remote sensing

MDPI

Article

RSMT: A Remote Sensing Image-to-Map Translation Model via Adversarial Deep Transfer Learning

Jieqiong Song, Jun Li, Hao Chen * and Jiangjiang Wu

College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China; sjq@nudt.edu.cn (J.S.); junli@nudt.edu.cn (J.L.); wujiangjiang08@nudt.edu.cn (J.W.)
* Correspondence: hchen@nudt.edu.cn

Abstract: Maps can help governments in infrastructure development and emergency rescue operations around the world. Using adversarial learning to generate maps from remote sensing images is an emerging field. As we now know, the urban construction styles of different cities are diverse. The current translation methods for remote sensing image-to-map tasks only work on the specific regions with similar styles and structures to the training set and perform poorly on previously unseen areas. We argue that this greatly limits their use. In this work, we intend to seek a remote sensing image-to-map translation model that approaches the challenge of generating maps for the remote sensing images of unseen areas. Our remote sensing image-to-map translation model (RSMT) achieves universal and general applicability to generate maps over multiple regions by combining adversarial deep transfer training schemes with novel attention-based network designs. Extracting the content and style latent features from remote sensing images and a series of maps, respectively, RSMT generalizes a pattern applied to the remote sensing images of new areas. Meanwhile, we introduce feature map loss and map consistency loss to reinforce generated maps' precision and geometry similarity. We critically analyze qualitative and quantitative results using widely adopted evaluation metrics through extensive validation and comparisons with previous remote sensing image-to-map approaches. The results of experiment indicate that RSMT can translate remote sensing images to maps better than several state-of-the-art methods.

Keywords: map translation; adversarial transfer learning; remote sensing image; attention mechanism

Citation: Song, J.; Li, J.; Chen, H.; Wu, J. RSMT: A Remote Sensing Image-to-Map Translation Model via Adversarial Deep Transfer Learning. *Remote Sens.* **2022**, *14*, 919. https://doi.org/10.3390/rs14040919

Academic Editor: Claudio Piciarelli

Received: 8 December 2021
Accepted: 11 February 2022
Published: 14 February 2022

1. Introduction

Using remote sensing images to generate maps enables people to understand the natural and cultural elements of the world entirely. With significant progress in remote sensing combined with deep learning methods, automatically generating maps from remote sensing images has become promising in the geographic information field. Traditional map-making methods are time consuming, especially for emergency scenarios such as fire disasters, earthquakes, and terrorist attacks. Automated map generation methods can translate remote sensing images to maps rapidly, unlike man-made cartography. Automated map generation cannot replace traditional map-making methods completely nowadays, since it only conducts style transferring for remote sensing images to maps rather than map vectorization, but it plays a critical role in emergency geographic services.

In essence, generating maps from remote sensing images is deemed an image-to-image translation task, which learns to map an image in a specific domain to an analogous image in a different domain. The current image-to-image translation methods such as Gc-GAN [1], CycleGAN [2] and MapGen-GAN [3] can translate remote sensing images to maps for a specific area. However, it is well known that the style of urban construction infrastructure varies greatly in different regions. In some metropolises of China, high-rise office buildings and apartments are relatively typical and dense, while in the United States, the majority of the population live in single-family houses with private gardens. Specifically, if we

train an image translation model (such as MapGenGAN [3]) to generate maps of some areas in Beijing, we should use remote sensing images and maps of the areas in or around Beijing as the training set. If we use a well-trained translator (MapGen-GAN) that adopts Beijing images as training data to generate maps of Los Angeles (LA), the results will be distorted and blurred, as illustrated in Figure 1. Meanwhile, we adopted a well-trained MapGen-GAN using Los Angeles datasets as the training set to generate maps for areas in Beijing, and the comparison results are shown in Figure 2, from which we can draw a similar conclusion. Apparently, the remote sensing image-to-map translation model trained with data from one region does not perform well in another region with a different style. General one-to-one mapping image-to-image translation models for map generation are limited by cognate training and testing datasets. They cannot learn the capability to generalize an unseen class based on prior knowledge. If we use both Beijing and Los Angeles datasets as training sets, the model may be confused because of the mixed inputs. Therefore, designing a generalizable translation model that can transfer the mapping from source class to target class is a challenging task. The source class we use in this paper contains several cities' remote sensing images (RS class) and corresponding maps (Map class), which refers to the training set. The target class comprises a few of unseen cities' remote sensing images and maps, indicating the testing set.

(**a**) (**b**) (**c**) (**d**)

Figure 1. Comparison of output maps generated by Los Angeles-dataset-trained MapGen-GAN and Beijing-dataset-trained MapGen-GAN testing on Los Angeles datasets.(**a**) Remote sensing image of Los Angeles(LA). (**b**) Man-made Google Map of Los Angeles. (**c**) LA map generated by trained MapGen-GAN using LA training set. (**d**) LA map generated by trained MapGen-GAN using Beijing training set.

(**a**) (**b**) (**c**) (**d**)

Figure 2. Comparison of output maps generated by Los Angeles-dataset-trained and Beijing-dataset-trained MapGen-GAN testing on Beijing datasets. (**a**) Remote sensing image of Beijing. (**b**) Man-made Google Map of Beijing. (**c**) Beijing map generated by trained MapGen-GAN using Beijing training set. (**d**) Beijing map generated by trained MapGen-GAN using LA training set.

In recent years, adversarial deep transfer learning has thrived with good effectiveness and strong practicability. Deep transfer learning based on the adversarial mechanism aims to find a representation suitable for both the source class and target class. As an attempt to train a map translation model that can generalize the mapping to the unseen target class, we seek a GAN (generative adversarial network)-based remote sensing image-to-map translation framework under the inspiration of transfer learning methods [4,5]. The

map translation model aims to learn a strong extensible applicability from transforming remote sensing images to maps. Furthermore, in the application of remote sensing image-to-map translation tasks, the translation model must focus on the critical areas of the image. Attention mechanisms can help the image-to-image translation model locate points of interest and improve generative adversarial networks' performance [6–8]. Hence, we employ the attention mechanism directly to translation model to estimate regions of interest, which helps networks pay more attention to geographic structure information of remote sensing images, generating more realistic maps.

