JARS Journal of Applied Remote Sensing

RESEARCH PAPER

Improved SmapGAN remote sensing image map generation based on multi-head self-attention and carafe

Zhipeng Ding,[a] Ben Wang,[a,*] Shuifa Sun,[a,*] Yongheng Tang,[b,c] Ren Zhuang,[a] and Wenbo Liu[a]

[a]Hangzhou Normal University, School of Information Science and Technology, Hangzhou, China
[b]Three Gorges University, College of Computer and Information Technology, Yichang, China
[c]Three Gorges University, College of Economics and Management, Yichang, China

ABSTRACT. The changes in ground roads, buildings, and occurrences of natural disasters lead to mismatches between the actual ground conditions and existing maps. Through style transfer between real-time remote sensing images and maps, map content can be rapidly generated and updated. However, in existing methods for generating maps from remote sensing images based on SmapGAN, we first found that using ResBlock as the style conversion module fails to establish long-distance relationships between features. In addition, the small receptive field of convolution layers in ResBlock leads to poor global information capture, resulting in inferior image restoration during upsampling. Second, using transpose convolution as the upsampling method can result in the issue of blurred content in the generated maps. To address these problems, we propose corresponding improvements: on one hand, a style conversion module combining multi-headed self-attention (MHSA) with residual modules, named MHSA-ResBlock, is introduced to address the difficulty in capturing long-distance relationships between features when dealing with a large number of pixel features, and to better capture global information in images. On the other hand, an upsampling method combining transpose convolution with the CARAFE upsampling operator, named TC-Carafe, is proposed to tackle the issues of content loss and blurring associated with traditional transpose convolution upsampling. Furthermore, experimental results show that the MHSA-ResBlock establishes inter-pixel feature relationships and leverages the advantages of fine-grained upsampling operations with TC-Carafe, thereby utilizing inter-pixel feature relationships and neighborhood information to further improve the quality of map generation. Compared to SmapGAN, our research method has shown improvements of 0.6133 and 0.0042 in PSNR and SSIM, respectively. In addition, it has reduced RMSE by 0.72, outperforming SmapGAN in all metrics.

 [DOI: 10.1117/1.JRS.18.014526]

Keywords: map; remote sensing image; SmapGAN; multi-headed Self-attention; CARAFE; style transfer

Paper 230659G received Dec. 1, 2023; revised Feb. 19, 2024; accepted Feb. 29, 2024; published Mar. 20, 2024.

1 Introduction

Maps play a crucial and indispensable role in the daily lives and work of the general public. They not only serve the purpose of spatial positioning and navigation but also provide rich geographic

*Address all correspondence to Ben Wang, 20170056@hznu.edu.cn; Shuifa Sun, watersun@hznu.edu.cn

1931-3195/2024/$28.00 © 2024 SPIE

information and spatial data resources. Traditional methods of map creation typically rely on manual surveys and vehicle GPS trajectory data. However, these methods have inherent limitations in the process of map updating. First, traditional map-making methods require a significant amount of human resources and time, resulting in slow map update rates. Second, manual surveys may introduce human errors, leading to disparities between the map and the actual terrain.[1] In addition, vehicle GPS trajectory data may be subject to environmental and equipment limitations, making it challenging to accurately reflect real-world conditions. Given the frequent changes in ground structures and roads, as well as the occurrence of natural disasters, there is a mismatch between the actual ground conditions and existing maps. Therefore, there is a pressing need for a rapid and accurate method for map generation.

In recent years, style transfer techniques have garnered widespread attention in the field of deep learning. Gatys et al.[2,3] first introduced a style transfer method based on convolutional neural networks, applying the VGG[4] network to style transfer. By constructing the Gram matrix, style features of any image can be extracted. Isola[5] et al. proposed the pix2pix style transfer model based on conditional generative adversarial networks (cGAN),[6] which realizes one-to-one image style transfer through supervised training of paired images. Wang et al.[7] extended pix2pix to pix2pixHD, improving the resolution of generated images using multi-scale generators and discriminators. Zhu[8] et al. introduced the CycleGAN network model, which ensures the consistency and accuracy of transformations through a cycle-consistency loss function, without the need for paired training data. As a result, many researchers have utilized this technology to perform style transfer between remote sensing images and maps. Building upon the ideas of Gatys et al.,[2,3] Isola et al.,[5] and Zhu et al.,[8] Song et al.[9] introduced the MapGen-GAN method to achieve unsupervised style transfer, rapidly converting remote sensing images to maps. In addition, Chen et al.[10] proposed the SmapGAN semi-supervised model for achieving style transfer transformations between remote sensing images and maps within the same region. This remote sensing-based map creation method not only allows for the rapid updating of map content and reduced labor costs but also minimizes discrepancies between maps and the actual terrain, thereby enhancing map accuracy and reliability. However, these methods still have shortcomings in semantic understanding and contextual feature extraction of remote sensing images, resulting in unclear content and lack of details in the generated maps. Therefore, further research is needed to develop an efficient and accurate generation method to enhance the quality of style transfer between remote sensing images and maps, improving the generated maps in terms of detail and clarity.

In response to the issues mentioned above, including content blurriness and missing details in the generated output, we have made two specific optimizations to the existing model as follows:

(1) The paper introduces a style converter based on the fusion of multi-headed self-attention (MHSA)[11] mechanism and ResBlock[12] and uses MHSA mechanism to capture the long range dependencies between features. By leveraging self-attention to weightedly aggregate input features and with the assistance of multiple attention heads, we can more finely capture feature relationships among geographical pixels.

(2) In the upsampling stage of the generator part of SmapGAN network, this paper proposes to use fusion transposed convolution and lightweight Carafe[13] operator for upsampling operation. Taking full advantage of the large receptive field in Carafe operator, the upsampling kernel prediction and feature reconstruction are performed on information at the granularity of individual pixels.

The upcoming sections in this paper include: In Sec. 2, introduction of previous related work is provided. In Sec. 3, the approach suggested in this study is presented. In Sec. 4, experimental results will be presented and analyzed. In Sec. 5, the advantages and limitations of this work are discussed. In Sec. 6, conclusions are drawn and an outlook for future work is provided.

2 Related Work

In this section, we will first introduce traditional map drawing methods. Next, we will present the methods of map drawing using semantic segmentation techniques. Finally, we will discuss research related to style transfer from remote sensing images to maps.

2.1 Traditional Mapping Methods

Hand-drawn maps are one of the most traditional methods of map creation. Cartographers use drawing tools and paints to manually draw maps, representing geographical information based on geographic features and labeling. This method requires high drawing skills and experience, and the production process is time-consuming and may involve subjectivity.[14]

In texture mapping drawing, cartographers use geographic data and aerial photographs to create maps by pasting textures of geographical features using specialized software tools.[15] This method is relatively simpler compared to hand-drawn maps, saving time, but it still requires manual operation and judgment.

2.2 Deep Learning-Based Map Making Method

With the development of computer vision and the technology of deep learning, map generation methods based on deep learning are gradually being applied, making map production more efficient. Currently, deep learning-based map generation methods can be mainly categorized into semantic segmentation-based methods and generative adversarial network (GAN)-based methods.

2.2.1 *Semantic segmentation-based approach to map making*

Map creation methods based on semantic segmentation techniques involve performing semantic segmentation on remote sensing images to recognize and label different geographic features and objects so as to make maps. UNet[16] combines the features of the encoder with those of the decoder using skip connections to preserve richer semantic information. FCN[17] uses convolutional layers instead of fully connected layers to achieve pixel-level segmentation. It also utilizes feature maps of different scales to handle multiscale semantic information. DeepLabv3+[18] introduced dilated convolutions into the encoder, employed Xception as the backbone feature extraction network, and incorporated depthwise separable convolutions into the atrous spatial pyramid pooling. Mask R-CNN[19] adds a branch to faster R-CNN,[20] which is used to generate an accurate mask for each object. It addresses pixel misalignment issues by using ROI align instead of ROI pooling. In addition, there are several variant models aimed at further enhancing segmentation performance. RoadNet,[21] AdaLSN,[22] SPIN Road Mapper,[23] X. Li.,[24] and CoANet[25] are models designed to perform road segmentation on remote sensing images. However, most of the above works can only segment and extract the same kind of objects, whereas the generation of maps involves more complex elements, such as terrain, buildings, and water bodies. Therefore, simply using semantic segmentation model is difficult to capture the maps of the whole structure and style.

