IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 October 2022, accepted 23 October 2022, date of publication 3 November 2022, date of current version 23 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3218834

RESEARCH ARTICLE

Remote Sensing Colorization Based on Bidirectional Macro-Micro Adaptive Enhancement Network

JINGYU WANG, CHENGLONG WANG, QICHENG YANG, CHENGYU ZHENG, JIE NIE, (Member, IEEE), AND MINGXING JIANG

Faculty of Information Science and Engineering, Ocean University of China, Qingdao 266100, China

Corresponding author: Mingxing Jiang (jiangmx@ouc.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 202042008, in part by the National Natural Science Foundation of China under Grant 62172376 and Grant 62072418, in part by the Major Scientific and Technological Innovation Project of Shandong under Grant 2019JZZY020705, and in part by the Key Research and Development Program of Qingdao Science and Technology Plan under Grant 21-1-2-18-xx.

ABSTRACT The demand for re-colorization of remote sensing images is urgent since image quality is extremely deteriorated by haze or other noises occurring in the atmospheric layer. The most challenging issue is to restore the color information with respect to preserving spatial consistency as well as to obtain object salience in context with extremely imbalanced space structure, where the former requires learning stable macroscopic semantics while the latter needs to recover microscopic pixels. In this paper, we propose a Bidirectional Macro-Micro Adaptive Enhancement (BMMAEnet) framework by adopting three modules, i.e., the Downward Micro Enhancement (DME) module, the Upward Adaptive Macro Enhancement(UAME) module, and Macro-Micro Balance (MMB) module. Firstly, the DME module is designed by adding micro details as well as suppressing macro context during the multi-branch downsampling process to supplement missing pixel details. Secondly, UAME is proposed by adaptive selecting proper level of features during multi-branch upsampling process to strengthen macro semantic constraints. In addition, MMB is designed by embedding attention-guided local details and global semantics into the decoding features to balance micro and macro information within each branch. Comprehensive comparison and ablation experiments are implemented and verify the proposed method performs overpass SOTA methods not only in pixel color value restoration performance but also in human perceptive understanding.

INDEX TERMS Colorization, DCGAN, multi-scale, remote sensing image.

I. INTRODUCTION

Nowadays, remote sensing data plays arising role in the scientific cognition of the earth process. Other than multi-banded images, the natural pseudo color image is considered as a critical means for morphology understanding conformed to human perception. However, satellite images suffer extremely from either low contrast or chromaticity unsaturation problem due to the limitation of remote sensors and the air environment, such as haze or clouds. Thus, color restoration of remote sensing images is eager demanded to enhance human understandability and reveal unknown earth processes [1].

With the development of deep learning models, image colorization techniques have been widely researched and made great progress in approaching human perception. Deep generative models are recently well-researched methods with satisfied colorization performance on ordinary images, especially Generative Adversarial Networks (GANs) and their multiple variants [2], [3], [4], [5] which force the generator to produce real images through a game between the generator and the discriminator.

Isola et al. [6] proposed PatchGAN for image colorization, which can generate chroma-rich coloring maps.

The associate editor coordinating the review of this manuscript and approving it for publication was Jerry Chun-Wei Lin.

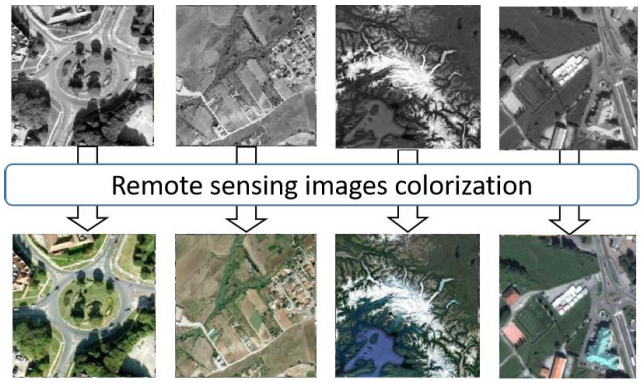


FIGURE 1. Examples of remote sensing image colorization. The first line shows grayscale images and the second line shows color images generated by the colorization method.

However, remote sensing images are quite different from ordinary images from the perspective of space layout. That is, the contextual structure of remote sensing image is consistent with the physical earth surface with extremely imbalanced objects distribution, for example, there are always micro-objects (cars or boats) surrounded by macro continuous texture regions. Under this condition, if we focus to learn the distribution of local pixels, it will deteriorate the ill-posed problem in macro texture regions, leading to generating abrupt pixels among consistency space. The single-scale PatchGAN ignores the spatial context and cannot guarantee the coloring spacial consistency of remote sensing images. Li. et al. [7] solved this problem by introducing multi-scale discriminators, it only optimizes the measurement of Jensen-Shannon divergence [8], but could not apply a strong constraint on macro scale-space stability. Wu et al. [9] proposed a method combining multi-scale convolution and SEnet [10] to increase contextual information while giving different weights to channels, which ensures spatial consistency to a certain extent. However, the features containing contextual information generated by convolution operations with different kernels sizes contain a number of redundant information and lose local details that are important for pixel coloring. On the other hand, SEnet focuses on the importance of channels rather than positions, which is not conducive to the microscopic details of coloring.

In addition, the encoder-decoder architecture used in the above methods loses local details information in the process of restoring to the original scale through upsampling. Although the representative structure U-Net [11] belonging to the encoder-decoder used as a generator in [6] and [7] recovers a part of details by adding skip connections, this naive connection way cannot promote the network to extract local detail features, making it impossible to generate color images with object salience.

To address the above problem, inspired by the idea of multi-scale, our network selects a DCGAN [4] including a multi-branch generator based on U-Net as the skeleton to learn macro-semantics and micro-pixels simultaneously which can preserve spatial consistency and obtain the object salience in coloring. Among them, the microscopic branch that takes the original grayscale image as input learns the pixel details, and the macroscopic that takes the downsampling grayscale image as the input learns the context information. Focus on making up for the details lost in the multi-branch downsampling process, the DME module is proposed to extract local details by suppressing the context information in the micro-branch and micro detail is enhanced by fusing the extracted missing local details with macro features. To enhance the macro-semantic constraints in the multi-branch upsampling process, the UAME module was proposed, which achieves macro-semantic enhancement by adaptively selecting appropriate context to constrained micro branches in the macro branches. In addition, for each branch, we propose the MMB module to fuse shallow features containing salient local details and embedding features containing global semantics with decoding features to balance micro-details and macro-semantics and obtain decoding features containing more useful information.

- We propose a Bidirectional Macro-Micro Adaptive Enhancement (BMMAE) network to guarantee both spatial consistency and object salience of remote sensing colorization by learning stable macro-semantics and recovering micro-pixels.
- We propose the UAME module to enhance macro-semantic constraints on the micro-scale by adaptively selecting contextual features to ensure color space consistency.
- We propose the MMB module, which balances local details and global semantics by using an attention-guided scheme during the generating process.
- We demonstrate the effectiveness of each module and the advancement of the whole network in qualitative and quantitative indicators through comparative and ablation experiments.

II. RELATED WORK

Image colorization can be divided into two categories [12], [13], [14]: user guided-based colorization and fully automatic image colorization. The user guided-based colorization includes scribbling-based [15], [16], [17], [18], [19] and exemplar-based [20], [21], [22], [23], [24], [25] methods that rely on human intervention. Compared with user-guided based colorization, the fully automatic colorization method does not need reference information and generates color images without human participation.

A. SCRIBBLE-BASED METHODS

The user provides local clues of color scribbles, which are then propagated to the corresponding area and finally realized the colorization of the whole image. Sangkloy et al. [15] proposed an adversarial method for sketch colorization which implements coloring based on the constraints of the user's scribbles and the boundaries of the sketch. Ren et al. [18] proposed a framework for semi-automatic learning based