The proposed model RSMT is an extension of our previous work MapGen-GAN [3], which solves the limitation that the map generation model is only applied to the datasets of the same region and improves the applicability of map generation model. RSMT is based on an adversarial transfer training scheme coupled with attention mechanism design. The adversarial deep transfer network is to generalize the mapping to unseen areas' remote sensing images of target class by learning to extract the content and style latent features from remote sensing images and maps, respectively. Attention mechanisms are added to the adversarial deep transfer network to predict the region of interest and learn more critical information by assigning different weights to the input of generator. We also find that adopting an additional feature map loss produced by discriminator is beneficial for computing the discriminative regions from both real and generated maps. In addition, we introduce map consistency loss to preserve the consistency of domain-specific features and enhance the transfer ability of the generator. Through extensive experimental verification on several map datasets of different cities worldwide, we adopt three widely used performance metrics to evaluate RSMT and several baseline methods. The qualitative and quantitative results verify that RSMT can translate remote sensing images to maps more competitively than other state-of-the-art methods. The main ideas and major contributions of this study are summarized as follows:

1. We propose a novel remote sensing image-to-map translation framework named RSMT which extensively achieves functional capabilities to generate maps for multiple regions using adversarial deep transfer learning schemes. RMST has the ability to learn generalized patterns by extracting the content and style representations from remote sensing images and maps, which solves the limitation that the previous map generation model only applied to the testing datasets of the same region with training sets.
2. RSMT uses spatial attention feature maps extracted from the discriminator to help the generator explicitly capture the point of interest in source classes and unseen target classes. To further improve the proposed model's performance, we proposed a feature map loss function based on the spatial attention computed by discriminator to preserve domain-specific features during training. Moreover, we also introduce a novel map identity loss to improve the transfer capability of generator.
3. To demonstrate the effectiveness of deploying spatial attention mechanism in remote sensing image-to-map translation tasks, we conduct extensive experiments on different datasets worldwide to validate the usability and applicability of the proposed method. The quantitative and qualitative results show that RSMT significantly outperforms the state-of-the-art models.

The rest of our work is structured as follows: We enumerate related works of map generation techniques and cutting-edge image-to-image translation methods based on generative adversarial network in Section 2. In Section 3, the proposed remote sensing image-to-map translation framework RSMT is presented. Furthermore, in Section 4, we conduct objective and subjective experiments to verify the validity of RSMT. In Section 5, conclusions and discussions are mentioned.

2. Related Work

We divide the current related work into three aspects. Firstly, we review the development of image-to-image translation methods. Secondly, we introduce attention mechanisms

used in GAN (generative adversarial network)-based models. Finally, deep transfer learning based on the adversarial mechanism are presented.

2.1. Image-to-Image Translation

In recent years, the Generative Adversarial Network (GAN) [9] has developed rapidly in computer vision, bringing impressive results for image generation. The GAN-based image-to-image translation methods have been widely used and achieved good results, aiming to learn the mapping between the source domain and target domain. In the unpaired image-to-image translation field, the most pioneering algorithm is DiscoGAN [10], DualGAN [11] and CycleGAN [2]. They propose frameworks to map source domain images to target domain images and use cycle-consistency loss which preserves some properties of the original images. However, the loss function of the transformation algorithm with cycle-consistency constraint has certain defects, which needs to assume that the two domains of the translation are bi-directional mapping. Therefore, Park et al. [12] propose a simple method based on contrastive learning, aiming to maximize the mutual information between the input image and the corresponding patch in target field, which pays attention to the content of the object rather than its appearance. Additionally, the attribute vector of previous image-to-image methods is binary, and the results' control is not satisfactory enough. RelGAN [13] uses the relative attribute vector to solve this problem. Instead of using encoder and reconstruction loss, Alharbi et al. [14] retain the structure and discrimination loss of traditional GAN, avoiding complex network structure and superabundant parameters.

Classical image-to-image models merely learn the one-to-one mapping between domains, and each input only corresponds to a single output image. Almahairi et al. [15] design a many-to-many mapping model called Augmented CycleGAN. The model can generate multiple output images with different styles for one input image by learning a mapping to capture the diversity of outputs. It takes the sample and latent variable of the source domain as input and outputs a sample of the target domain. In the meanwhile, Choi et al. [16] propose the StarGAN structure, which can train multi image-to-image in a network and load multiple datasets in the same network. It attaches a domain classifier to the discriminator and proposes a domain classification loss. The mask vector is used to make GAN ignore the unknown labels of multiple datasets and aggregate them on the known ones. Many other kinds of studies such as [17–19] utilize different methods to discuss how to transform one image into multiple images with different styles. The training datasets used in our work contain multiple remote sensing images and maps of different urban architectural styles. Although our model is similar to these methods using multiple remote sensing images and maps in source classes, we test remote sensing image-to-map translation model on previously unseen target classes.

2.2. Attentional Mechanism

At present, various kinds of deep learning tasks use attention mechanisms widely. Inspired by the human visual system, when people observe external things, they usually do not see things as a whole and tend to obtain essential parts of the observed things selectively according to their needs. Similarly, the attention mechanism is added to the deep learning model in order to predict regions of interest and learn more critical information by assigning different weights to the input, making the model judge accurately without bringing overheads to the calculation. It improves the performance in diverse tasks, such as machine translation [20,21], Seq2Seq model [22,23], image segmentation [24,25] and image caption [26,27].