2.2.2 *Map generation method based on generative adversarial networks*

Many scholars have drawn inspiration from GANs[26] and conducted research on style transfer techniques between remote sensing images and maps using GANs.

Liu et al.[27] proposed a framework called UNIT, which is an unsupervised image-to-image translation based on GANs and variational autoencoders and demonstrated that sharing latent space constraints include cycle-consistency constraints. Isola et al.[5] proposed the pix2pix style transfer model based on cGAN, achieving one-to-one image style transfer through supervised training of paired images. Wang et al.[7] proposed pix2pixHD based on pix2pix to generate high-resolution images by adopting a coarse-to-fine feature extraction strategy and discriminating on three different image scales. Zhu et al.[8] proposed CycleGAN network model, which ensures the consistency and accuracy of the transformation through the cycle-consistency loss function without the need for paired training data. Ganguli et al.[28] proposed GeoGAN, which added reconstruction loss and style transfer loss on top of the GAN loss. Song et al.[9] proposed MapGen-GAN, which enhances network depth by designing a novel generator called BRB-Unet. In addition, they incorporated cycle-consistency and geometric-consistency as part of the loss functions. Chen et al.[10] proposed SmapGAN, which designs a semi-supervised GAN for style transfer between regional remote sensing images and maps. They devised gradient loss and structural loss to optimize the generation process, aiming for maps that embody topological

relationships more effectively. Fu et al.[29] proposed utilizing Deeplabv3+ for initial feature extraction. Subsequently, the preliminarily extracted features were passed into a creative module, where they were fused with the original input image for in-depth feature extraction and high-quality map generation. Zhan et al.[30] proposed a new metric MoNCE in order to solve the existence of image blur and false shadows, which incorporates image contrast to learn a calibration metric for perceiving distances between multi-faceted images. Song et al.[31] proposed Semi-MapGen and designed extension loss and channel loss to improve the accuracy of map generation through knowledge extension learning strategies. Solano-Carrillo et al.[32] proposed the fully supervised model ATME, which enhanced the connection between the generator and the discriminator by focusing on the average entropy of the discriminator. They efficiently generated maps by integrating the high-quality generation ability of DMs and the sampling strength of GANs. Xu et al.[33] proposed SAM-GAN. SAM-GAN employs a cGAN as the generator and utilizes a multi-scale discriminator. It incorporates style loss, topological consistency loss, and the SeBlock attention mechanism to enhance the effectiveness of map generation. However, the most significant challenge at present is the lack of a thorough understanding of the semantics of remote sensing images and consideration of contextual features. Due to the rich semantic information present in remote sensing images, relying solely on simple feature mapping may not fully capture this semantic information. Consequently, the generated maps may exhibit some unreasonable content features, such as unclear content and missing details.

Currently existing models, such as the mentioned pix2pix,[5] pix2pixHD,[7] GeoGAN,[28] ATME,[32] SAM-GAN,[33] and the model proposed by Fu et al.[29] all fall under supervised models, requiring dependency on paired data or even semantic labels. However, in many practical scenarios, obtaining accurate paired data and labels is challenging. On the contrary, models such as UNIT,[27] CycleGAN,[8] and MapGen-GAN[9] adopt unsupervised approaches, capable of handling unpaired data. However, due to the lack of supervised training with paired data, their actual generation performance is not entirely satisfactory. Models such as SmapGAN[10] and Semi-MapGen[31] belong to semi-supervised models, combining the advantages of both supervised and unsupervised models. Despite having certain advantages, the style converter based on ResBlock primarily uses local receptive fields to capture spatial local relationships in the input data, leading to insufficient handling of long-distance dependency relationships. In addition, traditional transposed convolutions are susceptible to blurriness and information loss during the upsampling process, influenced by the local receptive fields of the convolutional kernels and padding. Therefore, this paper proposes a semi-supervised model that introduces MHSA into the style converter and combines traditional convolution with Carafe upsampling to solve the gaps and shortcomings of previous models.

3 Methods

In Fig. 1, G represents the generator, D represents the discriminator. X_{real} represents real remote sensing images, Y_{real} represents real maps. X_{fake} represents fake remote sensing images generated by Y_{real}, Y_{fake} represents fake maps generated by X_{real}. X'_{fake} represents fake remote sensing images generated by X_{real}, and Y'_{fake} represents fake maps generated by Y_{real}.

3.1 Map Generation Network Based on Multi-Headed Self-Attention and Carafe Upsampling

The generator of SmapGAN[10] utilizes an encoder-style converter-decoder structure. The encoder consists of two down-sampling layers with convolutional kernels of size 3×3 and a stride of 2. The style converter structure, as shown in Fig. 2, is composed of nine ResBlocks[12] with convolutional kernels of size 3×3. The decoder comprises two up-sampling layers with convolutional kernels of size 3×3 and a stride of 2.

However, this model has the following shortcomings: (1) It lacks the ability to capture long-range dependencies, resulting in generated images lacking coherence. (2) Lacking the ability to generate local details, the generated image may be missing fine textures and structures. This study, inspired by Srinivas et al.,[34] proposes a style converter that combines the MHSA with ResBlock. First, in the SmapGAN network style converter part, MHSA is used to better capture the key features in the map. Second, in the upsampling stage, the upsampling method TC-Carafe,

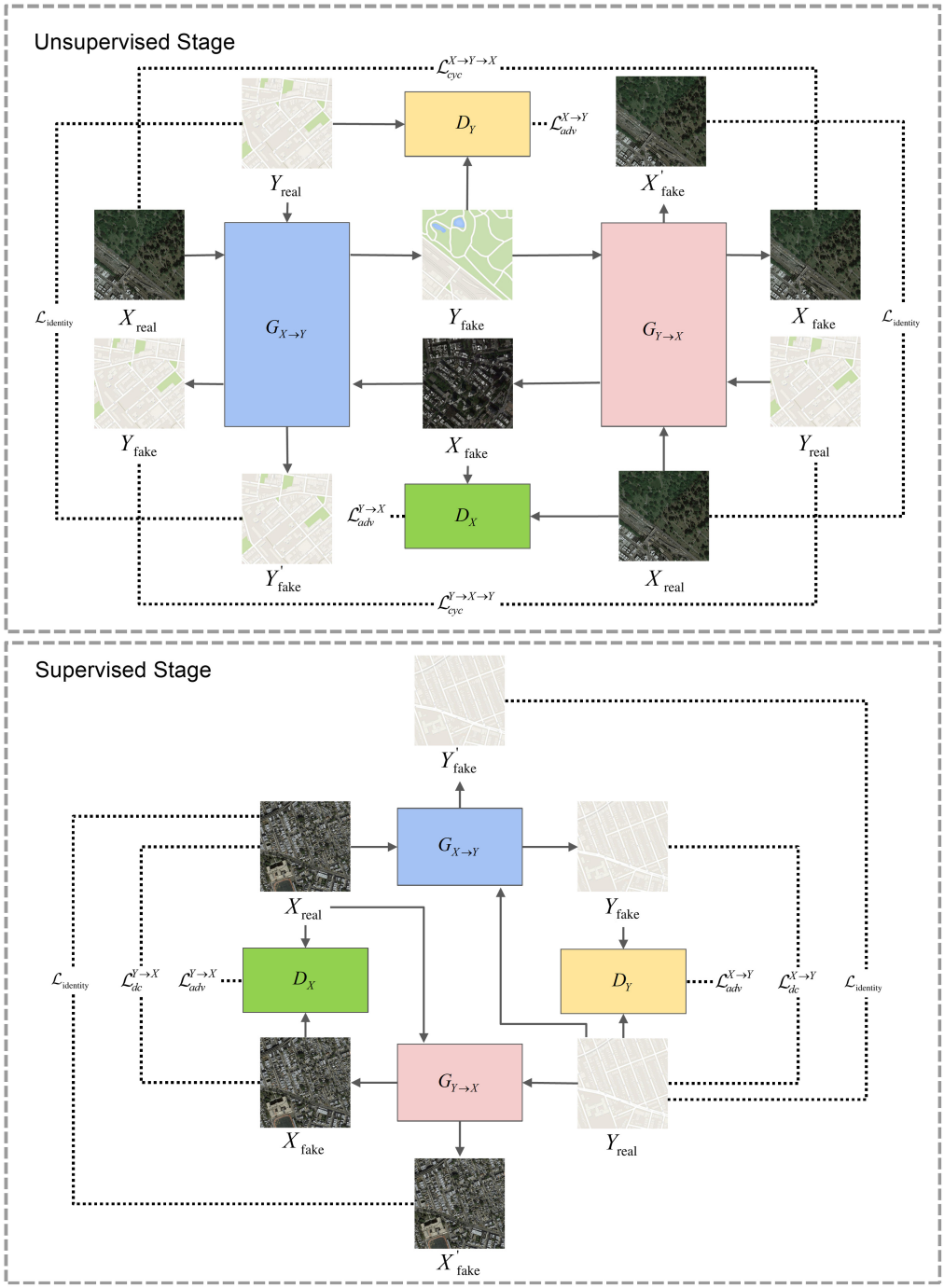

Fig. 1 Overall pipeline of our proposed semi-supervised map generation model. Each training iteration comprises an unsupervised phase and a supervised phase. The unsupervised phase utilizes unpaired data, whereas the supervised phase employs paired data. The dashed lines indicate the loss functions used, and arrows represent the workflow.

which fuses transposed convolution and CARAFE[13] operator, is used to enhance the details and reality of the generated image. The improved generator structure is shown in Fig. 3.