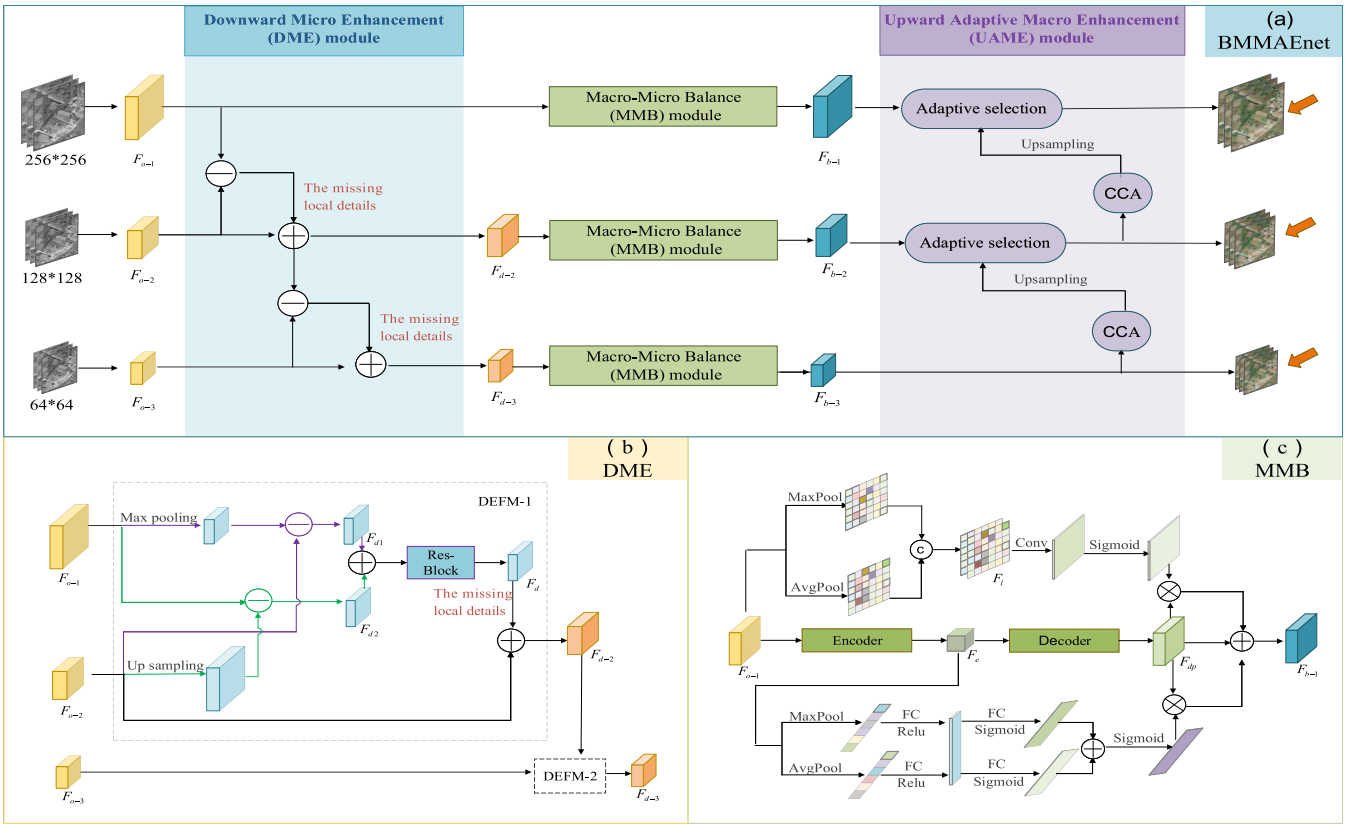


FIGURE 2. (a) Overview of the proposed framework for remote sensing colorization, which contains three branches with different scales, i.e.,micro branch, middle branch and macro branch. The framework consists of three major components: 1) the DME module proposed during the multi-branch downsampling process, 2) the DAME module proposed during the multi-branch upsampling process and 3) the MMB module proposed in each branch. (b) Illustration of the DME module which adds micro details as well as suppressing macro context during the multi-branch downsampling process to supplement missing pixel details. (c) Illustration of MMB modules which is designed by embedding attention-guided local details and global semantics into the decoding features to balance local and global features expression within each branch.

on CGAN to color sketches in two steps by predicting grass-colored sketches and repairing incongruent colors. These approaches are subjected to require a significant number of user inputs, especially when processing the images with complex textures, such as remote sensing images, which is difficult to achieve. Furthermore, selecting the appropriate color palette is not a light process.

B. EXEMPLAR-BASED METHODS

The Exemplar-based methods colorize the target grayscale image by referring to the color image. He et al. [24] proposed an exemplar-based colorization method to realize the transformation of the grayscale image to a color image using a Similarity Sub-net for measuring the semantic similarity of reference and target images and a Colorization sub-net for realizing final colorization. Li [25] et al. proposed a cross-scale texture matching method, which selects the appropriate matching scale locally and then performs global optimization and fusion, to solve the problem of poor coloring effect and robustness caused by the content of mismatching sample and target sizes. Although these methods require less user input than scribble-based methods, they still require high-quality sample images.

C. AUTOMATIC COLORING METHODS

The fully automatic coloring technology without any additional reference is widely used in recent years, especially the emergence of GAN and its variants, which provides an effective means to achieve diversity colorization.

Isola et al. [6] proposed the method that uses Conditional Generative Adversarial Network (CGAN) [3] to realize image colorization and this method is generalized to the colorization because it does not need to set up a specific loss function according to a specific problem. In [6], Isola et al. used U-Net as the generator to extract features and generate images and proposed a PatchGAN as the discriminator to process the high-frequency part of the image by the input divided picture blocks instead of the whole picture. Nazeri et al. extended the colorization process in [26] to high-resolution images and redefined the loss function, which replaces the correct probability of the minimization generator with the wrong probability of the maximization discriminator to solve the problem of unstable coloring process and slow convergence speed. Wang et al. [27] proposed a GAN-based radar image colorization network, and the network proposed denoising preprocessing process to further optimize the coloring effect.

In [6], [27], and [26], the generator architecture using the original U-Net with feature extraction and recovery is prone to loss of detailed information, which is a fatal challenge for the colorization task of assigning colors to each pixel. Therefore, Cao et al. [28] proposes a CGAN that uses convolutions with stride 1 in the generator for ensuring coloring details. Although Cao et al. paid attention to the coloring details, the consistent coloring space is ignored, resulting in discontinuous coloring. To relieve the problem of coloring space discontinuity, Iizuka et al. [29] proposed a CNN-based method of adding classification auxiliary tasks to help colorize through fusing global semantics features extracted in the auxiliary network and the local feature from the coloring main network, which guarantees spatial consistency to a certain extent. Similarly, Vitoria et al. [30] proposed an adversarial learning colorization method termed ChromaGAN, which learns colorization by combining class distribution studies and semantic understanding as well as considering classification loss. However, these methods only embed categorical semantic information when using semantics to colorize colorized images, resulting in semantic confusion and color bleeding in the final color image. To address these issues, Zhao et al. [31] proposed a Salience Map-guided colorization method SCGAN by predicting colorization and salience mapping jointly to reduce semantic confusion and color bleeding in colorized images. Unfortunately, this method based on other task aids is not suitable for remote sensing images, because the complex scenes of remote sensing images are difficult to obtain the ground truth of the auxiliary tasks. The above method of assisting coloring by adding global information reflects the thought of multi-scale actually which can supply constraints on details for ensuring spacial consistency.

More obviously, without extra tasks, Li et al. [7] proposed a multi-discriminator GAN, using multiple independent discriminators to supervise the color images produced by each feature extraction layer in the generator to address the problem of unstable spatial consistency in highly textured regions. However, multi-discriminator only optimizes the measure of Jensen-Shannon divergence [8] but could not apply a strong constraint to ensure space stability. Limmer et al. [32] proposed a multi-branch network applying a typical multi-scale structure, pyramid structure to realize colorization. Wu et al. [9] combined multi-scale convolution with a squeeze-excitation network (SEnet) based on DCGAN to preserve effective image features during image generation and adjust the channel weights during training. Based on this work, Wu et al. transferred the coloring task from RGB to YUV color space and used the multi-scale convolution to optimize the coloring effect [33]. Feng et. al [34] proposed the multi-scale residual receptive field net to extract the essential information of the original color image and then used U-net with an attention mechanism to recover the color. Subsequently, Feng et. al [35] proved the method using the Multi-scale Residual Block to extract features and the Information Recovery Architecture (IRA) with multi-scale information transfer is proposed to generate the color images. The above multi-scale coloring methods inevitably lead to the loss of micro-details in the process of using macro-context information to ensure spatial consistency [36], [37], [38]. Therefore, BMMAEnet is proposed to preserve coloring spatial consistency and obtain object salience at the same time by learning stable macroscopic semantics and recovering microscopic pixels.

III. METHOD

In this section, we introduce the BMMAEnet which recovers the spatially consistent color information while acquiring object salience in context with extremely imbalanced space structures. Firstly, we present the overall architecture of our approach. Secondly, we describe DME module, UAME module, and MMB module separately.

A. OVERVIEW OF THE BMMAEnet

Since coloring grayscale image needs to allocate reasonable color information for each pixel, abundant local detail information is required. However, focusing only on detail pixels can lead to a lack of coloring space consistency, resulting in the generated color images that violate human perception. Therefore, aiming at obtaining object salience while ensuring the consistency of the color space, we use a DCGAN-based multi-branch network as the backbone to implement colorization [4].