Recent studies have shown that combining attention learning with a GAN-based model can obtain more realistic images in image-to-image translation tasks. Zhang et al. [28] design a long-range dependency and attention-driven GAN, which is proficient in finding the dependency in images and coordinating the details of every position in the generated images. Then, image-to-image translation tasks employ attention learning. For example,

Emami et al. [29] use a segmentation annotation of the input images as additional supervision information and adopt the attention map to improve the generated images' quality. AttentionGAN [30] is improved based on CycleGAN to generate not only images but also attention maps, and then it combines attention maps and generated images with original images to obtain the final outputs. As a step towards remote sensing community, Zhang et al. [31] propose a channel attention mechanism adopted to re-scale channel features by considering the interdependence between channels. Our work further explores the plausible usage of attention learning in a GAN-based remote sensing image-to-map translation model.

2.3. Deep Transfer Learning Based on Adversarial Mechanism

Inspired by generative adversarial network (GAN) [9], deep transfer learning employing the adversarial mechanism aims to learn a representative pattern between source and target domains. Deep transfer learning is on the basis of assumption that "For effective transfer, good representation should be discriminative for the main learning task and indiscriminate between the source domain and target domain" [32]. Recently, adversarial deep transfer learning has thrived with its good effectiveness and strong practicability. A new deep architecture for domain adaptation was proposed by Ganin et al. [33], which is trained with a great deal of labeled data. Ajakan et al. [34] introduce an adversarial framework called DANN to transfer information for domain adaption. DANN is suited to the context of domain adaptation, in which training and testing datasets have similarities but come from disparate distributions. Tzeng et al. [35] use a new CNN architecture for transferring knowledge cross-domain and cross-task. For adversarial adaptation, Tzeng et al. [36] then first outline a new generalized framework called ADDA. ADDA provides a simplified and cohesive view by understanding the similarities and differences between recently proposed adaptation methods. At the same time, Luo et al. [37] propose a new framework that efficiently learns a transferable representation across different domains by generalizing the embedding to the new tasks. At present, deep transfer learning is mainly based on a supervised manner. Our framework is based on adversarial deep transfer learning, which is designed for remote sensing image-to-map translation tasks.

3. Methods

In this section, we first describe the overall architecture of our model RSMT, and then we detail how the generator and discriminator modules work in an adversarial manner. Finally, we formulate the minimax optimization problem by applying the adversarial loss function, feature map loss function, and map consistency loss function.

3.1. Overall Architecture

We design our model combined with a generator G and a spatial attention-guided discriminator D. The generator G contains a content encoder E_c, a style encoder E_s, and a decoder F_{cs}. Like [4], the proposed model RSMT aims to learn the mapping pattern from remote sensing images to maps of source classes and transfer the mapping to unseen target classes. Unlike other object class transformation problems, the model is never shown the style of remote sensing images in target class but is required to generate corresponding maps. During training, as shown in Figure 3, we use a remote sensing image from a random RS class as the input of the content encoder. Simultaneously, we extracted one map from each kind of map class (a total of K) as the input of the style encoder. The style encoder obtains K types of maps. The content encoder aims at extracting variant latent representations of remote sensing images, and the style encoder extracts class-specific latent representations of map style. A generalizable feature extractor is obtained by learning to extract latent patterns from diverse source classes, extending translation applicability to multiple regions. During testing, the unseen remote sensing images belonging to the target class are obtained from a new urban region. When we provide the well-trained model a few of remote sensing images from the target class, it has to generate the corresponding maps.

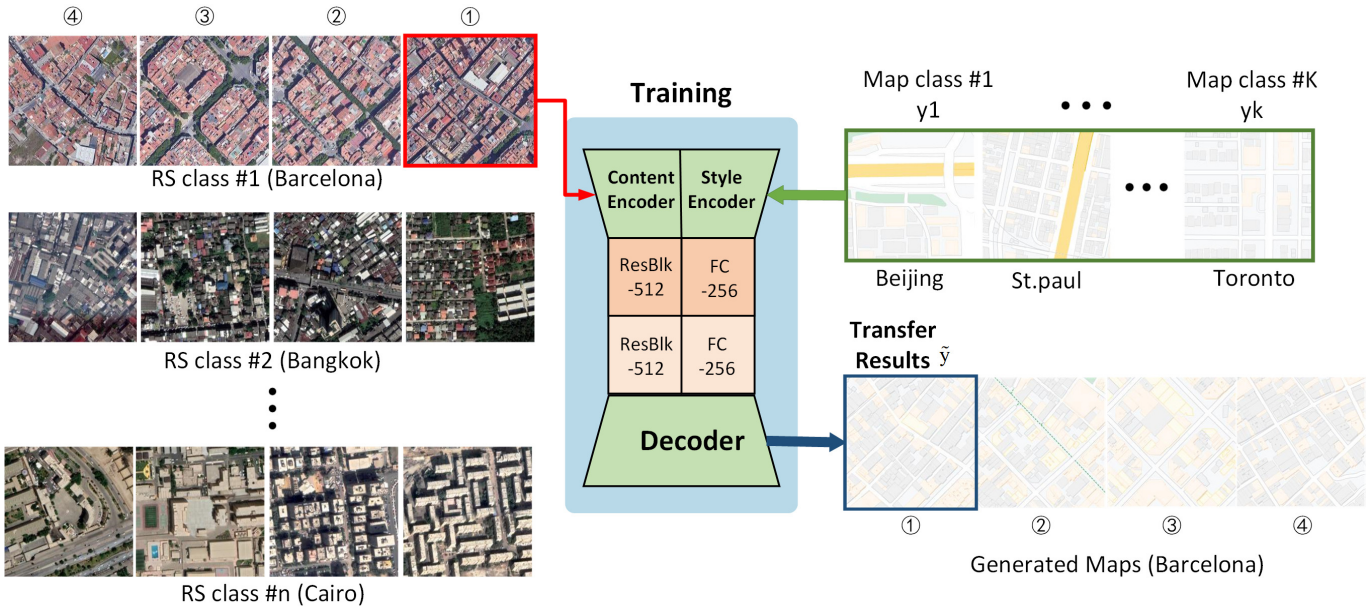


Figure 3. The flow path of RSMT training. The training set consists of paired remote sensing images and maps of various regions, like Barcelona, Bangkok, ..., Cairo (In this work, we use 12 cities' remote sensing images and maps as the training set). The generator takes two inputs: one is a remote sensing image, the other is a set of K random maps $\{y_1, y_2, \ldots, y_K\}$ obtained from map classes. The model aims to transfer the generation pattern to the unseen remote sensing images and generate realistic maps during testing. FC is the fully connection network.