The generator comprises an image encoder, a style converter composed of ResBlocks and the MHSA mechanism, and an image decoder. The discriminator follows the PatchGANs[8] architecture. The input of the network is a $C \times H \times W$ remote sensing image and reduces the size of the feature maps and eliminates redundant information using a 7×7 convolutional layer followed by two 3×3 downsampling layers. Pass the extracted feature information to a style

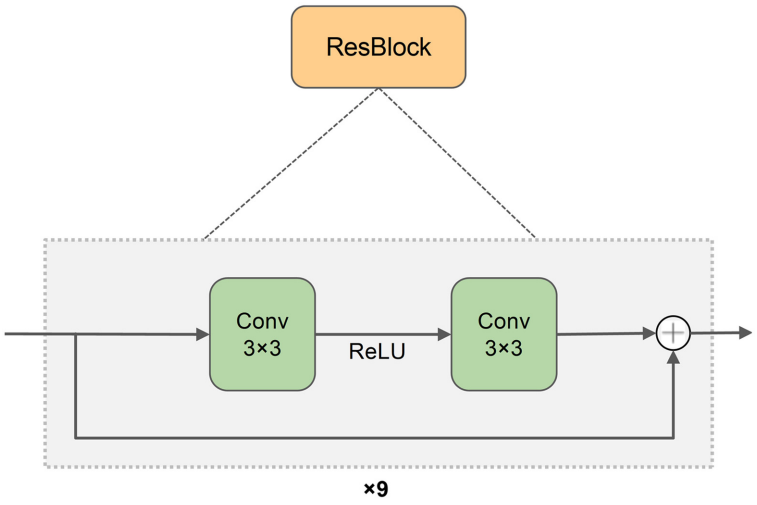

Fig. 2 Style converter structure based on ResBlock.

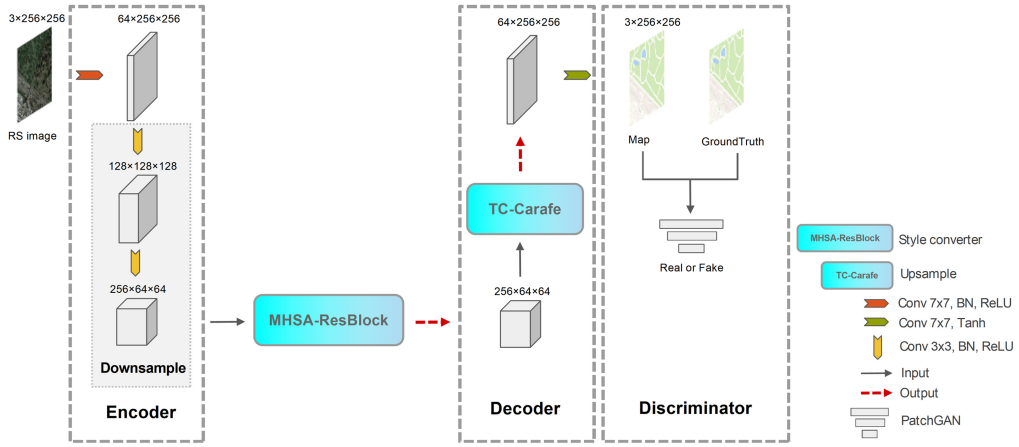

Fig. 3 Improved model structure after style converter and TC-Carafe. (Take the input remote sensing image as an example).

converter composed of 10 ResBlock and MHSA, transforming the feature vector of remote sensing images into the feature vector of maps. The module mainly captures the map features through MHSA and establishes the long-distance feature relationship. The restructured image features are subsequently transmitted to the decoder. First, they are fed into a traditional convolutional layer with a kernel size of 3×3 for initial image feature reconstruction. Next, the reconstructed image features are passed through the Carafe upsampling operator for fine-grained image feature reconstruction. Finally, a 7×7 convolution operation is performed to output a high-quality map.

3.2 Style Converter Based on Multi-Headed Self-Attention and ResBlock

This paper proposes the use of a 10-layer ResBlock as the style converter, and the last ResBlock is changed into a self-attention mechanism module with a convolution kernel size of 1×1 and four heads, which is used to capture the long-term dependency between pixels in the image. As shown in Fig. 4, by introducing the MHSA in parallel across multiple subspaces to learn feature representations, the model's non-linearity capability is increased. First, the downsampled features Feature_R are used as the input to the style converter. As the mapped features pass through the final layer of ResBlock, the channel count of the input features is reduced to AttentionIn (the channel count of the MHSA input), $\text{AttentionIn} = \text{DimOut}/\text{Factor}$. Next, the feature maps are fed into the MHSA layer, where each attention head consists of a set of query (Q), key (K), and

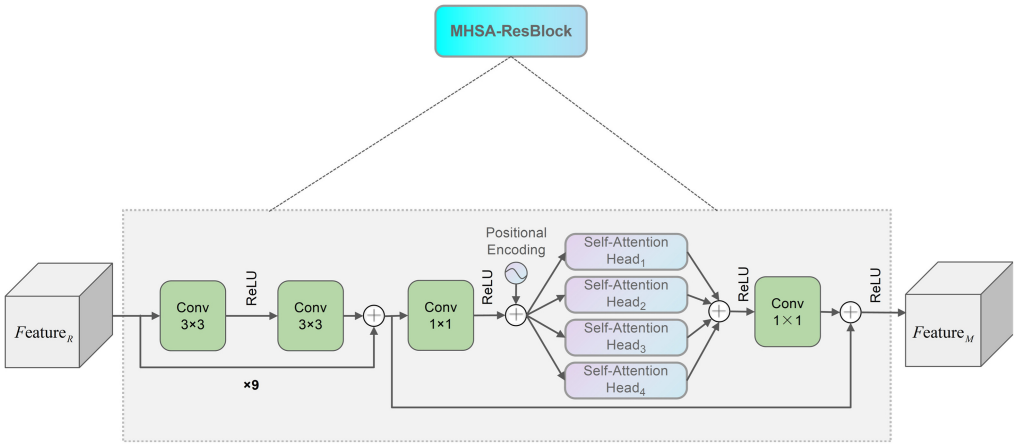

Fig. 4 MHSA-ResBlock style converter structure diagram.

value (V) vectors. Input features are mapped to Q, K, and V using one-dimensional convolutional layers. Similarity between Q and K is computed, and positional encoding is used to provide positional relationship information. The attention weight matrix is obtained by scaling with a factor of $\sqrt{d_k}$ and applying softmax operation, followed by multiplication with weight matrix V to obtain the weighted values. Finally, the results from all heads are concatenated, and a linear transformation is applied by multiplying with the weight matrix W^O to obtain the output of the MHSA. MHSA can be calculated[11] as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

$$\text{Head}_i = \text{Attention}(QW_i^q, KW_i^\kappa, VW_i^v), \tag{2}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{Head}_1, \text{Head}_2)W^O. \tag{3}$$

In this study, the number of heads is set to 4, with each self-attention head having a dimension of 64. The number of channels DimOut of the output feature is set to 256, and the projection factor is set to 4. Absolute positional encoding is used in the attention layer to provide direct positional information, facilitating the model's learning of relevant features. After passing through the MHSA layer, a 1×1 convolutional layer is applied to increase the channel count of AttentionOut (the feature channels from the MHSA layer) to DimOut. The features obtained from the 1×1 convolutional layer are then added to the original input features and passed through a ReLU activation function. Finally, this results in fine-grained feature extraction, denoted as Feature$_M$. The purpose of this skip connection is to merge the original input features with those processed through the MHSA layer and convolutional layer, preserving the information from the original input and introducing the features deeply extracted by the MHSA layer, thus enhancing the model's representational capacity. This design contributes to further optimizing the quality of generated images and reducing information loss.

