K**	**932K**	**965K**

The bold entities represents the method proposed in this paper.

shows the parameters of ReconNet+res, MSRResNet, RAMS, Joinput-CiNet, Meta-CiNet, RCAN, and Meta-TR, under sampling ratios 1/4, 1/16, and 1/36. It can be seen that, as the sampling ratios decreases, the parameters of all networks will increase, which is mainly due to the increase of layer numbers in the upsampling module. Furthermore, we can find that, Meta-TR achieves the third-least number of parameters among all methods, but achieves the best reconstruction results (according to Sections III-C and III-D). This shows that Meta-TR can achieve better reconstruction results with a lower number of parameters, which is more conducive to model deployment and application. This subsection reflects that Meta-TR achieves an excellent balance between network performance and parameter quantity.

Finally, the reconstruction running time comparison of different SCI methods is presented. As shown in Table V, except TVAL-3, the reconstruction time of other SCI methods for an image of size (384*384) is kept between 20 and 60 ms, which can basically meet the needs of real-time imaging. This subsection illustrates that Meta-TR still achieves a good balance between reconstruction time and quality.

TABLE V
COMPARISON OF THE RUNNING TIME OF VARIOUS SCI METHODS ON (384*384) IMAGES, WITH A SAMPLING RATIO OF 1/16

Algorithm	Running time	Algorithm	Running time
TVAl-3	4123ms	Joinput-CiNet	24.94ms
ReconNet+res	27.72ms	MetaCiNet	27.40ms
MSRResNet	41.67ms	RCAN	57.37ms
RAMS	22.26ms	**Meta-TR**	**43.92ms**

The bold entities represents the method proposed in this paper.

IV. CONCLUSION

In this article, we propose a SCI network employing MA and swin transformer. The proposed Meta-TR uses the swin transformer as the network backbone to extract global information inside the image block by using self-attention, which improves the depth of network information extraction while ensuring that the amount of parameters is not overloaded. Furthermore, we design a MA module to extract key information from the image degradation metadata in SCI through Squeeze-Excitation structure, and perform channel modulation in the feature maps of Meta-TR. By using this module, additional prior information can be used to guide the network training process, which improves the network reconstruction quality and interpretability. Extensive experiments on remote sensing benchmark datasets with different bands and different sampling ratios confirm the superiority of the proposed Meta-TR in both reconstruction metrics and visual effects.

REFERENCES

[1] D. Donoho, “Compressed sensing,” *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[2] C. Cui and J. Ke, “Spatial compressive imaging deep learning framework using joint input of multi-frame measurements and degraded maps,” *Opt. Exp.*, vol. 30, no. 2, 2022, Art. no. 1235.

[3] L. Zhang, J. Ke, S. Chi, X. Hao, T. Yang, and D. Cheng, “High-resolution fast mid-wave infrared compressive imaging,” *Opt. Lett.*, vol. 46, no. 10, 2021, Art. no. 2469.

[4] J. Sun, H. B. Li, and Z. B. Xu, “Deep ADMM-Net for compressive sensing MRI,” *Adv. Neural Inf. Process. Syst.*, vol. 29, pp. 10–18, 2016.

[5] M. Tello Alonso, P. Lopez-Dekker, and J. Mallorqui, “A novel strategy for radar imaging based on compressive sensing,” *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 12, pp. 4285–4295, Dec. 2010.

[6] M. Duarte et al., “Single-pixel imaging via compressive sampling,” *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 83–91, Mar. 2008.

[7] C. Deng, D. Jing, Y. Han, S. Wang, and H. Wang, “FAR-Net: Fast anchor refining for arbitrary-oriented object detection,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, Art. no. 6505805.

[8] L. Tang, W. Tang, X. Qu, Y. Han, W. Wang, and B. Zhao, “A scale-aware pyramid network for multi-scale object detection in SAR images,” *Remote Sens.*, vol. 14, no. 4, pp. 973–996, 2022.

[9] A. Bordone Molini, D. Valsesia, G. Fracastoro, and E. Magli, “DeepSUM: Deep neural network for super-resolution of unregistered multitemporal images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3644–3656, May 2020.

[10] F. Salvetti, V. Mazzia, A. Khaliq, and M. Chiaberge, “Multi-image super resolution of remotely sensed images using residual attention deep neural networks,” *Remote Sens.*, vol. 12, no. 14, 2020, Art. no. 2207.