Detailed information of RSMT are present in Figure 4. Since discriminator classifies images into real or fake, it is apparent that the discriminator can capture discriminative features in latent space. Hence, we introduce spatial attention features computed by discriminator and assist generator focus on relevant parts. The discriminator provides attention as the spatial feature map [38]. It indicates the areas that the discriminator pays attention to, and discriminates the input map as real or fake. Then, we feed the extracted spatial attention feature maps to the generator, forcing it to pay more attention to the discernable regions of source classes, generating more realistic maps.

Formally, we denote the data distribution as $x \sim P(x)$ and $y \sim P(y)$, where x denotes remote sensing images and y denotes map images (maps). The spatial attention feature map obtained from discriminator is defined as $M_{D_m}(y)$. Unlike generators for other existing image-to-image translation models [12,39], which only input one image at a time, we split the input of generator into two parts. One takes a remote sensing image x combined with a spatial attention feature map $M_{D_m}(y)$ as input, and the other takes a set of K maps $\{y_1, y_2, \ldots, y_K\}$ belonging to K Map classes as input. By combining the input $M_{D_m}(y) \odot x$ with K maps $\{y_1, y_2, \ldots, y_K\}$, the generator learns correlative latent features between RS classes and Map classes, and produces the output image $\tilde{y}$ via:

$$\tilde{y} = G(M_{D_m}(y) \odot x, \{y_1, y_2, \ldots, y_K\}) \tag{1}$$

where $\odot$ is matrix dot product [29].

3.2. Generator

The remote sensing image-to-map translation model RSMT contains a modified generator and an attention-based discriminator. Generator G is composed of a content encoder E_c, a style encoder E_s and a decoder F_{cs}. As depicted in Figure 4, content encoder E_c consists of four 2D convolutional layers with instance normalization and ReLU nonlinearity followed by several residual blocks. It extracts a feature map from the input map x, which is element-wised with spatial attention feature map $M_{D_m}(y)$. The content encoder E_c is

designed to encode the regions of interest part for the input $M_{D_m}(y) \odot x$ and produces a feature map z_x (content latent code). On the other hand, the style encoder E_s maps a set of K maps of different regions $\{y_1, y_2, \ldots, y_K\}$ to an intermediate latent vector. It consists of five 2D convolutional layers with ReLU nonlinearity that followed by a mean operation to acquire the vector z_y (style latent code).

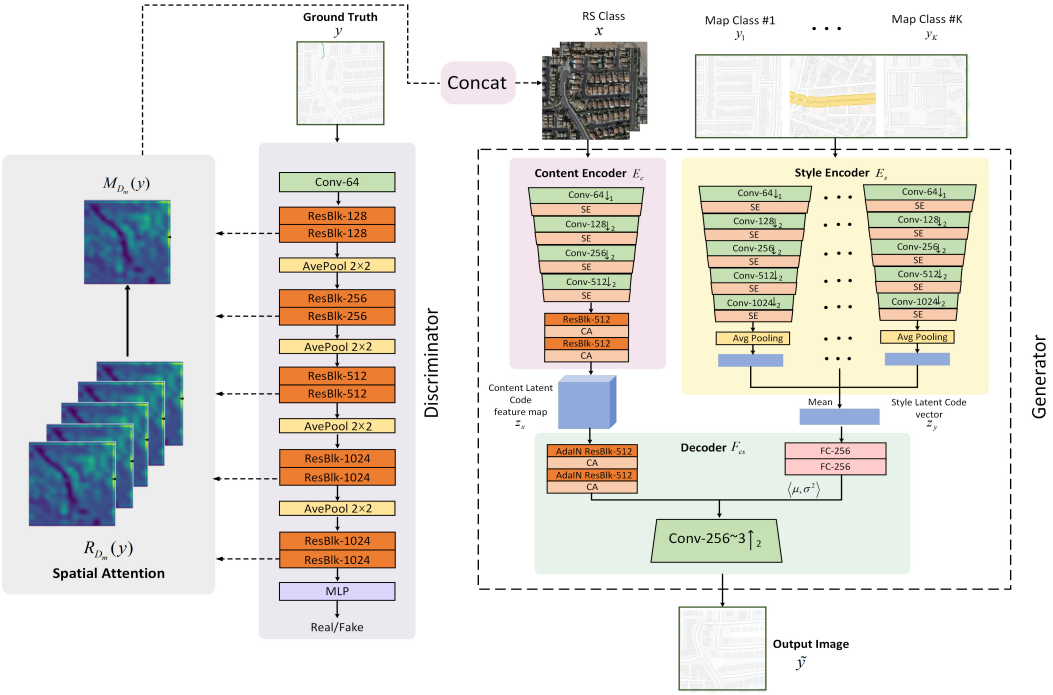


Figure 4. Details of the RSMT architecture. At first, the discriminator computes the spatial attention feature map of the corresponding map y (paired with remote sensing image x) by deploying $M_{D_m}(y)$, and then feeds it back into the generator concating with the remote sensing image x. The content encoder E_c encodes the input $M_{D_m}(y) \odot x$ to the content latent code z_x, and the style encoder E_s extracts K maps $\{y_1, y_2, \ldots, y_K\}$ belonging to K Map classes to the style latent code z_y. Finally, the decoder F_{cs} uses vector $\langle \mu, \sigma^2 \rangle$ as affine transformation parameters to obtain the output map $\tilde{y}$. CA is the channel attention, FC is the fully connected network, SE is the Squeeze-and-Excitation network.