As shown in Fig. 2, the generator of this network consists of three branches with different scales, defined as a micro branch, middle branch, and macro branch, each of which uses the U-Net structure to extract micro features containing more local details or macro-features containing more contextual information, respectively. The input of the middle branch and macro branches are obtained by downsampling the original grayscale image with average pooling. It is worth noting that essential local details are inevitably lost in the multi-branch downsampling process. Therefore, to compensate for the lost pixel details, we propose the DME module that extracts micro details by suppressing macro-context and incorporates the micro details into the macro-branch to guarantee coloring object salience. What's more, in the multi-scale structure, macro-semantics are used to constrain micro-pixels by post-fusing multi-scale branches. The naive fusion method, such as concatenation and addition, cannot give full play to the strong constraining effect of macro branches during the multi-branch upsampling process. Therefore, we propose the UAME module, which adaptively selects the proper level of features to strengthen macro semantic constraints, thereby ensuring the consistency of the coloring space.

Additionally, on each branch, to balance micro and macro information, we propose the MMB module, which enhances the representational power of decoded features by fusing the prominent local details of low-level features and the global semantics of embedding features with decoded features. Each branch in the generator outputs a color image as the supervision condition of the network while only the generated color

images from the micro branch are fed into the discriminator for adversarial training.

B. DME MODULE

Focusing on supplementing the essential pixel details lost during the multi-branch downsampling process and ensuring the coloring object salience, we propose the DME module shown in Fig. 2(b). We take DME-1 as an example to further illustrate this module.

The inputs of the DME-1 are the feature $F_{o-1} \in \mathbb{R}^{H\times W\times C}$ and $F_{o-2} \in \mathbb{R}^{h\times w\times C}$ generated from the corresponding grayscales, in which the F_{o-2} from downsampled middle scale loses some pixel details that are crucial for colorization, therefore, we use a dual-step scheme to enhance micro-detail. The first step of DME module is to perform a maximum pooling operation on F_{o-1} to obtain local salient features in the size of $h \times w \times C$ and perform element-wise subtraction between the obtained salient features and F_{o-2} to generate the feature F_{d1}. The second step is to up sample F_{o-2} through uppooling to generate a feature in the size of $H \times W \times C$, and perform element-wise subtraction of the upsampled feature with F_{o-1} which contains more pixel details to generate the feature F_{d2}. The subtraction operation in these two steps suppresses the macro context and removes the complicated redundant information, as well as obtains the lost micro details.

Then, operate convolution operation on the output feature of the second step to generate a feature in the size of $h \times w \times C$, which is performed the element-wise addition operation with the feature generated in the first step to realize the supplement of missing micro details in macro-branch. The micro-details are extracted by a dual-step operation aiming at ensuring the sufficiency and integrity of micro-detail mining.

Finally, the fused features of F_{d1} and F_{d2} are processed through the residual block [39] to obtain the local details feature $F_d \in \mathbb{R}^{h\times w\times C}$ which is operated the element-wise addition with the middle branch feature F_{o-2} to generate the feature $F_{d-2} \in \mathbb{R}^{h\times w\times C}$ with enhanced micro-details. The DME operation can be symbolized as:

$$\begin{aligned} F_{d1} &= \mathrm{Max}\left(F_{o-1}\right) - F_{o-2}, \\ F_{d2} &= F_{o-1} - Up\left(F_{o-2}\right), \\ F_{d-2} &= F_{o-2} + RES\left(\alpha F_{d1} + \beta F_{d2}\right) \end{aligned} \tag{1}$$

where $Max(\cdot)$ and $Up(\cdot)$ denote the max pooling and up pooling, respectively; $RES(\cdot)$ denotes residual block which consists of convolution, ReLU, and shortcut connection; and α and β are the learnable parameters that are initialized as 1.

Similarly, the F_{o-2} and the feature F_{o-3} from the macro branch are input into DEM-2, and they generate the feature $F_{d-3} \in \mathbb{R}^{h'\times w'\times C}$ containing the supplementary details.

With the DME module, the loss of microscopic details during downsampling is reduced, providing a favorable guarantee for the object salience of colorization accordingly.

C. UAME MODULE

Focus on the great challenge to ensure the consistency of the coloring space under the extremely unbalanced spatial distribution of remote sensing images, we propose the UAME module including the adaptive selection scheme and the criss-cross attention to ensure color space consistency by enhancing macro-semantic constraints during the multi-branch upsampling process.

The inputs of the UAME module are the feature F_{b-1}, F_{b-2} and F_{b-3} output from the MMB module in each branch. Firstly, the feature F_{b-3} is performed criss-cross attention [40] process and then obtain features $F_{a-3} \in \mathbb{R}^{h'\times w'\times C}$. The criss-cross attention combines the two cross-forms of attention to obtain the global dependence of any two positions in the feature and realizes the non-local effect while reducing the time and space complexity, which effectively ensures spatial consistency. In addition, to enhance the macro-semantic constraints and fuse the different branches appropriately, we upsample F_{a-3} and then process the upsampled F_{a-3} with the F_{b-2} by adaptive selection calculation. specifically, concatenate the upsampled feature F_{a-3} and F_{b-2} and use two different 1×1 convolution kernels to obtain the transfer information obtained from the macro and middle branches respectively. The operation can be symbolized as:

$$\begin{aligned} T^{a-3} &= Conv\left(Up\left(F_{a-3}\right) \mid F_{b-2}\right), T^{a-3} \in R^{1\times h\times w}, \\ T^{b-2} &= Conv'\left(Up\left(F_{a-3}\right) \mid F_{b-2}\right), T^{b-2} \in R^{1\times h^{\times w}} \end{aligned} \tag{2}$$

where T^{a-3} and T^{b-2} denote the transfer information obtained from the macro branch feature F_{a-3} and the middle branch feature F_{b-2}.

Then, the obtained transfer features are normalized by a softmax function to obtain the infomaiton transfer gate, which can be symbolized as:

$$\begin{aligned} W_{a-3}^{(i,j)} &= \frac{e^{T^{a-3}}}{e^{T^{a-3}} + e^{T^{b-2}}}, \\ W_{b-2}^{(i,j)} &= \frac{e^{T^{b-2}}}{e^{T^{a-3}} + e^{T^{b-2}}} \end{aligned} \tag{3}$$

where $W_{a-3}^{(i,j)}$ and $W_{b-2}^{(i,j)}$ represent the transfer weight of feature F_{a-3} and F_{b-2} at i^{th} row pixel index and j^{th} column pixel index.

Thus, the fusion feature obtained by the adaptive selection scheme can be calculated as follows:

$$F_{fu} = T^{a-3} \cdot W_{a-3} + T^{b-2} \cdot W_{b-2} \tag{4}$$

Then, the F_{fu} will be fused with the feature F_{b-1} from the micro branch by CCA and adaptive selection calculation, which is similar to the above fusion operations between the macro branch and the middle branch, to achieve the full net fusion.

D. MMB MODULE

To simultaneously guarantee color space consistency and object salience, as shown in Fig. 2(c), we propose the MMB