[11] R. Hang, Q. Liu, and Z. Li, “Spectral super-resolution network guided by intrinsic properties of hyperspectral imagery,” *IEEE Trans. Image Process.*, vol. 30, pp. 7256–7265, 2021.

[12] L. Guo, R. Yang, Z. Zhong, R. Zhang, and B. Zhang, “Target recognition method of small UAV remote sensing image based on fuzzy clustering,” in *Proc. Neural Comput. Appl.*, 2021, pp. 1–17.

[13] H. Yang, J. Kong, H. Hu, Y. Du, M. Gao, and F. Chen, "A review of remote sensing for water quality retrieval: Progress and challenges," *Remote Sens.*, vol. 14, no. 8, 2022, Art. no. 1770.

[14] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[15] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[16] X. Huang and M. Yan, "Nonconvex penalties with analytical solutions for one-bit compressive sensing," *Signal Process.*, vol. 144, pp. 341–351, 2018.

[17] F. Hou, J. Chen, and G. Dong, "Compressed sensing with nonconvex sparse regularization and convex analysis for duct mode detection," *Mech. Syst. Signal Process.*, vol. 145, 2020, Art. no. 106930.

[18] Z. Zha, X. Yuan, B. Wen, J. Zhou, J. Zhang, and C. Zhu, "From rank estimation to rank approximation: Rank residual constraint for image restoration," *IEEE Trans. Image Process.*, vol. 29, pp. 3254–3269, 2020.

[19] Z. Zhao, Y. Han, T. Xu, X. Li, H. Song, and J. Luo, "A reliable and real-time tracking method with color distribution," *Sensors*, vol. 17, no. 10, 2017, Art. no. 2303.

[20] C. Deng, S. He, Y. Han, and B. Zhao, "Learning dynamic spatial-temporal regularization for UAV object tracking," *IEEE Signal Process. Lett.*, vol. 28, pp. 1230–1234, 2021.

[21] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, and A. Ashok, "Reconnet: Non-iterative reconstruction of images from compressively sensed measurements," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 449–458.

[22] W. Shi, F. Jiang, S. Liu, and D. Zhao, "Image compressed sensing using convolutional neural network," *IEEE Trans. Image Process.*, vol. 29, pp. 375–388, 2020.

[23] Z. Zhang, Y. Liu, J. Liu, F. Wen, and C. Zhu, "AMP-Net: Denoising-based deep unfolding for compressive image sensing," *IEEE Trans. Image Process.*, vol. 30, pp. 1487–1500, 2021.

[24] W. Shi, F. Jiang, S. P. Zhang, and D. B. Zhao, "Deep networks for compressed image sensing," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2017, pp. 877–882.

[25] W. Shi, F. Jiang, S. H. Liu, and D. B. Zhao, "Scalable convolutional neural network for image compressed sensing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12290–12299.

[26] H. Yao, F. Dai, S. Zhang, Y. Zhang, Q. Tian, and C. Xu, "DR2-Net: Deep residual reconstruction network for image compressive sensing," *Neurocomputing*, vol. 359, pp. 483–493, 2019.

[27] J. Zhang and B. Ghanem, "ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1828–1837.

[28] W. Cui, S. Liu, F. Jiang, and D. Zhao, "Image compressed sensing using non-local neural network," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2021.3132489.

[29] Y. Sun, Y. Yang, Q. Liu, J. Chen, X. Yuan, and G. Guo, "Learning non-locally regularized compressed sensing network with half-quadratic splitting," *IEEE Trans. Multimedia*, vol. 22, pp. 3236–3248, 2020.

[30] Z. Y. Zha, X. Yuan, J. T. Zhou, J. T. Zhou, B. H. Wen, and C. Zhu, "The power of triply complementary priors for image compressive sensing," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 983–987.

[31] T. An, X. Zhang, C. Huo, B. Xue, L. Wang, and C. Pan, "TR-MISR: Multiimage super-resolution based on feature fusion with transformers," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1373–1388, 2022.

[32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[33] Z. Chen et al., "Deep-learned regularization and proximal operator for image compressive sensing," *IEEE Trans. Image Process.*, vol. 30, pp. 7112–7126, 2021.

[34] Y. Han, C. Deng, Z. Zhang, J. Li, and B. Zhao, "Adaptive feature representation for visual tracking," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 1867–1870.