Following the few-shot unsupervised image-to-image translation (FUNIT) [4], the decoder is composed of two adaptive instance normalization residual blocks [40] using adaptive instance normalization (AdaIN) [41] as the normalization layer and three upscale convolutional layers followed by ReLU nonlinearity. AdaIN first normalizes the activations of content latent code z_x in each channel. Meanwhile, the decoder F_{cs} computes the style latent code z_y generated from style encoder to a mean and variance vector $\langle \mu, \sigma^2 \rangle$ via two fully connected layers. Then, we use vector $\langle \mu, \sigma^2 \rangle$ as affine transformation parameters in adaptive instance normalization residual blocks to scale the activations, where μ is the biases and σ^2 is the scaling factor. The parameters $\langle \mu, \sigma^2 \rangle$ are applied to each residual block of the decoder to obtain global appearance information. With the content encoder E_c, style encoder E_s and decoder F_{cs}, the output map can be described as:

$$\begin{aligned} \tilde{y} &= F_{cs}(z_x, z_y) \\ &= F_{cs}(E_c(M_{D_m}(y) \odot x), E_s(\{y_1, y_2, \ldots, y_K\})) \end{aligned} \tag{2}$$

3.3. Discriminator with Spatial Attention Mechanism

Introducing attention learning into the GAN model can make remote sensing image-to-map translation more realistic. Inspired by several methods [25,38] that transfer a teacher network's attention to a student network, the proposed model RSMT deploys a spatial

attention mechanism to share knowledge between the discriminator and generator aiming at the relevant parts during translation. We employ the discriminator not only to classify the map as real or fake but also generate attention feature map fed back to the generator. The attention feature map indicates the discriminative area of the discriminator, making it distinguish the input image correctly. Feeding attention feature maps into the generator propels the network give higher weight to the region with obvious distinction, which can retain some specific features of the domain to a greater extent.

Like [29], we use normalized spatial attention feature maps produced from the discriminator to transmit the map's specific features. Specifically, by feeding the discriminator a map y, we obtain the spatial attention map $M_{D_m}(y)$. The size of spatial attention map $M_{D_m}(y)$ is identical to the input remote sensing image x. As illustrated in [38], $R_{D_m}(y)$ denotes the sum of the absolute values of activation maps in each spatial location in a residual block across the channel dimension, which can be described as:

$$R_{D_m}(y)=N(\sum_{j=1}^{C}\left|A_j(y)\right|) \tag{3}$$

where $A_j(y)$ denotes the activation of y in the j^{th} feature plane, and C denotes channels' number in the output of each residual block layer in discriminator for the input map y. $N(\cdot)$ is to normalize the input to [0,1] range and upsample the activation maps to match the original image size. $R_{D_m}(y)$ indexes the value of the potent information at each spatial location to discriminate an input map as fake or real directly. Discriminator's detailed architecture is shown in Figure 4, while there are five residual blocks in it. Considering that different layers of the discriminator network pay attention to different features, L attention maps are extracted from various layers in latent space for L residual blocks. Hence, the spatial attention map $M_{D_m}(y)$ obtained from discriminator is described as:

$$M_{D_m}(y) = N(\sum_{i=1}^{L} R_{D_m}(y)) = N(\sum_{i=1}^{L} N(\sum_{j=1}^{C}\left|A_{ij}(y)\right|)) \tag{4}$$

3.4. Loss Function

The remote sensing image-to-map model RSMT is trained by solving a minimax optimization problem:

$$\min_G \max_D \mathcal{L}_{GAN}(G,D) + \lambda_{FM}\mathcal{L}_{FM}(D) + \lambda_{MC}\mathcal{L}_{MC}(G) \tag{5}$$

where $\mathcal{L}_{GAN}(G,D)$, $\mathcal{L}_{FM}(D)$, $\mathcal{L}_{MC}(G)$ are the adversarial loss, feature map loss, and map consistency loss, respectively. In the following, we describe the expressions of these three loss functions detailedly.

3.4.1. Adversarial Loss

We summarize the principle of RSMT as transforming a remote sensing image x belonging to the RS class to a map y belonging to the Map class. The model trains the generator G in a generative adversarial network in order to learn the mapping from the RS class to Map class. By learning patterns from the real map y, we also train the discriminator D to distinguish whether the generated map is real or fake. The target of generator G is to minimize the loss against an adversary discriminator D. In contrast, discriminator D tries to maximize the objective function. The equation is expressed as:

$$\begin{aligned}\min_G \max_D \mathcal{L}_{GAN}(G,D) = {} & \mathbb{E}_{y\sim P(y)}[\log D(y)] \\ & + \mathbb{E}_{x\sim P(x),y\sim P(y)}[\log(1-D(G(x,y)))]\end{aligned} \tag{6}$$

In addition, we utilize the architecture of an attention-guided discriminator and introduce an attention feature map loss for the spatial attention feature computed by

the discriminator. The spatial attention feature indicates the discriminative area of the discriminator, making it distinguish the input image correctly. The discriminator needs to classify the similar region, while the feature map loss penalizes it to pay attention to distinct locations in both real map y and generated map $\tilde{y}$. Thereby, the feature map loss function between the real map $y \sim P(y)$ and generated maps $\tilde{y} \sim P(\tilde{y})$ is computed as follows:

$$\mathcal{L}_{FM}(D) = \mathbb{E}_{\tilde{y}\sim P(\tilde{y}),y\sim P(y)}[\|M_{D_m}(\tilde{y}) - M_{D_m}(y)\|_1] \tag{7}$$

where $M_{D_m}(\cdot)$ is the deployment of spatial attention maps obtained from discriminator, which is described in Equation (4) explicitly. $\|\cdot\|_1$ is the L_1-normalization [29]. The features of images by acting on L_1-normalization often achieve good results like [42]. Thus, we use L_1 to calculate the feature map loss and enhance performances of remote sensing image-to-map translation. We explain the accountable results of adding feature map loss in the ablation study in Section 4.