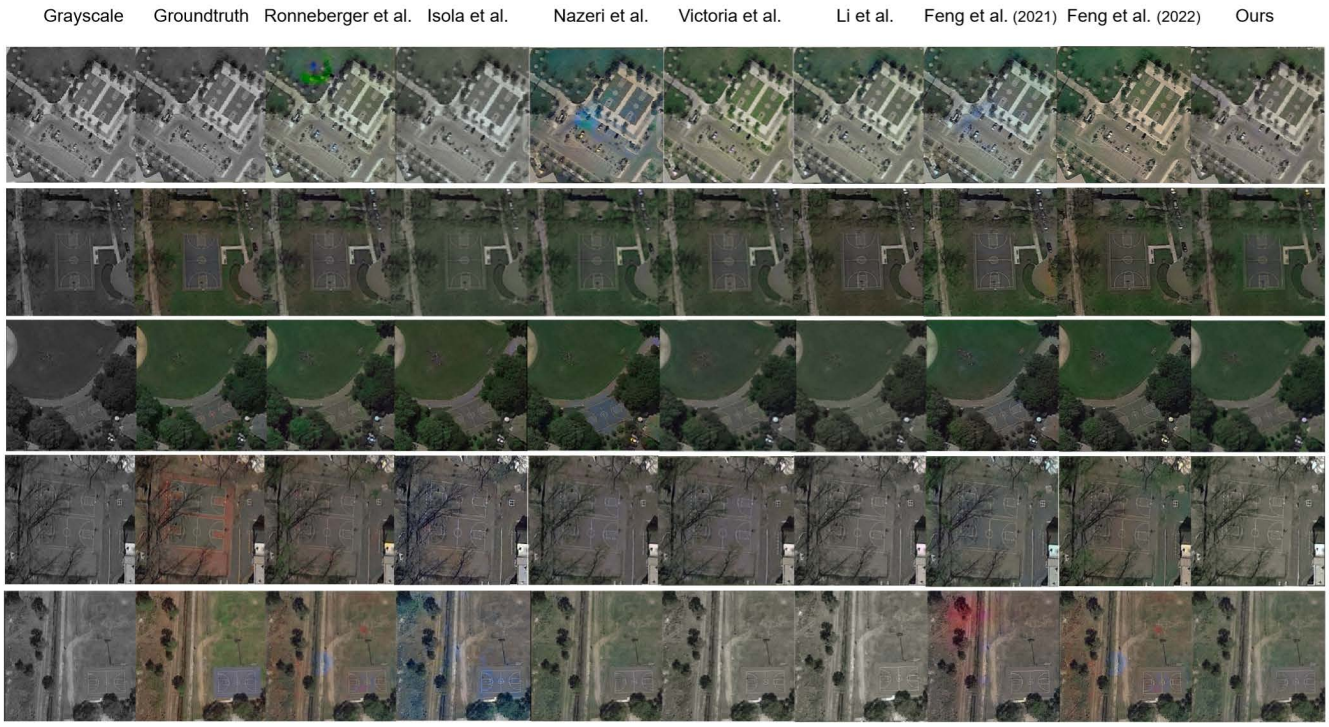


FIGURE 3. The generated images of the contrast experiments about the playground. By comparing the images generated by different methods, it is obvious that the images generated by our method have stronger spatial consistency and stronger object salience and are more in line with human understanding. For example, the image in the first row and in the last column is more realistic which is different from the image with apparent light spots generated by Nazeri et al.'s method.

module by attention-guided local details and global semantics embedding to balance local details and global semantics.

The MMB module in each branch uses U-Net as the skeleton to restore local details lost in the upsampling process through skip connection. However, different positions in shallow features have different contributions to the recovery of local details and direct concatenation of the low-level feature and high-level feature cannot reasonably utilize essential local details information. Therefore, before fusing the low-level and high-level features, we first use the attention mechanism [41], [42] to highlight the local details with more contribution for the object salience in colorization. Take the MMB module at the micro branch as an example, the low-level feature F_{o-1} is fed into the encoder to generate feature F_e containing global semantics which is used to generate the deep layer feature F_{dp} through the decoder. On this basis, firstly, we obtain two features in size of $H \times W \times 1$ by operating the maximum pooling and average pooling on low-level feature $F_{o-1} \in \mathbb{R}^{H\times W\times C}$ in the location dimension and concatenate the obtained features to generate feature $F_l \in \mathbb{R}^{H\times W\times 2}$ highlighting the efficient local information. Then, the convolution layer and Sigmoid [43] layer are performed on F_l to obtain contribution weight features in the size of $H \times W \times 1$ of different spatial positions. Resize the spatial weight feature to $H \times W \times C$ and then multiplied it with the decording feature F_{dp} to generate the feature $F'_l \in \mathbb{R}^{H\times W\times C}$ which embeds information containing rich local details. The above operation can be symbolized as:

$$\begin{aligned} F_l &= Max(F_{o-1})|Avg(F_{o-1}), \\ F'_l &= \sigma(Conv(F_l)) \end{aligned} \tag{5}$$

where $Max(\cdot)$ denotes max pooling, $Avg(\cdot)$ denotes average pooling, $Conv(\cdot)$ denotes convolution operation and σ denotes Sigmoid function.

Then, to avoid the problem of weakening the representation ability of global semantics that can guarantee spatial continuity due to the focus on recovering global details, we also fuse the embedding feature containing global semantics into high-level decoding features. Specifically, we perform channel-wise max-pooling and average-pooling operations on the embedding feature $F_e \in \mathbb{R}^{h\times w\times c}$, respectively. Then, perform full connection and activation function operations on the two generated features, and add the obtained results to generate the feature $F_c \in \mathbb{R}^{1\times 1\times c}$. Afterward, the F_c is processed by several fully collection layers and Sigmoid function to obtain the contribution weights of different channels. The feature $F'_c \in \mathbb{R}^{h\times w\times c}$ is generated by multiplying the channel weight with the deep layer decoding feature F_{dp}, which enhances the semantic representation of the feature from a global perspective, thus ensuring the coloring spacial consistency. The above operation can be symbolized as:

$$\begin{aligned} F_c &= FC(Max(F_e))) + FC(Avg(F_e)), \\ F'_c &= \sigma(FC(F_c)) \end{aligned} \tag{6}$$

FIGURE 4. The generated images of the contrast experiments about the building. By comparing the images generated by different methods, it is obvious that the images generated by our method have stronger spatial consistency and stronger object salience and are more in line with human understanding.

where $FC(\cdot)$ represents the fully connection operation and activation function.

Finally, the features F_s', F_c', and F_{dp} are fused by adding element-wise to generate the feature F_{b-1} which achieves a balance between local details and global semantics. The above operation can be symbolized as:

$$F_{b-1} = F_s' + F_c' + F_{dp} \tag{7}$$

The MMB module in the macro and middle branches are similar and we will repeat no more.

E. OBJECTIVE FUNCTION

To encourage the similarity between the real color image and the generated image as well as promote the generated images to cater to human perspective understanding, we propose an objective function suitable for the colorization task which consists of GAN loss and Huber loss. The overall loss function is defined by equ (8):

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{GAN} + \sum_{i=1}^{3} \lambda_i \mathcal{L}_{\text{Huber}} \tag{8}$$

where $\lambda_1 = \alpha$, $\lambda_2 = \beta$ and $\lambda_3 = 100 - \alpha - \beta$.

1) DCGAN LOSS

In the BMMAEnet, the generator is used to restore the color information of the grayscales and the discriminator obtains the probability that the input belongs to generated images. Especially, the optimization of the generator is defined to minimize the probability of the discriminator correctly predicting the generated image, while the optimization of the discriminator is defined to maximize the probability of correctly distinguishing the generated image from the original image. According to [26], the generator loss is given by equ (9):

$$\mathcal{L}^{(G_i)}\left(\theta_D, \theta_{G_i}\right) = -\mathbb{E}_{x_i\ P(x_i)}\left[\log\left(D\left(G\left(x_i\right)\right)\right)\right] \tag{9}$$

The loss function of discriminator is given by equ (10):

$$\begin{aligned}\mathcal{L}^{(D)}(\theta_D, \theta_{G_i}) = {} & \mathbb{E}_{y\sim P(y)}[\log(D(y|x_i))] \\ & + \mathbb{E}_{x_i\sim P(x_i)}[\log(1 - D(G(x_i|x_i)))]\end{aligned} \tag{10}$$

where i represents the i^{th} branch of the generator, and $i = 1, 2, 3$. The generator and discriminator are trained alternately to realize network optimization.

2) COLOR ERROR LOSS

To encourage less blurring, previous methods mix the traditional loss such as L_1 distance [6] or L_2 distance [44] with the GAN objective. However, L_1 always has a large gradient which makes the model difficult to train and learn. Meanwhile, L_2 has poor robustness to outliers, which leads to excessive outliers contribution and reduces the overall performance of the model. To avoid the disadvantages of the above two, we choose Huber (or smooth L1) [45] as the color error loss and add it to the objective function of the generator to guarantee the accuracy of the coloring. The Huber loss is