[35] B. Zhao, B. Zhao, L. Tang, Y. Han, and W. Wang, "Deep spatial-temporal joint feature representation for video object detection," *Sensors*, vol. 18, no. 3, pp. 774–794, 2018.

[36] C. Cui and J. Ke, "A multiple degradation maps network for spatial compressed sensin," *Comput. Opt. Sens. Imag.*, pp. C M2E–2, 2021.

[37] C. Cui and J. Ke, "Multi frame super resolution technology based on deep learning and compressed sensing," in *Proc. 4th Opt. Young Scientist Summit*, 2021, vol. 11781, pp. 77–81.

[38] E. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.

[39] L. Li, Y. Fang, L. Liu, H. Peng, J. Kurths, and Y. Yang, "Overview of compressed sensing: Sensing model, reconstruction algorithm, and its applications," *Appl. Sci.*, vol. 10, no. 17, 2020, Art. no. 5909.

[40] E. Candes and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[41] L. Feng, H. Sun, Q. Sun, and G. Xia, "Compressive sensing via nonlocal low-rank tensor regularization," *Neurocomputing*, vol. 216, pp. 45–60, 2016.

[42] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 1–11, 2017.

[43] S. Li, Q. Guo, and A. Li, "Pan-sharpening based on CNN pyramid transformer by using no-reference loss," *Remote Sens.*, vol. 14, no. 3, pp. 624–647, 2022.

[44] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022, Art. no. 5615611.

[45] P. Lv, W. Wu, Y. Zhong, F. Du, and L. Zhang, "SCViT: A spatial-channel feature preserving vision transformer for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022, Art. no. 4409512.

[46] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[47] W. Dierckx et al., "PROBA-V mission for global vegetation monitoring: Standard products and image quality," *Int. J. Remote Sens.*, vol. 35, no. 7, pp. 2589–2614, 2014.

[48] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[51] C. B. Li, W. T. Yin, and Y. Zhang, "User's guide for TVAL3: TV minimization by augmented lagrangian and alternating direction algorithms," *CAAM Rep.*, vol. 20, no. 46–47, pp. 4–20, 2009.

[52] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.

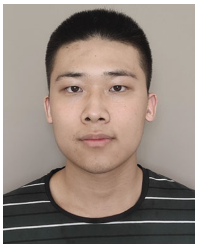
[53] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.

[54] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3262–3271.

[55] M. Aquilina, C. Galea, J. Abela, K. Camilleri, and R. Farrugia, "Improving super-resolution performance using meta-attention layers," *IEEE Signal Process. Lett.*, vol. 28, pp. 2082–2086, 2021.

Can Cui received the B.S. degree in the field of optoelectronic information science and engineering, in 2020, from the School of Optics and Photonics, Beijing Institute of Technology, Beijing, China, where he is currently working toward the M.S. degree in the field of optical engineering.

His current research interests include super resolution, remote sensing image processing, and compressive imaging.



Linhan Xu received the B.S. degree in the field of electronic science and technology from Xidian University, Xi'an, China, in 2021. He is currently working toward the M.S. degree in optical engineering with the Computational Imaging Laboratory, Beijing Institute of Technology, Beijing, China.

His current research interests include compressed sensing and infrared target recognition.

Boyu Yang received the B.S. degree in electronic science and technology from the Hefei University of Technology, Hefei, China, in 2021. He is currently working toward the M.S. degree in the field of optical engineering from the School of Optics and Photonics, Beijing Institute of Technology, Beijing, China.

His current research interests include event camera, speckle imaging, and video synthesis.

Jun Ke (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the University of Electronic and Science Technology of China, Chengdu, China, in 1996 and 1999, respectively, the M.S. degree in mathematics from Purdue University, in 2002, and the Ph.D. degree in optical and computer science from the Department of Electrical and Computer Engineering (ECE), University of Arizona (UA), in 2010.

She is currently the Associate Professor with the School of Optics and Photonics, Beijing Institute of Technology, Beijing, China. Her research interests include optical science and computational imaging.

Dr. Ke is a senior member of The Optical Society (OSA) and a member of Society of Photo-Optical Instrumentation Engineers (SPIE). She is the reviewer of *Journal of the Optical Society of America* (JOSA), *Applied Optics*, *Optics Letters*, *Optics Express*, etc.