3.4.2. Map Consistency Loss

The map consistency loss helps generator reinforce the translation capability. When the map samples of source class are considered as the input to generator, we normalize generator to be closer to a consistent mapping. Specifically, during training, we put one map in content encoder E_c and K maps in style encoder E_s at the same time. Inspired by [2], generator G is used to generate map-style images. When a map is sent to G, it should still generate a map, which proves that the generator has the ability for map translation tasks. Thereby, map consistency loss penalizes the differences between the real map y and generated map $G(y, \{y_1, \ldots, y_K\})$. We define the map consistency loss as:

$$\mathcal{L}_{MC}(G) = \mathbb{E}_{y\sim P(y)}[\|y - G(y, \{y_1, \ldots, y_K\})\|_1] \tag{8}$$

where $\|\cdot\|_1$ is the L_1-normalization.

The ablation study experimental results in Section 4.6 prove that the map consistency loss helps the translation model RSMT improve the performance of generating maps.

4. Results

In the experimental section, we present the well-selected datasets and hardware equipment used in experiments firstly. Next, we compare RSMT with previous image-to-image methods quantitatively by using three evaluation metrics. Then, we find that the training ability of RSMT is related to the input Map class conditions by conducting different numbers of maps in style encoder. In the end, the ablation study is employed to analyze each component's effect in RSMT.

4.1. Datasets and Experimental Setups

Unlike other object class transformation problems, the paired datasets used for the remote sensing image-to-map translation tasks in this work are available from Google Maps, and the use of paired datasets can significantly improve the efficiency of training. To make it fair, we use paired datasets to train all comparable methods in this work. To verify the proposed map translation model's effectiveness, we explicitly selected paired datasets of 15 representative cities distributed around the world, such as Beijing and Bangkok in Asia; Los Angeles and Toronto in North America; Oslo and Pairs in Europe; Cairo in Africa; St.Paul in South America; Melbourne in Oceania, etc. All the maps were obtained from Google Map at zoom-17. We divided the datasets into a source class set representing a training set and a target class set representing testing set for transfer learning, which contains 12 and 3 cities, respectively. The source class contains 15,234 remote sensing images and paired 15,234 maps, including datasets of Beijing, Bangkok, Cairo, St.Paul, Toronto, Melbourne, Oslo, Pairs, Teheran, Barcelona, Las Vegas, and San Diego, while the target class contains

3000 remote sensing images composed of Los Angeles, Riyadh, and Vancouver. All the images' sizes are 256 × 256.

Details of the network design are shown in Figure 4. We set λ_{FM} = 10 and λ_{Id} = 1 by fine tuning. An Adam optimizer was used in both the generator and discriminator, the initial parameters, with the same learning rate of 0.0002. We adopted Python 3.7 and the PyTorch deep learning framework to accomplish the whole experiment. Two NIVIDA GTX 1080Ti GPU with 2 × 12GB GPU memory were employed to train the proposed model.

4.2. Evaluation Metrics

To evaluate the performance of our model RSMT and previous image-to-image methods quantitatively, we used three evaluation metrics: RMSE (Root Mean Square Error), SSIM (Structural Similarity), and ACC (Pixel Accuracy).

4.2.1. Root Mean Square Error

Root Mean Square Error (RMSE) [43] reflects the degree of difference between variables. It is an objective evaluation indicator of image quality based on pixel Error. Differences between the fusion image and reference image are measured by RMSE. The smaller the RMSE is, the better the quality of fusion image is. We express RMSE's general function as follows:

$$RMSE = \sqrt{\frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (I(i,j) - K(i,j))^2} \tag{9}$$

where M and N represent the length and width of the image, respectively; $K(i,j)$ and $I(i,j)$ are the pixel values at the pixel (i,j) of the original image and the image to be evaluated, respectively.

4.2.2. Structural Similarity

As an implementation of structural similarity (SSIM) theory [44], the structural similarity index defines structural information concerning image composition as an attribute of luminance, contrast, and structure. The mean value is used to estimate luminance, contrast is estimated by standard deviation, and covariance is measured by structural similarity.

Given an image x and a compared image y, the SSIM can be computed as follows:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{10}$$

where μ_x is the mean value of x, μ_y is the mean value of y, σ_x^2 represents the variance of x, and σ_y^2 represents the variance of y. Constant $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are calculated to maintain stability. L is the range of pixel values. Generally, $k_1 = 0.01$, and $k_1 = 0.03$. SSIM ranges from -1 to 1. The smaller the difference between the two images, the closer SSIM is to 1.

4.2.3. Pixel Accuracy

The abbreviation of pixel accuracy is ACC(%). Given a pixel i with the RGB value (r_i, g_i, b_i) of a ground-truth map and the generated map with RGB value $(r_i^{'}, g_i^{'}, b_i^{'})$, if $max(\mid r_i - r_i^{'} \mid, \mid g_i - g_i^{'} \mid, \mid b_i - b_i^{'} \mid) < \delta$ [1], we estimate this to be the precise generated map that is similar to the ground truth. We set δ = 5 in this paper.

4.3. Baselines

We evaluate the RSMT's performance by three state-of-the-art image-to-image approaches and FUNIT [4]. For the experiment, all methods work with paired datasets.