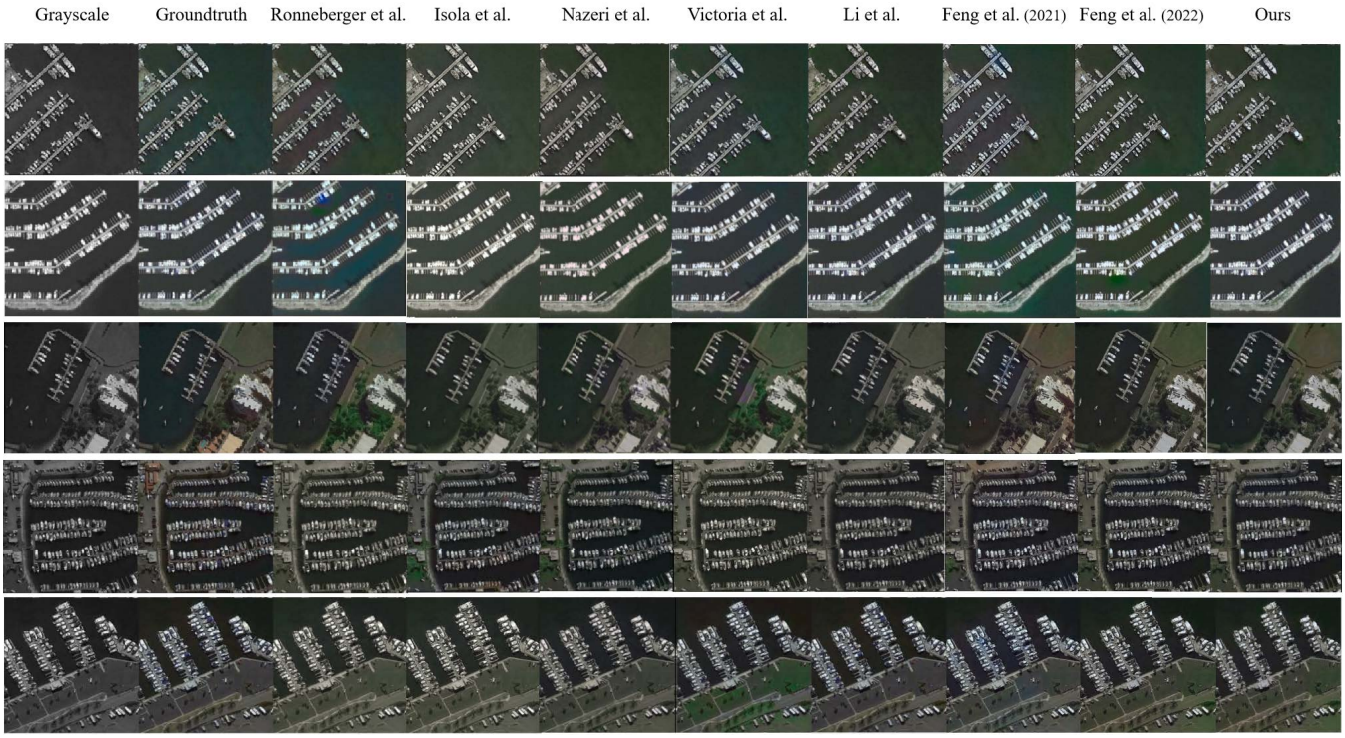


FIGURE 5. The generated images of the contrast experiments about the harbour. By comparing the images generated by different methods, it is obvious that the images generated by our method have stronger spatial consistency and stronger object salience and are more in line with human understanding.

given by equ (11):

$$\mathcal{L}_{\text{Huber}} = \begin{cases} \frac{1}{2}\left(G\left(x_i\right) - y_i\right)^2 & \text{for } \left|G\left(x_i\right) - y_i\right| < \delta \\ \delta\left|G\left(x_i\right) - y_i\right| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases} \tag{11}$$

where, $i = 1, 2, 3$. We set $\delta = 1$ in equ (11).

IV. EXPERIMENTS

In this section, we introduce the data set, evaluation metrics, and comparative networks used in the experiment and show the essential experimental details. Moreover, we validate our method by comparing experiment and ablation experiments on both qualitative and quantitative metrics.

A. DATASET AND IMPLEMENTATION DETAILS

1) NWPU DATA SET

NWPU Data Set [46] is one of the latest data sets of remote sensing images. The NWPU-resisc45 data set is a remote sensing image scene classification benchmark created by Northwestern Polytechnical University (NWPU). The data set consists of 45 scene categories and each category consists of 700 images. In total, this data set has 31,500 images in the size of 256×256. Our network uses 500, 100, and 100 images in each category as training sets, validation sets, and test sets respectively.

2) AID DATASET

AID Data Set [47] is one of the commonly used remote sensing image data Sets created by Huazhong University of Science and Technology and Wuhan University. It contains 30 categories of images such as desert, port, and buildings. Each category contains between 220 and 420 images and in total 10,000 images in the size of 600*600. We selected about 60 percent of the images in each category as the training set and the rest is evenly divided into the test set and validation set. All images are pre-processed before use.

B. TRAINING DETAILS

The network we proposed is trained on an NVIDIA Tesla V100-SXM2-16GB GPU using the TensorFlow framework [48]. We set 8 as the batch size and trained 30 epochs to choose the best result. We used the Adam algorithm [49] with a learning rate of 0.003.

C. EVALUATION METRICS

We evaluate the experimental results using four metrics: Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM), accuracy, and Amazon Mechanical Turk (AMT) [6], [50]. PSNR, SSIM, and accuracy are quantitative metrics. AMT is a qualitative metric and the higher the score of AMT, the closer the colored image is to conforming to people's perceptions.

FIGURE 6. The generated images of the ablation experiments. The images generated by the multi-branch structure have stronger spatial consistency than the images generated by the single-branch structure. For example, the first row and the third column of the pictures contain distinguishable blue areas from the surrounding area which disappear in the fifth column. After adding the DME module, the generated images have stronger object salience, such as the image shown in the first row in the sixth column. Inconsistent colors disappear after adding the UAMB module, as shown in the seventh column of the first row, demonstrating the role of this module in enhancing color space consistency. The addition of the MMB module further optimizes the generation results by balancing object salience and spatial consistency, as shown in the first row and the last column.

1) ACCURACY

We calculate the accuracy by calculating the ratio of the number of pixels in a certain range between the original image and the colored image to the total number of pixels. The formula is as follows:

$$acc(x, y) = \frac{1}{n}\sum_{p=1}^{n}\prod_{l=1}^{3} 1_{[0,\varepsilon_l]}(|h(x)^{(p,l)} - y^{(p,l)}|) \quad (12)$$

In equ (12), ε represents the threshold which is set to 2 and 5 and the different thresholds can get the description of the varying degrees of coloring accuracy. If the difference between the corresponding pixels of the generated image and the original image in all three channels is within the threshold, the position is considered to have been colored successfully.

2) PSNR

We use Peak Signal to Noise Ratio (PSNR), one of the most common and widely used image evaluation metrics, as one of the qualitative evaluation metrics. It is based on the calculation based on the error between the corresponding pixels and the larger the PSNR value, the better the generated color image. The formula of PSNR is as follows:

$$PSNR = 10 \times \log_{10}\left[\frac{(L)^2}{MSE}\right], \quad MSE = \frac{1}{N} \parallel y - \hat{y} \parallel_F^2 \quad (13)$$

where, L denotes the dynamic range of pixel values, N denotes the number of the pixels, and $\| \cdot \|_F^2$ denotes the Frobenius norm.

3) SSIM

We use Structural Similarity (SSIM) as one of the qualitative evaluation indicators for colorization results, which measures image similarity from three aspects including brightness, contrast, and structure. The value range of SSIM is from 0 to 1 and the larger the value, the better the generated color image effect. The formula of SSIM is as follows:

$$SSIM = \frac{\left(2\mu_y\mu_{\hat{y}} + c_1\right)\left(\sigma_{y\hat{y}} + c_2\right)}{\left(\mu_y^2 + \mu_{\hat{y}}^2 + c_1\right)\left(\sigma_y^2 + \sigma_{\hat{y}}^2 + c_2\right)},$$
$$c_1 = (k_1 L)^2, \quad c_2 = (k_2 L)^2 \quad (14)$$

where μ_y and $\mu_{\hat{y}}$ represent the mean of y and $\hat{y}$, σ_y^2 and $\sigma_{\hat{y}}^2$ represent the variance of y and $\hat{y}$ respectively, $\sigma_{y\hat{y}}$ represents the covariance of y and $\hat{y}$, k_1 is set to 0.01, and k_2 is set to 0.03.

4) AMAZON MECHANICAL TURK(AMT) PERCEPTION TEST

We chose the Amazon Mechanical Turk(AMT) perception test to evaluate the network since the ultimate goal of remote sensing image colorization is to make the generated images look real and meet people's sensory needs. We looked for

TABLE 1. Contrast experiments results on NWPU data set.

Method	acc2	acc5	PSNR	SSIM(%)	AMT
Ronneberger et al.	36.05 ± 0.17	88.42 ± 0.13	27.84 ± 0.12	82.16 ± 0.05	42.25
Isola et al.	37.56 ± 0.12	86.61 ± 0.21	28.52 ± 0.06	85.27 ± 0.16	46.75
Nazeri et al.	37.02 ± 0.13	88.05 ± 0.07	29.94 ± 0.10	84.98 ± 0.12	59.93
Li et al.	38.30 ± 0.09	87.12 ± 0.19	30.27 ± 0.07	85.73 ± 0.19	70.12
Wu et al.	37.91 ± 0.18	86.81 ± 0.15	30.55 ± 0.09	86.46 ± 0.07	71.48
Feng et al. (2021)	38.07 ± 0.05	86.73 ± 0.13	31.32 ± 0.14	87.42 ± 0.08	74.34
Feng et al. (2022)	39.61 ± 0.13	88.96 ± 0.15	32.37 ± 0.17	**89.32 ± 0.14**	80.45
Ours	**41.38 ± 0.14**	**89.27 ± 0.11**	**33.84 ± 0.06**	89.27 ± 0.10	**82.71**

TABLE 2. Contrast experiments results on AID data set.