3.3 Upsampling Method Combining Transposed Convolution and the Carafe Operator

In the upsampling stage, as shown in Fig. 5, this paper proposes a combination of traditional transposed convolution and the Carafe method for upsampling operations. First, the feature map Feature$_M$ is upsampled using a conventional 3×3 transposed convolution as the first layer of upsampling. The output features are denoted as F_1. In the second layer, the Carafe upsampling operator is employed to enhance image generated details and clarity. Carafe upsampling offers advantages, such as a large receptive field, lightweight design, and fast computation speed. The Carafe module consists of three convolution layers and is divided into two parts: upsampling kernel prediction and feature recombination. In the upsampling kernel prediction module, the input data are a feature map of $C \times H \times W$ size. First, a 1×1 convolution layer is used to

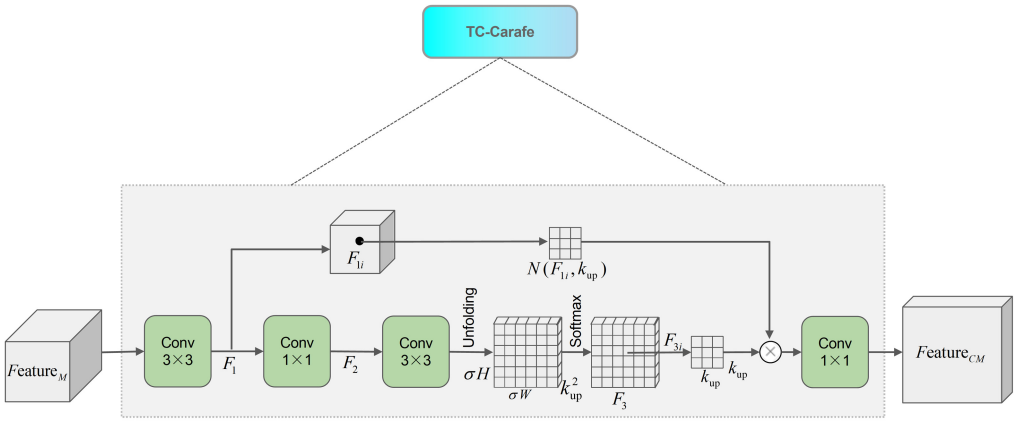

Fig. 5 TC-Carafe structure diagram.

reduce the channel number from C to C_n. The features after compression are denoted as F_2. Second, a 3×3 convolution kernel is used for upsampling kernel prediction, where the input is $H \times W \times C_n$ and the output is $H \times W \times \sigma^2 k_{\text{up}}^2$. The parameters are set as $\sigma = 2$, $k_{\text{up}} = 3$. This operation increases the encoder's receptive field and effectively utilizes contextual information over a larger area. Simultaneously, the channel dimension is unfolded in the spatial dimensions, resulting in an upsample kernel of shape $\sigma H \times \sigma W \times k_{up}^2$. Subsequently, spatial normalization is applied to the reconfigured kernels of size $k_{\text{up}} \times k_{\text{up}}$. It is expressed by the equation:

$$F_3 = \text{KPM}(\text{softmax}(\text{Unfolding}(H \times W \times \sigma^2 k_{\text{up}}^2))) \ , \tag{4}$$

where KPM represents performing the upsampled kernel prediction operation, Unfolding represents performing the spatial expansion on it, softmax represents performing the normalization operation on it, and F_3 represents the result of the upsampled kernel prediction.

In the feature recombination module, the dot product operation is performed between the upsampled prediction kernel and the $k_{\text{up}} \times k_{\text{up}}$ region centered at a feature point of the input feature map to achieve feature recombination. Finally, the convolution layer with 1×1 convolution kernel is used to compress the channel and output the Feature$_{\text{CM}}$. It is expressed by the equation:

$$\text{Feature}_{\text{CM}} = \text{Compress}(\text{CARM}(F_{3i} \otimes N(F_{1i}, k_{\text{up}}))), \tag{5}$$

where CARM represents the feature recombination operation, $N(F_{1i}, k_{\text{up}})$ represents the region of size $k_{\text{up}} \times k_{\text{up}}$ centered at the F_{1i} feature point in the input feature map, and F_{3i} represents the upsampling kernel of this point predicted by the upsampling kernel prediction module.

3.4 Loss Function

The paper adopts topological consistency loss,[10] content loss, adversarial loss,[6] and identity loss[35] as the loss functions. Among them, content loss is divided into cycle consistency loss[8] and direct loss.[5] The model in this paper is a semi-supervised model, meaning it has both unsupervised and supervised stages. In the unsupervised stage, due to the use of unpaired data, the network needs to ensure that the generated images retain the characteristics of the original images, hence employing cycle consistency loss as the content loss. In contrast, in the supervised stage, paired data are used, so direct loss is used as the content loss. Except for the difference in content loss between the two stages, the rest of the loss functions remain unchanged in both stages.

(1) Topological consistency loss:[10] Used to ensure the correct topological relationships of $G_{X \to Y}$ (remote sensing to map). Here, L_{grall} represents the image gradient L1 loss, and L_{grastr} represents the image gradient structural loss. Its equation is as follows:

$$L_{\text{Top}}^{X\to Y} = L_{\text{gral1}} + L_{\text{grastr}} = E_{x\sim P_X}[\|G(G_{X\to Y}(x)) - G(y)\|_1] + E_{x\sim P_X}\left[2 - \frac{1}{N-1}\sum_{j=0}^{N-2}\frac{|\sigma_{G_j(y)G_j(G_{X\to Y}(x))}| + C_1}{\sigma_{G_j(y)}\sigma_{G_j(G_{X\to Y}(x))} + C_1} - \frac{1}{M-1}\sum_{i=0}^{M-2}\frac{|\sigma_{G_i(y)G_i(G_{X\to Y}(x))}| + C_2}{\sigma_{G_i(y)}\sigma_{G_i(G_{X\to Y}(x))} + C_2}\right]. \tag{6}$$

(2) Content loss: Aimed at ensuring the similarity in content between the generated map and ground truth, where $L_{\text{cyc}}^{X\to Y\to X}$ and $L_{\text{cyc}}^{Y\to X\to Y}$ are cycle losses,[8] and $L_{dc}^{X\to Y}$ and $L_{dc}^{Y\to X}$ are the direct losses. In the unsupervised stage, cyclic loss is employed. In the supervised stage, direct loss is used. Its equation is as follows:

$$L_{\text{cyc}}^{X\to Y\to X} = \lambda_{L1u}L_{L1u} = \lambda_{L1u}E_{x\sim P_X}[\|G_{Y\to X}(G_{X\to Y}(x) - x)\|_1], \tag{7}$$

where $L_{\text{cyc}}^{X\to Y\to X}$ represents the cycle loss from the remote sensing image. This cycle loss is calculated using $L1$ loss to measure the differences between pixels, ensuring content cyclic consistency between the generated map and the remote sensing image:

$$L_{\text{cyc}}^{Y\to X\to Y} = \lambda_{L1u}L_{L1u} + L_{\text{Top}}^{X\to Y} = \lambda_{L1u}E_{y\sim P_Y}[\|G_{X\to Y}(G_{Y\to X}(y)) - y\|_1] + L_{\text{Top}}^{X\to Y}, \tag{8}$$

where $L_{\text{cyc}}^{Y\to X\to Y}$ represents the cycle loss from the map. The introduction of the topological consistency loss $L_{\text{Top}}^{X\to Y}$ helps maintain the cyclic consistency of the generated image's topological structure with that of the target image:

$$L_{dc}^{Y\to X} = \lambda_{L1}L_{L1} = \lambda_{L1}E_{y\sim P_Y}[\|G_{Y\to X}(y) - x\|_1], \tag{9}$$

where $L_{dc}^{Y\to X}$ represents the direct loss from the map to the remote sensing image. It ensures content consistency through the use of the $L1$ loss function:

$$L_{dc}^{X\to Y} = \lambda_{L1}L_{L1} + L_{\text{Top}}^{X\to Y} = \lambda_{L1}E_{x\sim P_X}[\|G_{X\to Y}(x) - y\|_1] + L_{\text{Top}}^{X\to Y}, \tag{10}$$

where $L_{dc}^{X\to Y}$ represents the direct loss from the remote sensing image to the map. It maintains topological consistency between the generated map and the remote sensing image through the $L_{\text{Top}}^{X\to Y}$.