- CycleGAN: The cycle-consistency constrains are proposed by CycleGAN [2], which makes a mapping from X to Y, G_{XY}:$X \rightarrow Y$ and its inverse mapping from Y to X, G_{YX}:$Y \rightarrow X$. The objective function restricts to $G_{YX}(G_{XY}(x)) = x$, and $G_{XY}(G_{YX}(y)) = y$.
- DualGAN: DualGAN [11] conducts dual learning for image-to-image translation by applying Wasserstein GAN loss [45].
- MapGen-GAN: MapGen-GAN [3] is our previous work for map translation tasks. The framework is based on circularity and geometrical consistency constraints, transforming remote sensing images to maps directly and reducing the translation's semantic distortions.
- FUNIT: FUNIT [4] is an image-to-image translation framework based on few-shot learning that can translate images for previous unseen target classes. It can train multiple image-to-image processes in a network and even load multiple datasets in the same network.

4.4. Comparisons with Baselines

We compare our model with previous image-to-image methods in both objective and subjective ways to show the effectiveness of RSMT. The evaluation and analysis are presented from these two perspectives.

4.4.1. Objective Evaluation

This section compares the remote sensing image-to-map model RSMT with four previous methods to validate the RSMT's effectiveness. We evaluate the performance of RSMT by putting $K = 10$ maps of different regions $\{y_1, y_2, \ldots, y_K\}$ into the style encoder E_s and use three kinds of image quality measurement indexes. Since baselines we choose are representative image-to-image translation methods that only learn a one-to-one mapping between two domains, to be fair, we extend the source class to the same as RSMT (12 different cities mixed). We use same target class to test the translation results between our method and baselines to compare the generalization ability.

Table 1 presents the mixed training for the previous image-to-image methods compared with FUNIT and RSMT on three testing datasets. The highest score is bold, and the runner-up is underlined. Clearly, the proposed RSMT framework significantly outperforms all baselines for the remote sensing image-to-map translation task on three performance metrics. RSMT achieves about 46% accuracy with parameter $\delta = 5$, which is almost two times the improvement over the CycleGAN. In particular, RSMT yields a 22–18% improvement in RMSE and a 34–55% increment in SSIM among the previous image-to-image methods. The experiment results demonstrate that previous image-to-image methods cannot learn the capability to generalize an unseen class based on prior knowledge. That is because general one-to-one mapping image-to-image models only focus on a particular domain and do not have the ability to learn generalized features. RSMT learns the content and style of source class images through two kinds of encoders, which can apply the learned latent mapping patterns to the remote sensing images that have never been seen before.

Table 1. Scores for different methods on three testing datasets by three evaluation metrics. RSMT keeps $K = 10$, compared with other four models by mixed training. (bold: best; underline: runner-up).

Method	Los Angeles			Riyadh			Vancouver		
	RMSE	**SSIM**	**ACC (%)**	**RMSE**	**SSIM**	**ACC (%)**	**RMSE**	**SSIM**	**ACC (%)**
CycleGAN	30.2375	0.4931	22.1454	34.2375	0.4683	20.8642	27.9847	0.5543	23.8743
DualGAN	32.4654	0.4865	22.5231	35.1483	0.4526	19.3203	28.4632	0.5499	24.0580
MapGenGAN	33.7613	0.5046	25.6213	34.0984	0.4659	21.6732	27.0352	0.5460	26.8993
FUNIT	<u>24.8874</u>	<u>0.6673</u>	<u>34.4627</u>	<u>27.4577</u>	<u>0.6475</u>	<u>32.7814</u>	<u>23.6244</u>	<u>0.6698</u>	<u>35.9849</u>
RSMT	**20.2485**	**0.6822**	**45.7756**	**23.5894**	**0.6622**	**42.6528**	**20.0353**	**0.6964**	**43.5721**

The qualitative results of the mixed training comparison are displayed in Figure 5. We took two generated maps from each of the three cities. The first line of the figure is the input of remote sensing images, including Los Angeles, Riyadh, and Vancouver in order. As the baselines are not designed for the transfer learning settings, they failed in challenging translation task. On the other hand, our method can successfully translate remote sensing images to maps of novel classes and generate more impressive translations empirically. The general image-to-image translation methods only generate the buildings' rough outlines, and they do not approach the details. Comparatively, our method identifies the accurate contours of objects, making the geographical layout more regular. RSMT performs better than others due to the attention-guided discriminator collaborating with the generator, which makes the generator more exquisite to roads' transformation. It is evident that our model can translate the maps more similarly than FUNIT, which does not adopt the attention mechanism regarding the generation of roads and buildings. The comparisons verify the effectiveness of integrating the attention mechanism to the discriminator and constraining the network with feature map loss.

4.4.2. Subjective Evaluation

From Table 1, we can seethat the accuracy scores of all models are less than 50%, which is because these evaluation metrics are from a pixel-wise perspective. Maps are considered as an abstraction of human recognition of the world. From a subjective point of view, the slight changes in resolution and size will not affect the application of generated maps in emergency practice. However, these changes will degrade the quantitative accuracy of the generated maps. Therefore, we employed twenty persons with experience in cartography to separately grade the above methods' outputs subjectively in terms of similarity, fidelity, and availability to further discuss the performance of RSMT and baselines. The three indicators are rated on a scale from 1 to 10. Similarity means comparisons between the generated map and ground truth intuitively; a score of 10 means that the generated map is exactly the same as the ground truth, while a score of 1 indicates that the generated map is entirely different from the ground truth. Fidelity refers to whether the generated map matches the corresponding remote sensing image; a score of 10 represents an exact match between the generated map and the corresponding remote sensing image, while a score of 1 means that the generated map does not match the remote sensing image at all. Aside from the remote sensing image and ground truth, availability is to evaluate whether the generated maps can be used in real-world map services; 10 points means the generated map can be applied directly in the emergency map service; in contrast, 1 point means the generated map is completely unusable. As shown in Table 2, we calculate the average value of each method marked by each person on three testing sets. The winner is RSMT, and CycleGAN obtains the lowest score. The anthropogenic evaluation demonstrates that although the generated maps of RSMT do not score highly on pixel-wise quantitative metrics, it does not affect generated maps applied in practical emergency map services. Furthermore, we intend to improve the quantitative accuracy of generated maps for future studies.