Method	acc2	acc5	PSNR	SSIM(%)	AMT
Ronneberger et al.	22.89 ± 0.14	78.82 ± 0.06	21.85 ± 0.15	76.82 ± 0.16	38.64
Isola et al.	23.07 ± 0.11	78.04 ± 0.16	22.44 ± 0.17	77.82 ± 0.14	42.80
Nazeri et al.	23.42 ± 0.16	79.36 ± 0.09	22.26 ± 0.14	77.82 ± 0.12	45.53
Li et al.	24.15 ± 0.19	79.18 ± 0.11	24.35 ± 0.06	78.82 ± 0.07	58.64
Wu et al.	26.53 ± 0.08	79.69 ± 0.20	24.25 ± 0.12	79.82 ± 0.09	64.81
Feng et al. (2021)	28.37 ± 0.16	80.68 ± 0.14	24.32 ± 0.11	80.39 ± 0.18	74.79
Feng et al. (2022)	29.53 ± 0.10	81.91 ± 0.19	25.24 ± 0.10	82.14 ± 0.06	78.51
Ours	**31.37 ± 0.15**	**83.05 ± 0.13**	**26.97 ± 0.13**	**83.97 ± 0.17**	**73.97**

20 people to participate in the AMT test, which showed participants 10 real images and 20 generated images in disorder one by one, and asked participants to score 1 to 100 points for each image. The higher the score, the participants think the displayed image is closer to the real image. The final experimental data is the average number of points each participant gave to each picture.

D. COMPARATIVE NETWORKS

1) RONNEBERGER et al. 's METHOD [11]

Use U-Net structure to color remote sensing image by feature extraction and recovery of the input image, and each layer in the extraction process is added to the corresponding restoration process layer.

2) ISOLA et al. METHOD [6]

Use CGAN to achieve multiple generation tasks. The generator is conditional on an input image and uses the U-Net structure and PatchGAN is used as a classifier in the discriminator.

3) NAZERI et al. METHOD [26]

Use DCGAN to achieve remote sensing image colorization. Using U-Net structure as the generator and grayscale image is taken as the initial condition of the generator.

4) LI et al. METHOD [7]

Use different scale multi-discriminator GAN to discriminate each stage results, and the discriminator input of each layer integrates the results of the upper layers.

5) WU et al. METHOD [33]

Use multi-scale DCGAN to color remote sensing images, and multi-scale is implemented by convolution operation with different kernel sizes in the generator.

6) FENG et al. METHOD [34]

Use multiple multi-scale residual receptive filed blocks (MRRFB) to extract features, and then use U-Net as the basic structure to build a color information recovery network (CIR-Net) to recover color information.

7) FENG et al. METHOD [35]

Use the Multi-scale Residual Block to extract features, and then the final coloring is achieved through the Information Recovery Architecture (IRA) with multi-scale information transfer.

E. CONTRAST EXPERIMENTS

Table 1 and Table 2 show the experimental results of the method we proposed compared to the state-of-art methods on NWPU Data Set and AID Data Set respectively. Experimental results demonstrate that our approach outperforms all the other approaches with a large margin on all qualitative and quantitative metrics. For instance, compared with the state-of-the-art colorization method proposed by Wu et al. [33], our method improves acc2 by 3.5% and 4.8% on the NWPU Date Set and AID Data Set, respectively.

Fig. 3, Fig. 4, and Fig. 5 show the qualitative comparison of images, belonging to the playground, building, and harbor classes respectively, generated by our proposed method and other methods. It is not difficult to find that the color images generated by our method have stronger object salience due to the DME module which enhances the local details during the downsampling process.

In addition, images generated by Nazeri's method have evident light spots different from surrounding colors caused by unbalanced spatial distribution, like the first row in Fig. 3, which go against human visual perception. Compared with other methods, the images generated by our method have a strong spacial consistency which benefits from constraints of

macro semantics on micro details while the proposed UAME module further strengthens the macroscopic semantic constraints by adaptive selecting proper transfer information during multi-branch upsampling process.

Another reason causing our approach to be superior is that our proposed MMB module balances micro and macro information by using valuable local details and global semantics, further forcing the network to generate more realistic color images with spatial consistency as well as object salience.

We show several more intuitive examples to prove that our proposed method can generate color images with stronger spatial consistency and stronger object salience than others. In Fig. 7, The first and fourth columns show images generated by our methods and others, respectively. The red box shows that the images generated from our method have stronger space consistency, for example, the image in the third row and the fourth column has an obvious blue spot which is inconsistent with the surroundings but does not exist in our image. In addition, we circle the object in the generated image with a yellow box and enlarged it for comparison. It can be found that the object in the image generated by our method has a clear outline, higher color saturation, and stronger salience.

We make the accuracy metrics into box plots shown in Fig. 8. By observing the box plot, we can find that our accuracy metrics have a higher average compared with the contrast methods. In addition, the most compact distribution of our results shows that our method can provide more stable coloring than other methods.

F. ABLATION EXPERIMENTS

To verify the effectiveness of the proposed network, we conducted ablation experiments on single-branch and multi-branch networks using NWPU Data Set and AID Data Set respectively. Table 3 and Table 4 are the overall experimental results that incrementally evaluate the effectiveness of each module and the visual effects are shown in Fig. 6. Subsequently, we will analyze each module in detail through experimental data and visual effects shown in generated color images.

1) MULTI-BRANCH

We used Nazeri et al. 's method with the single branch as the baseline for the ablation experiment, which is represented as 1-B. The middle branch and the macro branch are successively added to the generator to form networks 2-B and 3-B. Table 3 and Table 4 show that the multi-branch network has better coloring performance than the single-branch network. This is because the macroscopic semantic constraints of macroscopic branches on microscopic branches ensure the spatial continuity of generated images to some extent and reduce the appearance of abrupt points, which promotes the generated images better conform to the sensory experience of people.

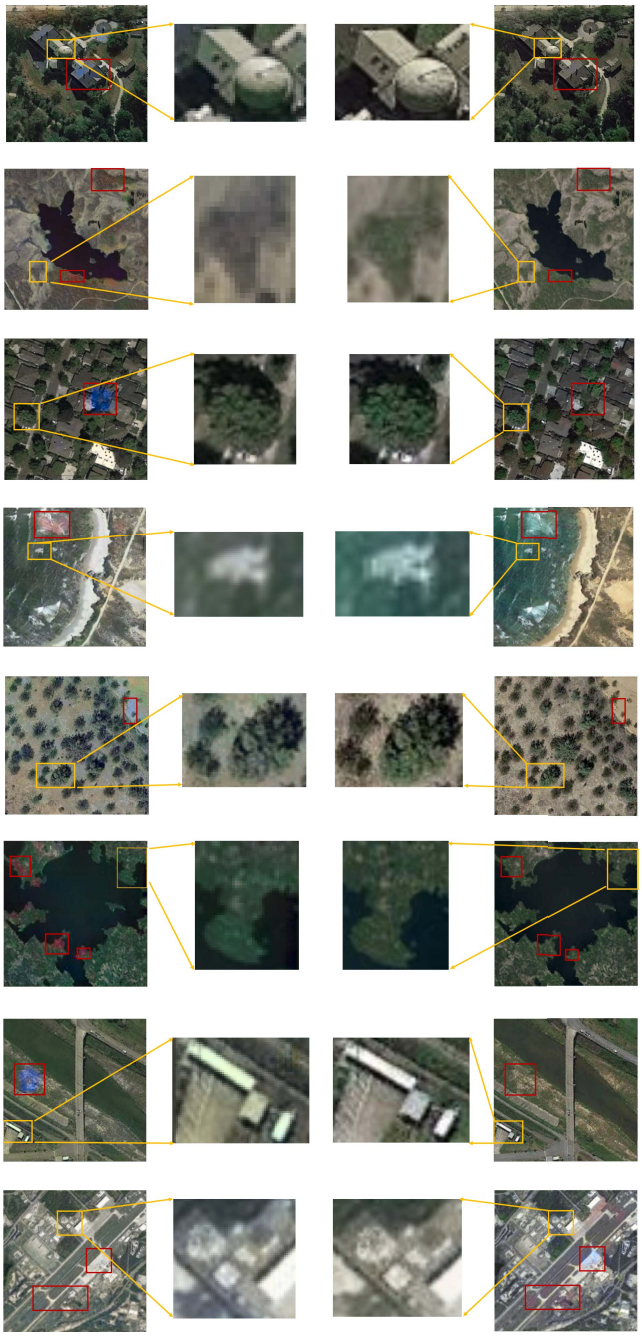


FIGURE 7. The comparison of the generated images. The first and fourth columns show images generated by our methods and others, respectively. The area is circled using red and yellow boxes proving the stronger spatial consistency and the stronger object salience of our generated images respectively.