(3) Adversarial loss:[6] It assesses the disparity between generated images and real images using a discriminator. The generator G aims to minimize the loss function value, whereas the discriminator D aims to maximize it. The equation is as follows:

$$L_{\text{adv}}^{X\to Y} = E_{y\sim P_Y}[\log\ D_Y(y)] + E_{x\sim P_X}[\log(1 - D_Y(G_{X\to Y}(x)))], \tag{11}$$

where $L_{\text{adv}}^{X\to Y}$ represents the adversarial loss from remote sensing images to maps. $G_{X\to Y}$ is the generator for remote sensing images to maps, and the generated images are input to D_Y for discrimination:

$$L_{\text{adv}}^{Y\to X} = E_{x\sim P_X}[\log\ D_X(x)] + E_{y\sim P_Y}[\log(1 - D_X(G_{Y\to X}(y)))], \tag{12}$$

where $L_{\text{adv}}^{Y\to X}$ represents the adversarial loss from maps to remote sensing images. $G_{Y\to X}$ is the generator for remote sensing images to maps, and the generated images are input to D_X for discrimination.

(4) Identity loss:[35] Used to ensure consistency between the transformed image and the original image. For example, when passing a map into generator $G_{X\to Y}$, the generated map should ideally remain as consistent as possible with the input map, preserving the map's content and colors. The equation for this is

$$L_{\text{identity}} = |G_{Y\to X}(x) - x| + |G_{X\to Y}(y) - y|. \tag{13}$$

4 Experiment and Result Analysis

4.1 Datasets

As shown in Fig. 6, the dataset used for the experiments consists of 2194 remote sensing images selected from Google Maps, along with their corresponding map data.[8] The pixel size of the

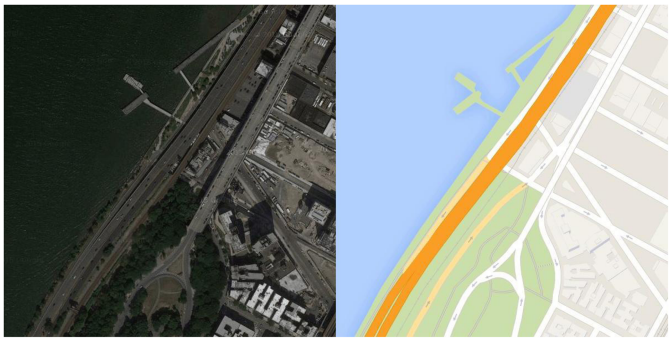

Fig. 6 Remote sensing images and map samples.

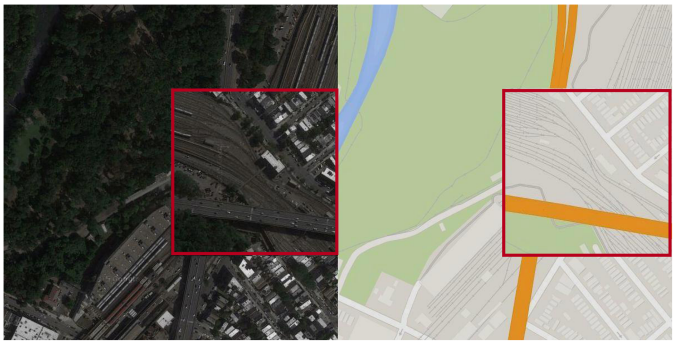

Fig. 7 Examples of matting and rotating 90 deg on part of the dataset.

images is 256×256. The dataset is divided into 829 paired images for supervised training and 829 unpaired images for unsupervised training. Both the validation and test sets consist of 268 images each.

Due to the fully supervised training approach in ATME,[32] it uses paired matching samples for training. Since ATME is a supervised model, paired data are used in the training phase. The model presented in this paper is a semi-supervised model, meaning that unpaired data are employed during the unsupervised phase, whereas paired data are used during the supervised phase. This is unfair for semi-supervised models. To ensure experimental fairness, this paper adopts the training set processing method for both supervised and semi-supervised models proposed by Song et al.[31] Data perturbation is applied to a portion of the training set data intended for the ATME supervised model. We conducted training on the ATME model using paired data with 10% data perturbation and 20% data perturbation, respectively. As shown in Fig. 7, we randomly selected 10% and 20% of training samples in the training set for data perturbation, including matting and rotation of 90 deg.

4.2 Experimental Setup

The GPU used for all experiments in this study is the NVIDIA RTX 3060. During training, the batch size is set to 1, and the adam optimizer parameters are set to $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The initial learning rate is 0.0002, which remains constant from epoch 1 to 100. From epoch 101 to 200, the learning rate gradually decays to 0. The training for all models comprises a total of 200 epochs.

4.3 Evaluation Metrics

(1) Peak signal-to-noise ratio (PSNR):[36] with higher PSNR values indicating better image quality. Here, MAX_I represents the maximum pixel value of the image, and the unit is in decibels (dB). The equation for calculating PSNR is

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}_\mathrm{I}^2}{\mathrm{MSE}}\right) = 20 \cdot \log_{10}\left(\frac{\mathrm{MAX}_\mathrm{I}}{\sqrt{\mathrm{MSE}}}\right). \tag{14}$$

(2) Structural similarity index (SSIM):[37] SSIM primarily assesses the similarity between two images in terms of structure, brightness, and contrast. When one of the images is an undistorted reference image, and the other is a distorted version, SSIM serves as a quality metric for the distorted image. The SSIM ranges from 0 to 1, where a value closer to 1 indicates a higher similarity between two images. The SSIM equation is as follows:

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\delta_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\delta_x^2+\delta_y^2+C_2)}, \tag{15}$$

where μ_x represents the average of x, μ_y represents the average of y, δ_x^2 represents the variance of x, δ_y^2 represents the variance of y, and δ_{xy} represents the covariance between x and y. C_1 and C_2 are constant values used to stabilize the computation.

(3) Root-mean-square error (RMSE):[38] First, calculate the squared difference between the actual and predicted values, then sum them up, take the average, and finally, calculate the square root. The range of RMSE values is from 0 to $+\infty$, and a lower RMSE indicates greater similarity between the images. RMSE is calculated as follows:

$$\mathrm{RMSE}=\sqrt{\mathrm{mse}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}, \tag{16}$$

where n represents the number of samples, y_i represents the ground truth values, and $\hat{y}_i$ represents the predicted values.

4.4 Performance and Comparison

The experiment selected SmapGAN,[10] CycleGAN,[8] Pix2pixHD,[7] CUT_MoNCE,[30] and ATME[32] comparing with the method, and use the PSNR, SSIM and RMSE evaluation metric for testing and evaluation.

As shown in Table 1, in the comparative experiments, we compared our model with other models using three image quality evaluation metrics. Compared to the baseline SmapGAN, our method exhibits improvements in PSNR, SSIM, and RMSE metrics. Specifically, there is an improvement of 0.6133 in PSNR, 0.0042 in SSIM, and a reduction of 0.72 in RMSE. Bold indicates the highest score, italics indicates the second-best, and bold italics indicates the upper bound fully supervised model ATME. ATME-10% represents 10% perturbed samples included in the training data. ATME-20% denotes 20% perturbed samples included in the training data. ATME-full signifies the usage of the original unperturbed paired data for fully supervised training. Since ATME utilizes a fully supervised approach, it is understandable that our semi-supervised approach may be slightly lacking in comparison. We treat ATME as a performance upper bound. Despite our method showing a slight drawback in PSNR and RMSE metrics, it outperforms the ATME-full model by 0.0046 in SSIM. After perturbing a portion of the training data, our method significantly surpasses ATME-10% and ATME-20% across all metrics.

Table 1 The comparison results of our approach with SmapGAN, ATME-full, ATME-10%, ATME-20%, Pix2pixHD, CUT_MoNCE, and CycleGAN on objective evaluation metrics.