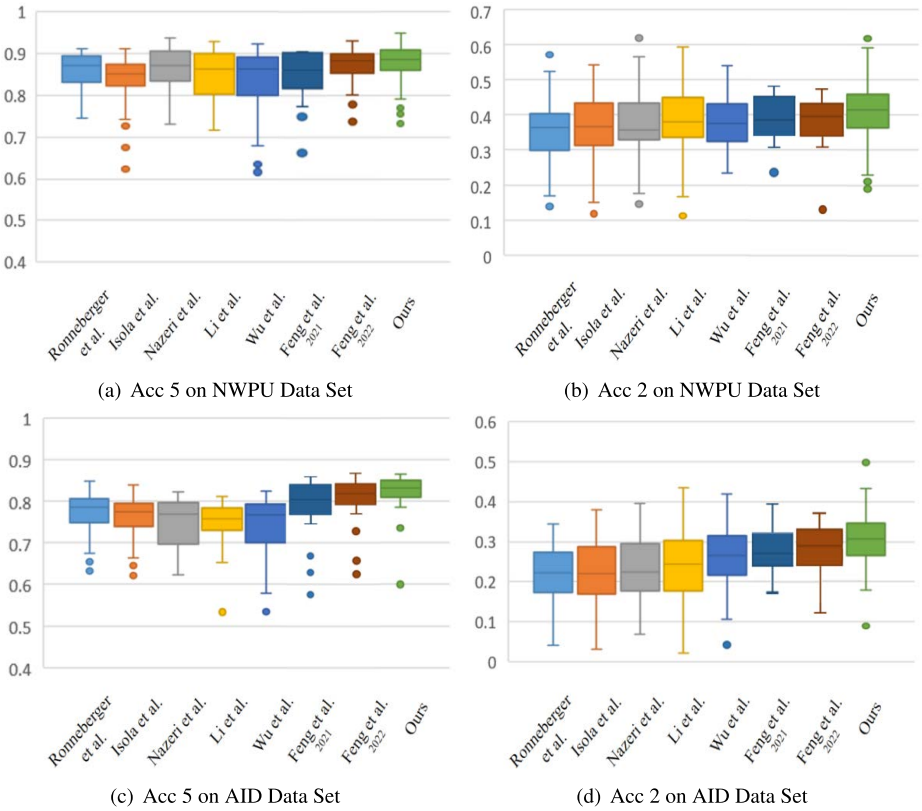
2) DME MODULE

DME is proposed to enhance the micro details missing during the downsampling process. We added DME to 2-B and 3-B respectively and the experimental result was improved to a certain degree.

It is worth noting that DME improved 3-B significantly better than 2-B which demonstrates the importance of supplementing missing details for the downsampled sample. This

TABLE 3. The ablation experiments results on NWPU data set.

Method	Acc2	Acc5	PSNR	SSIM(%)	AMT
1-B	37.02 ± 0.13	88.05 ± 0.07	29.94 ± 0.10	84.98 ± 0.12	59.93
1-B+MMB	38.07 ± 0.05	88.32 ± 0.03	30.15 ± 0.14	85.54 ± 0.13	67.12
2-B	38.19 ± 0.23	86.85 ± 0.05	29.83 ± 0.17	85.78 ± 0.19	63.69
2-B+DEM	38.61 ± 0.10	86.92 ± 0.21	30.09 ± 0.12	86.98 ± 0.21	67.72
2-B+DEM+UAME(o/w AS)	38.79 ± 0.08	87.43 ± 0.13	31.25 ± 0.14	87.05 ± 0.13	69.56
2-B+DEM+UAME	39.03 ± 0.18	87.98 ± 0.06	31.48 ± 0.16	87.64 ± 0.08	72.06
2-B+DEM+UAME+MMB	39.51 ± 0.12	88.51 ± 0.21	31.96 ± 0.12	88.03 ± 0.12	73.52
3-B	38.63 ± 0.15	87.19 ± 0.17	30.11 ± 0.06	87.35 ± 0.19	66.24
3-B+DEM	39.59 ± 0.27	87.61 ± 0.13	30.78 ± 0.07	87.86 ± 0.17	70.75
3-B+DEM+UAME(o/w AS)	40.23 ± 0.23	88.13 ± 0.09	31.59 ± 0.07	88.39± 0.06	73.48
3-B+DEM+UAME	40.68 ± 0.06	88.64 ± 0.12	32.48 ± 0.07	88.85 ± 0.15	77.79
3-B+DEM+UAME+MMB	**41.38 ± 0.14**	**89.27 ± 0.11**	**33.84 ± 0.06**	**89.27 ± 0.10**	**82.71**



(a) Acc 5 on NWPU Data Set (b) Acc 2 on NWPU Data Set

(c) Acc 5 on AID Data Set (d) Acc 2 on AID Data Set

FIGURE 8. Box plot of accuracy distributions for the method we proposed and comparative methods, which shows our method has higher and more stable average accuracy metrics compared with others.

phenomenon further validates the importance of our proposed DME module for enhancing micro details in the multi-branch downsampling process.

3) UAME MODULE

We continue to add the UAME module on the 2-B network and 3-B network respectively to verify its effectiveness, and the experimental results after the addition of UAME modules were improved compared with those before. We divided the UAME into two parts, i.e., CCA and the adaptive selection and we add them to the previous experiment setting successively. The experiment data shows that the CCA is useful for the consistency of coloring since it builds global dependency between each feature vector. While the introduction of the adaptive selection further improves the effect because it selects appropriate features during the multi-branch upsampling process adaptively which further ensures spatial consistency of coloring. Therefore, the UAME module solves the problem of weak macro semantic constraints caused by the inefficient selection of macro constraint features in the method using concatenating for realizing multi-branch fusion.

TABLE 4. The ablation experiments results on AID data set.

Method	Acc2	Acc5	PSNR	SSIM(%)	AMT
1-B	23.42 ± 0.16	79.36 ± 0.09	22.26 ± 0.12	77.82 ± 0.17	45.53
1-B+MMB	24.88 ± 0.13	78.82 ± 0.04	23.13 ± 0.17	78.42 ± 0.13	49.53
2-B	24.75 ± 0.22	78.56 ± 0.20	23.24 ± 0.08	78.82 ± 0.08	49.46
2-B+DEM	25.63 ± 0.17	79.42 ± 0.18	23.96 ± 0.05	79.54 ± 0.15	53.90
2-B+DEM+UAME(o/w AS)	26.38 ± 0.20	80.07 ± 0.06	24.53 ± 0.13	80.12 ± 0.13	58.98
2-B+DEM+UAME	27.57 ± 0.06	81.16 ± 0.09	24.80 ± 0.19	80.85 ± 0.18	63.41
2-B+DEM+UAME+MMB	28.02 ± 0.11	81.69 ± 0.17	25.26 ± 0.12	80.97 ± 0.14	65.34
3-B	26.85 ± 0.18	79.94 ± 0.22	24.06 ± 0.20	79.02 ± 0.19	55.76
3-B+DEM	28.12 ± 0.24	81.23 ± 0.15	24.85 ± 0.14	79.74 ± 0.08	61.37
3-B+DEM+UAME(o/w AS)	29.06 ± 0.08	81.59 ± 0.11	25.53 ± 0.11	80.78 ± 0.07	65.72
3-B+DEM+UAME	29.85 ± 0.19	82.26 ± 0.24	26.69 ± 0.17	82.12 ± 0.18	70.75
3-B+DEM+UAME+MMB	**31.37 ± 0.15**	**83.05 ± 0.13**	**26.97 ± 0.13**	**83.97 ± 0.17**	**73.97**

4) MMB MODULE

MMB improves the generation effect by embedding local details and global semantics to balance micro and macro information within each branch. On the basis of the above experiments, we continue to add MMB modules on 1-B, 2-B, and 3-B networks respectively to clarify its effectiveness. After adding MMB, the qualitative and quantitative metrics have been improved obviously, which is because the MMB module supplements the valuable local details and global semantics to the decoding features and balances the micro and macro features while enriching the decoding features, which ensures the object salience and spatial consistency of the coloring.

G. DISCUSS

The BMMAEnet we proposed can color the grayscale image with slight object occlusions, such as haze or thin clouds because slight occlusion will reduce the color contrast while it will not cause local semantic loss. However, in the case of missing image areas and semantic ambiguity caused by cloud blocking, cloud removal [51] should be considered as the prerequisite for colorization and then recovery of color information.

V. CONCLUSION

In this work, we strive to embrace challenges toward realistic remote sensing image colorization. Focusing on the extremely unbalanced spatial structure of remote sensing images, we propose a novel BMMAEnet including three modules, i.e., the DME module, UAME module, and MMB module, to realize colorization. During restoring the color information, the DME module aims at obtaining object salience in context by supplementing missing pixel details, while the UAMB module enhances macro semantic constraints by selecting fusion information adaptively to preserve spatial consistency. In addition, the MMB module further ensures spatial consistency and object salience and achieves a relative balance between them. The method we proposed is verified on NWPU Data Set and AID Data Set, and the results demonstrate the method has been improved compared with state-of-the-art methods in both qualitative and quantitative metrics.