Model	PSNR↑	SSIM↑	RMSE↓
CycleGAN	24.5271	0.8157	18.3162
CUT_MoNCE	24.0537	0.7880	18.8089
Pix2pixHD	27.1008	0.8499	13.0374
SmapGAN	27.5014	*0.8742*	12.4684
ATME-10%	*27.7678*	0.8644	*12.1932*
ATME-20%	27.4103	0.8594	12.6653
(Ours)	**28.1147**	**0.8784**	**11.7484**
ATME-full	***28.3162***	***0.8738***	***11.2939***

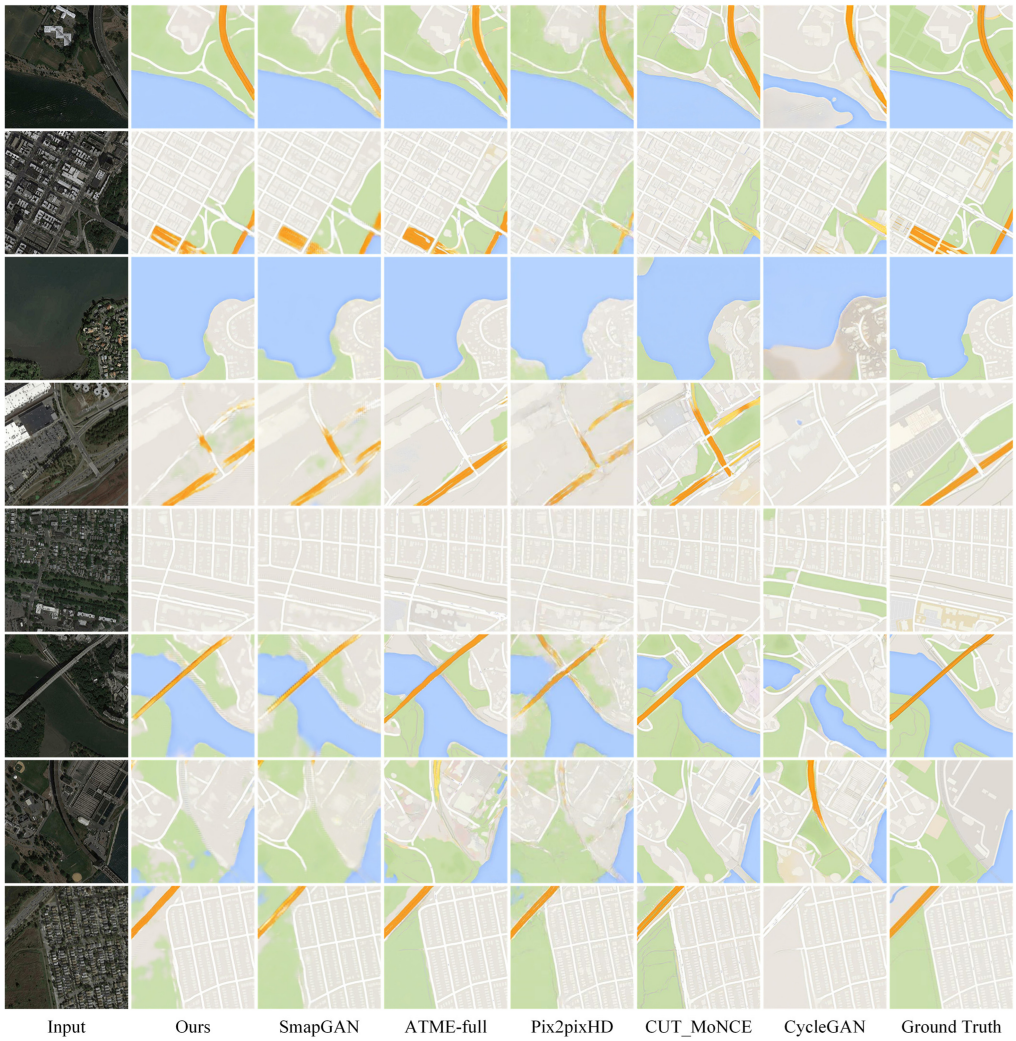

Fig. 8 Comparison of generation effect between the proposed method and SmapGAN, ATME-full, Pix2pixHD, CUT_MoNCE, and CycleGAN.

In Fig. 8, the first column shows the input remote sensing images, the second column presents the generated map images by our model. The following columns 3 to 7 display map images generated by various comparative models. Column 8 showcases the ground truth maps corresponding to the remote sensing images.

From the perspective of content details, the content generated by SmapGAN is relatively consistent with the original image, but there are several disruptions in major roads and minor issues of inadequate content generation. Compared to SmapGAN, ATME-full has made progress in generating major roads, but there are still deficiencies in the details. For example, in the generated map image from the first remote sensing image, the depiction of a three-way circular road is not complete, and there is a minor issue of over-generation. CUT_MoNCE and CycleGAN suffer from over-generation issues. For example, in the sixth map generated from the remote sensing image, land is incorrectly generated in the sea area, and in the seventh map, multiple roads are wrongly generated in the gray region.

In generating special roads (highlighted in orange), CUT_MoNCE and ATME-full produce highly saturated colors for special roads but still have some defects. For instance, in the map image corresponding to the second remote sensing image generated by ATME-full, the orange roads are not clearly separated but appear merged. In the sixth image, the generated orange roads exhibit bending issues. In the map corresponding to the fourth remote sensing image generated by CUT_MoNCE, normal roads are mistakenly identified as special roads, resulting in erroneous generation. CycleGAN performs the worst in generating special roads, failing to correctly

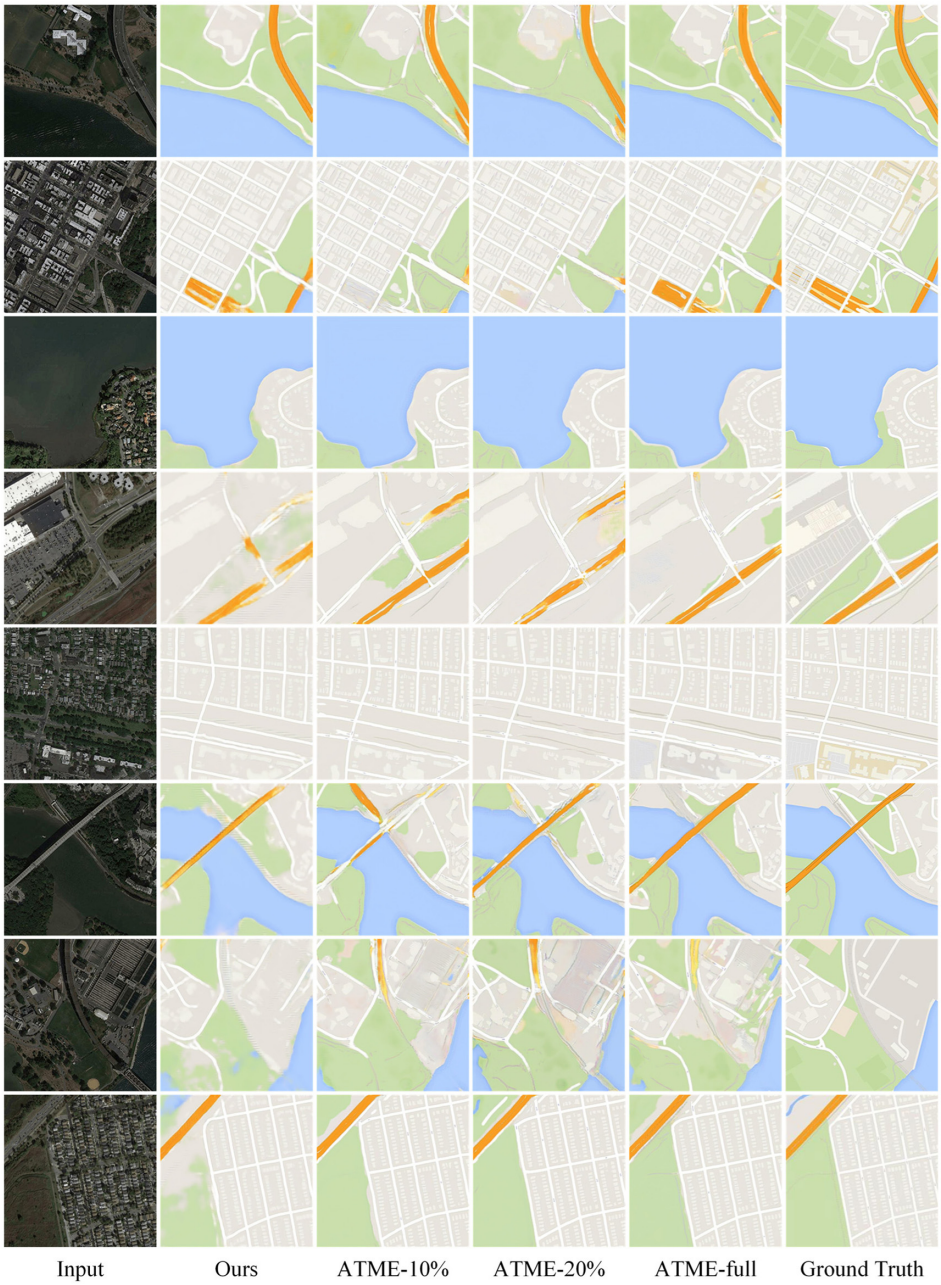

Fig. 9 Comparison of the proposed method with ATME-10%, ATME-20%, and ATME-full in terms of generation effect.

identify multiple special roads. SmapGAN and Pix2pixHD both have the problem of poor special road generation effect.