Although the pure data-driven method of BMMAEnet has achieved satisfactory coloring results, the interpretability is weak. Therefore, in the future, we will consider introducing a physical model [52] to assist in colorizing based on the existing work which can enhance the interpretability of the model while improving the coloring effect.

REFERENCES

[1] D. Hou, W. Zhang, K. Chen, S.-J. Lin, and N. Yu, "Reversible data hiding in color image with grayscale invariance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 363–374, Feb. 2019, doi: 10.1109/TCSVT.2018.2803303.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[3] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *Comput. Sci.*, pp. 2672–2680, 2014.

[4] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Comput. Conf.*, 2015.

[5] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," Tech. Rep., 2017.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[7] F. Li, L. Ma, and J. Cai, "Multi-discriminator generative adversarial network for high resolution gray-scale satellite image colorization," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 3489–3492.

[8] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1858–1860, Jul. 2003.

[9] M. Wu, X. Jin, Q. Jiang, S.-J. Lee, L. Guo, Y. Di, S. Huang, and J. Huang, "Remote sensing image colorization based on multiscale SEnet GAN," in *Proc. 12th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2019, pp. 1–6.

[10] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2015, pp. 234–241.

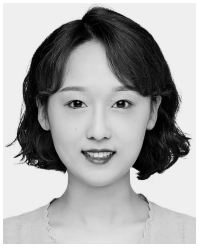
[12] I. Zeger and S. Grgic, "An overview of grayscale image colorization methods," in *Proc. Int. Symp. ELMAR*, Sep. 2020, pp. 109–112.

[13] H. Wang and X. Liu, "Overview of image colorization and its applications," in *Proc. IEEE 5th Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Mar. 2021, pp. 1561–1565.
[14] S. Anwar, M. Tahir, C. Li, A. Mian, F. S. Khan, and A. W. Muzaffar, "Image colorization: A survey and dataset," 2020, *arXiv:2008.10774*.
[15] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5400–5409.
[16] Z. Dou, N. Wang, B. Li, Z. Wang, H. Li, and B. Liu, "Dual color space guided sketch colorization," *IEEE Trans. Image Process.*, vol. 30, pp. 7292–7304, 2021.
[17] Z. Cheng, F. Meng, and J. Mao, "Semi-auto sketch colorization based on conditional generative adversarial networks," in *Proc. 12th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2019, pp. 1–5.
[18] H. Ren, J. Li, and N. Gao, "Two-stage sketch colorization with color parsing," *IEEE Access*, vol. 8, pp. 44599–44610, 2020.
[19] H. Ren, J. Li, and N. Gao, "Automatic sketch colorization with tandem conditional adversarial networks," in *Proc. 11th Int. Symp. Comput. Intell. Design (ISCID)*, Dec. 2018, pp. 11–15.
[20] F. Pierre, J.-F. Aujol, A. Bugeau, N. Papadakis, and V.-T. Ta, "Exemplar-based colorization in RGB color space," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 625–629.
[21] B. Li, F. Zhao, Z. Su, X. Liang, Y.-K. Lai, and P. L. Rosin, "Example-based image colorization using locality consistent sparse representation," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5188–5202, Nov. 2017.
[22] D. Varga and T. Szirányi, "Twin deep convolutional neural network for example-based image colorization," in *Proc. Int. Conf. Comput. Anal. Images Patterns*. Springer, 2017, pp. 184–195.
[23] B. Li, Y.-K. Lai, and P. L. Rosin, "Example-based image colorization via automatic feature selection and fusion," *Neurocomputing*, vol. 266, pp. 687–698, Nov. 2017.
[24] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplar-based colorization," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–16, 2018.
[25] B. Li, Y.-K. Lai, M. John, and P. L. Rosin, "Automatic example-based image colorization using location-aware cross-scale matching," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4606–4619, Sep. 2019.
[26] K. Nazeri, E. Ng, and M. Ebrahimi, "Image colorization using generative adversarial networks," in *Proc. Int. Conf. Articulated Motion Deformable Objects*. Springer, 2018, pp. 85–94.
[27] P. Wang and V. M. Patel, "Generating high quality visible images from SAR images using CNNs," in *Proc. IEEE Radar Conf. (RadarConf)*, Apr. 2018, pp. 0570–0575.
[28] C. Yun, Z. Zhou, W. Zhang, and Y. Yong, "Unsupervised diverse colorization via generative adversarial networks," in *Machine Learning and Knowledge Discovery in Databases*. Cham, Switzerland: Springer, 2017.
[29] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color! Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 110.1–110.11, 2016.
[30] P. Vitoria, L. Raad, and C. Ballester, "ChromaGAN: Adversarial picture colorization with semantic class distribution," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2445–2454.
[31] Y. Zhao, L. M. Po, K. W. Cheung, W. Y. Yu, and Y. Rehman, "SCGAN: Saliency map-guided colorization with generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3062–3077, Aug. 2020.
[32] M. Limmer and H. P. A. Lensch, "Infrared colorization using deep convolutional neural networks," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 61–68.
[33] M. Wu, X. Jin, Q. Jiang, S.-J. Lee, W. Liang, G. Lin, and S. Yao, "Remote sensing image colorization using symmetrical multi-scale DCGAN in YUV color space," *Vis. Comput.*, vol. 37, no. 7, pp. 1707–1729, Jul. 2021.
[34] J. Feng, Q. Jiang, X. Jin, S.-J. Lee, S. Huang, and S. Yao, "Remote sensing image colorization based on deep neural networks with multi-scale residual receptive filed," *J. Comput.-Aided Des. Comput. Graph.*, vol. 33, no. 11, pp. 1658–1667, Nov. 2021.
[35] J. Feng, Q. Jiang, C.-H. Tseng, X. Jin, L. Liu, W. Zhou, and S. Yao, "A deep multitask convolutional neural network for remote sensing image super-resolution and colorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.
[36] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Computer Vision—ECCV 2016*. Cham, Switzerland: Springer, 2016.
[37] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[38] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8759–8768.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[40] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.
[41] J. Park, S. Woo, J. Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*.
[42] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Computer Vision—ECCV 2018*. Cham, Switzerland: Springer, 2018.
[43] D. J. Finney, *Probit Analysis: A Statistical Treatment of the Sigmoid Response Curve*. Cambridge, U.K.: Cambridge Univ. Press, 1952.
[44] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.
[45] M. Schmidt, G. Fung, and R. Rosales, "Fast optimization methods for L1 regularization: A comparative study and two new approaches," in *Proc. Eur. Conf. Mach. Learn.* Springer, 2007, pp. 286–297.
[46] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[47] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[48] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.
[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[50] 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.
[51] Z. Xu, K. Wu, W. Wang, X. Lyu, and P. Ren, "Semi-supervised thin cloud removal with mutually beneficial guides," *ISPRS J. Photogramm. Remote Sens.*, vol. 192, pp. 327–343, Oct. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271622002350
[52] Z. Xu, K. Wu, L. Huang, Q. Wang, and P. Ren, "Cloudy image arithmetic: A cloudy scene synthesis paradigm with an application to deep-learning-based thin cloud removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

JINGYU WANG is currently pursuing the Ph.D. degree with the Faculty of Information Science and Engineering, Ocean University of China. Her current research interests include remote sensing image colorization and computer vision.

CHENGLONG WANG is currently pursuing the Ph.D. degree with the Faculty of Information Science and Engineering, Ocean University of China. His research interests include multi-modal big data mining and multimedia content analysis.

QICHENG YANG is currently pursuing the M.S. degree with the Faculty of Information Science and Engineering, Ocean University of China. His research interest includes remote sensing segmentation.



CHENGYU ZHENG is currently pursuing the Ph.D. degree with the College of Information Science and Engineering, Ocean University of China. Her main research interest includes processing and analysis of remote sensing data.

JIE NIE (Member, IEEE) received the Ph.D. degree in computer science from the Ocean University of China, Qingdao, China, in 2011. From September 2009 to September 2010, she was a Visiting Scholar with the School of Electrical Engineering, University of Pittsburgh, Pittsburgh, PA, USA. From 2015 to 2017, she was a Postdoctoral Researcher with Tsinghua University, Beijing, China. She is currently working with the Ocean University of China. Her current research interests include social media and multimedia content analysis.

MINGXING JIANG received the M.S. degree from the Hefei University of Technology, Hefei, China, in 2013. He is currently pursuing the Ph.D. degree in communication and information systems with Shanghai University, Shanghai, China. He is also a Visiting Student with the Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, Hefei University of Technology. His research interests include multimedia quality assessment, affective computing, and machine learning.