As shown in Fig. 9, when applying 10% and 20% data perturbations to the training set of the ATME model, our model's generation quality significantly surpasses that of ATME-10% and ATME-20%. While our approach slightly lags behind ATME-full in terms of clarity and special road generation, the generated content details and structure in our model are superior to ATME-full.

Compared to other models, this paper introduces a style converter based on ResBlock and a multi-head self-attention mechanism. In addition, it adopts a combination of traditional transpose convolution and the CARAFE operator during the upsampling stage. This approach enhances the learning and processing capabilities of structural features, aiming to avoid pronounced deformations and content loss. It maximally preserves the spatial details of map roads and provides a visual effect superior to the baseline.

4.5 Ablation Study

To verify the effectiveness of the proposed method, we use the proposed style converter MHSA-ResBlock and the TC-Carafe module were used separately to evaluate their impact on the results when combined. Figure 10 and Table 2 present the generation results and metric evaluations from the ablation study.

According to the results in Table 2, first, only using MHSA-ResBlcok improves the PSNR metric by 0.2885, improves the SSIM metric by 0.0029, and reduces the RMSE metric by 0.4346. Second, when only TC-Carafe was added, there was an improvement of 0.3633 in the PSNR metric, an improvement of 0.0016 in the SSIM metric, and a reduction of 0.0901 in the RMSE metric. Finally, when both were combined, there was an improvement of 0.6133 in the PSNR metric, an improvement of 0.0042 in the SSIM metric, and a reduction of 0.72 in the RMSE metric. Based on the results from Fig. 10 and Table 2, the generation effect of MHSA-ResBlcok alone is limited, which indicates that although the MHSA mechanism can better capture feature information and establish long-distance dependencies, the traditional deconvolution layer usually does not make full use of global information because the operation is based on a small local receptive field. Thus, the effect of map generation is limited. Similarly, the standalone introduction of TC-Carafe did not significantly improve the generation results. Due to the lack of accurate feature information, TC-Carafe cannot fully leverage its advantages in

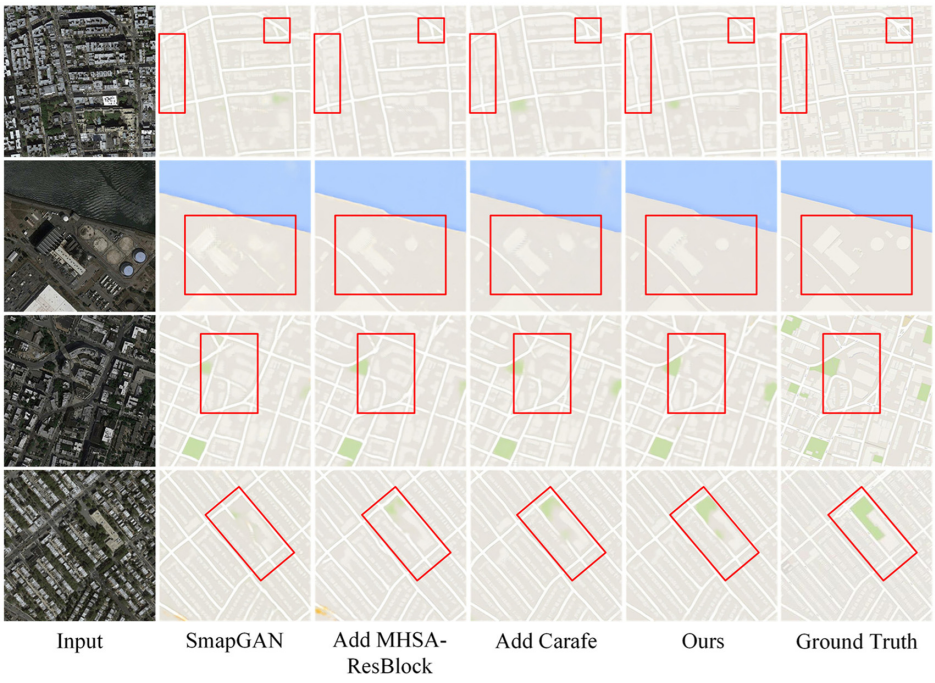

Fig. 10 The map generation results from ablation study are shown in Fig. 9. "Add MHSA-ResBlock" represents the generation result after using MHSA-ResBlock alone as the style converter. "Add TC-Carafe" represents the generation result after using TC-Carafe alone. "Ours" represents the generation results when both MHSA-ResBlock and TC-Carafe are combined.

Table 2 By ablation study of four evaluation index score.

SmapGAN	MHSA-ResBlock	TC-Carafe	PSNR↑	SSIM↑	RMSE↓
✓	—	—	27.5014	0.8742	12.4684
✓	✓	—	27.7899	0.8771	12.0338
✓	—	✓	27.8647	0.8758	12.3783
✓	✓	✓	**28.1147**	**0.8784**	**11.7484**

fine-grained upsampling. However, model performance improvement should consider the comprehensive impact of multiple factors. To further leverage the strengths of each module, this paper combines the MHSA-ResBlcok with TC-Carafe. Through the MHSA mechanism, correlations between each pixel and other pixels are computed, allowing different location features to interact and integrate, providing TC-Carafe with more accurate feature information. At the same time, TC-Carafe, through upsampling kernel prediction and feature recombination, can better preserve and restore image details, making the generated maps clearer and more complete. Therefore, by combining the MHSA-ResBlcok and TC-Carafe, their weaknesses can complement each other, resulting in a better performance improvement.

5 Discussion

This paper improves the generation results significantly beyond the baseline model by refining the style converter and upsampling methods in existing models. However, there are still some shortcomings that need further improvement, especially in the generation of greenbelt, where issues, such as blurriness and omissions, may occur. This indicates that when generating greenbelt, the model requires more refined handling to enhance the clarity and accuracy of the generated images.

According to the model parameter data in Table 3, it is shown that the model parameters of our method are slightly higher than the baseline model SmapGAN. However, as observed in Fig. 8, our method provides superior visual results. Compared to the fully supervised ATME-full, although our method falls slightly short in PSNR and RMSE metrics, it surpasses the ATME-full model by 0.0046 in SSIM. Due to the inclusion of MHSA in the MHSA-ResBlock proposed in this paper, MHSA requires computing the relationships between all input features at each layer, incurring significantly higher computational complexity compared to the convolution operations in ResBlock. Therefore, there has too much time on the model training in this paper. It is worth noting that our model has significantly fewer parameters compared to ATME. ATME employs fully supervised training, utilizing paired samples extensively for training, thus we consider ATME fully supervised training as the performance upper bound. In contrast, our training approach only utilizes 50% of the paired samples from the dataset. Therefore, our model's slight underperformance in PSNR and RMSE compared to the fully supervised ATME-full is acceptable.

However, by perturbing a portion of the training data, our model's generation performance is significantly superior to ATME-10% and ATME-20%. The experiments demonstrate that after applying perturbations, such as introducing matting and rotating parts of the image by 90 deg, to the training data, ATME's performance decreases in various aspects. During the image data acquisition process, obtained remote sensing images and maps often exhibit spatial disparities and perturbations to some extent. These image perturbations can impact the model's performance, leading to a reduction in generation quality. However, our semi-supervised approach demonstrates a clear advantage in handling image perturbations. Despite a slight shortfall compared to the fully supervised ATME model, our method still showcases feasibility within a limited set of paired datasets.

Table 3 The comparison of parameters, training time, and testing time between our method and different models in the paper.

Model	Params	Training time	Testing time
CUT_MoNCE	11.378 M	42 h	31 s
CycleGAN	22.766 M	31 h	32 s
Pix2pixHD	45.874 M	16 h	15 s
SmapGAN	22.756 M	30 h	31 s
ATME	35.785 M	23 h	25 s
(Ours)	22.936 M	62 h	38 s

6 Conclusions

This paper addresses two issues in the method of generating maps from remote sensing images based on SmapGAN. First, the problem of feature information loss arises from the inability of the ResBlock based style converter to establish long-distance dependencies between features. To address this issue, we propose using MHSA-ResBlock in the style converter part of the SmapGAN network, which combines residual modules with MHSA to capture long-range dependencies between features, providing more accurate feature information for the upsampling operation. Second, to tackle the issue of blurred maps generated during upsampling due to the small receptive field of traditional transpose convolution, we propose a novel upsampling method named TC-Carafe, which combines traditional transpose convolution with the Carafe operator. This method performs upsampling through kernel prediction and feature recombination, resulting in clearer and more complete generated maps. Experimental results demonstrate that this research has achieved effective outcomes in enhancing map generation quality. The generation effects are superior to those of SmapGAN, CycleGAN, Pix2pixHD, CUT_MoNCE, and the ATME that is after perturbation of the training data. In future studies, our research will focus on map generation at different spatial scales and investigate how to improve the quality of generated maps while reducing the number of model parameters and improving training efficiency.

Code and Data Availability

The publicly available dataset used in this study is from http://efrosgans.eecs.berkeley.edu/cyclegan/datasets.

Acknowledgments

The author would like to thank the open source for providing the model and dataset. Simultaneously, we appreciate the valuable comments from the editors and reviewers. This work was supported by Zhejiang Provincial Foundation (Grant No. LS21F020003).

References

1. P. Haunold and W. Kuhn, "A keystroke level analysis of manual map digitizing," *Lect. Notes Comput. Sci.* **716**, 406–420 (1993).
2. L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Adv. Neural Inf. Process. Syst. 28*, C. Cortes et al., Eds., Curran Associates, Inc. (2015).
3. L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 2414–2423 (2016).
4. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent.* (2015).
5. P. Isola et al., "Image-to-image translation with conditional adversarial networks," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 5967–5976 (2017).
6. M. Mirza and S. Osindero, "Conditional generative adversarial nets," in *Computer Science*, pp. 2672–2680 (2014).
7. T.-C. Wang et al., "High-resolution image synthesis and semantic manipulation with conditional GANs," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 8798–8807 (2018).
8. J.-Y. Zhu et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE Int. Conf. Comput. Vision (ICCV)*, pp. 2242–2251 (2017).
9. J. Song et al., "MapGen-GAN: a fast translator for remote sensing image to map via unsupervised adversarial learning," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **14**, 2341–2357 (2021).
10. X. Chen et al., "SMAPGAN: generative adversarial network-based semisupervised styled map tile generation method," *IEEE Trans. Geosci. Remote Sens.* **59**, 4388–4406 (2021).
11. A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst. 30*, I. Guyon et al., Eds., Curran Associates, Inc. (2017).
12. K. He et al., "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 770–778 (2016).
13. J. Wang et al., "Carafe: content-aware reassembly of features," in *IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, pp. 3007–3016 (2019).
14. B. Elizabeth, "The new nature of maps: essays in the history of cartography (book)," *Engl. Hist. Rev.* **118**(478), 1093–1094 (2003).

15. C. Vega-Garcia, "Making maps: a visual guide to map design for GIS (second edition)," *Photogramm. Rec.* **27**(139), 389–390 (2012).
16. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
17. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 3431–3440 (2015).
18. L.-C. Chen et al., "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2018).
19. K. He et al., "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(2), 386–397 (2020).
20. S. Ren et al., "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1137–1149 (2017).
21. Y. Liu et al., "RoadNet: learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images," *IEEE Trans. Geosci. Remote Sens.* **57**(4), 2043–2056 (2019).
22. C. Liu et al., "Adaptive linear span network for object skeleton detection," *IEEE Trans. Image Process.* **30**, 5096–5108 (2021).
23. W. G. C. Bandara, J. M. J. Valanarasu, and V. M. Patel, "Spin road mapper: extracting roads from aerial images via spatial and interaction space graph reasoning for autonomous driving," in *Int. Conf. Rob. and Autom. (ICRA)*, pp. 343–350 (2022).
24. X. Li et al., "Topology-enhanced urban road extraction via a geographic feature-enhanced network," *IEEE Trans. Geosci. Remote Sens.* **58**(12), 8819–8830 (2020).
25. J. Mei et al., "CoANet: connectivity attention network for road extraction from satellite imagery," *IEEE Trans. Image Process.* **30**, 8540–8552 (2021).
26. A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Computer Science* (2015).
27. M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Adv. Neural Inf. Process. Syst. 30*, I. Guyon et al., Eds., Curran Associates, Inc. (2017).
28. S. Ganguli, P. Garzon, and N. Glaser, "GeoGAN: a conditional GAN with reconstruction and style loss to generate standard layer of maps from satellite images," arXiv:1902.05611 (2019).
29. Y. Fu et al., "Translation of aerial image into digital map via discriminative segmentation and creative generation," *IEEE Trans. Geosci. Remote Sens.* **60**, 4703715 (2022).
30. F. Zhan et al., "Modulated contrast for versatile image synthesis," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 18259–18269 (2022).
31. J. Song et al., "Semi-MapGen: translation of remote sensing image into map via semisupervised adversarial learning," *IEEE Trans. Geosci. Remote Sens.* **61**, 4701219 (2023).
32. E. Solano-Carrillo et al., "Look ATME: the discriminator mean entropy needs attention," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit. Workshops (CVPRW)*, pp. 787–796 (2023).
33. J. Xu et al., "SAM-GAN: supervised learning-based aerial image-to-map translation via generative adversarial networks," *ISPRS Int. J. Geo-Inf.* **12**(4), 159 (2023).
34. A. Srinivas et al., "Bottleneck transformers for visual recognition," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 16514–16524 (2021).
35. Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," arXiv:1611.02200 (2016).
36. X. Gao et al., "Image quality assessment based on multiscale geometric analysis," *IEEE Trans. Image Process.* **18**(7), 1409–1423 (2009).
37. Z. Wang et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.* **13**(4), 600–612 (2004).
38. C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Clim. Res.* **30**, 79 (2005).

Zhipeng Ding received his bachelor's degree in software engineering from Xinxiang University, Henan, China, in 2022. Currently, he is studying for a master's degree in software engineering at Hangzhou Normal University, Hangzhou, China. His research interests include computer vision and deep learning.

Ben Wang received his MS degree from Zhejiang University of Technology, China, in 2002, and his PhD from the University of Essex in 2007. He is now a professor in the School of Information Science and Technology (SIST), Hangzhou Normal University, Hangzhou, China. His research interests include digital forensics, the internet of things (IoT), and intelligent information processing.

Shuifa Sun received his BS degree in applied physics and in radio technique from Tianjin University, Tianjin, China, in 1999; his MS degree in telecommunication and information systems from Zhejiang University of Technology, Hangzhou, China, in 2002; and his PhD in information and telecommunication engineering from Zhejiang University, Hangzhou, China, in 2005. He is currently a professor in the College of Computer and Information Technology, China Three Gorges University, Yichang, China. He is also in SIST, Hangzhou Normal University, Hangzhou, China. His research interests include intelligent information processing, computer vision, digital forensics, and multimedia information processing.

Yongheng Tang received his master's degree in computer science from Liaoning University of Petrochemical Technology in China in 2021. Currently, he is pursuing a doctoral degree in management science and engineering at Three Gorges University in China. His research interests include intelligent information processing and computer vision.

Ren Zhuang received his bachelor's degree in mechatronic engineering from Shaoxing University, Zhejiang, China, in 2020. Currently, he is studying for a master's degree in software engineering at Hangzhou Normal University, Hangzhou, China. His research interests include natural language processing and deep learning.

WenBo Liu received his bachelor's degree in internet of things project from Normal University of LiuPanShui, GuiZhou, China, in 2022. He is currently pursuing his master's degree in the School of Information Science and Technology, Hangzhou Normal University, ZheJiang, China. His current research interests focus on deep learning, IOT perception, and intelligent sensor systems.