Table 2. Average scores of subjective evaluation from human perceptual for RSMT and baselines on three testing datasets. (bold: best; underline: runner-up).

Method	Los Angeles			Riyadh			Vancouver		
	Similarity	**Fidelity**	**Availability**	**Similarity**	**Fidelity**	**Availability**	**Similarity**	**Fidelity**	**Availability**
CycleGAN	3.55	3.70	2.85	2.65	3.40	3.35	3.55	3.30	3.45
DualGAN	4.15	4.10	4.20	3.60	3.75	3.55	4.15	4.00	3.90
MapGenGAN	5.45	5.30	5.75	4.45	4.20	4.80	4.60	4.05	4.55
FUNIT	<u>7.70</u>	<u>7.95</u>	<u>8.30</u>	<u>7.35</u>	<u>7.60</u>	<u>8.15</u>	<u>7.40</u>	<u>7.35</u>	<u>7.95</u>
RSMT	**8.95**	**9.25**	**9.20**	**8.65**	**8.50**	**9.00**	**9.05**	**8.75**	**8.90**

4.5. Numbers of Input Map Classes During Training

In this section, we intend to figure out whether the generation ability of RSMT is related to the number of input Map classes conditions. As mentioned in Section 3, during training, we used a remote sensing image from a random RS class as the input of the content encoder. Simultaneously, we extracted one map from each kind of Map class (a total of K) as the input of the style encoder. The style encoder obtains K types of maps. In this part, we changed the number of input maps classes K fed to style encoder E_s from 1 to 12 with an interval of 2. Then, we evaluated the scores of testing datasets by three different evaluation metrics separately.

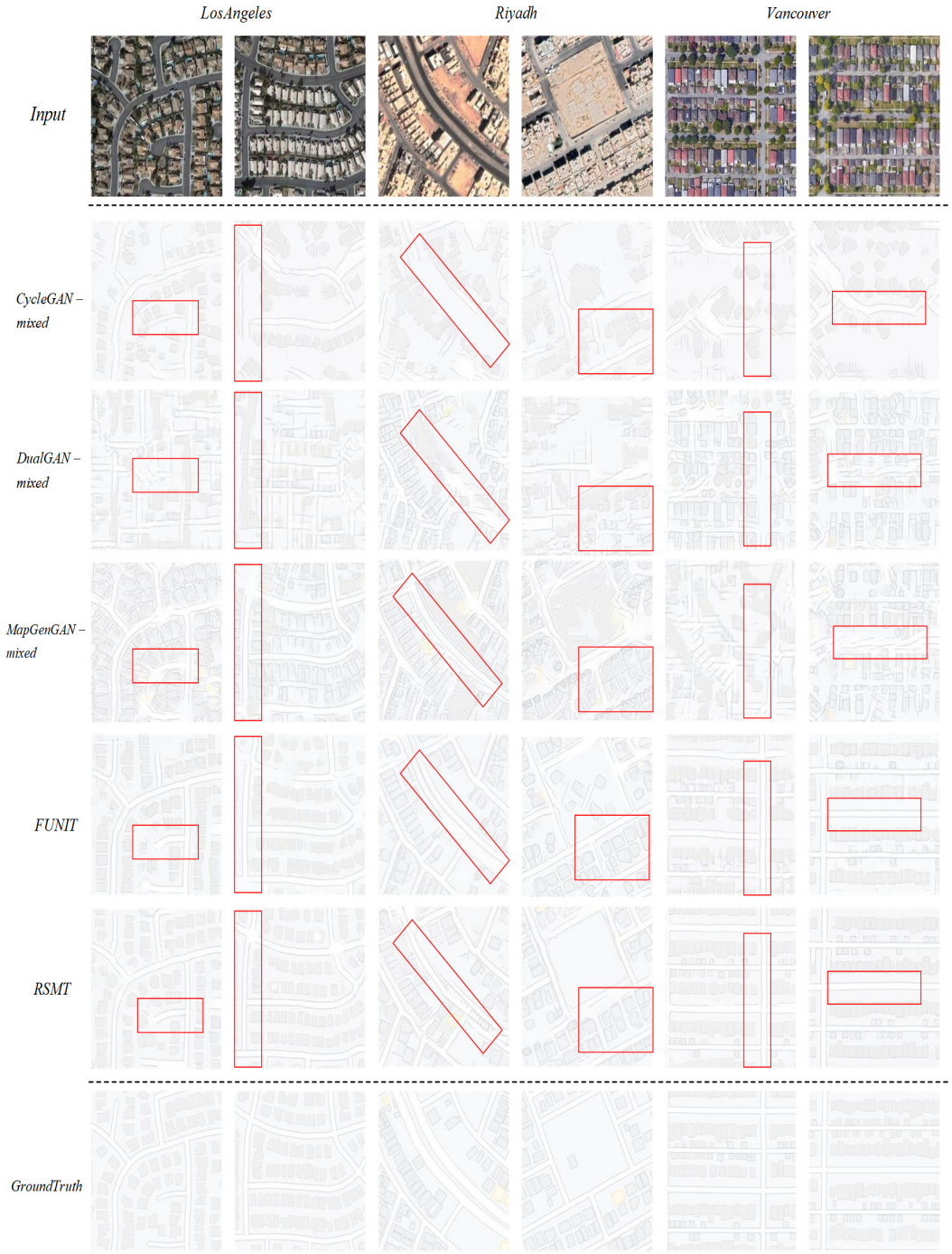


Figure 5. Selected results of remote sensing image-to-map translation performance for three testing datasets. Images from top to bottom: input remote sensing images; translation results of baselines; translation results of our model RSMT; the ground truth.

In Figure 6, we display our model's performance using different numbers of available Map classes during training and test RSMT-*K* on five testing datasets: Los Angeles, Riyadh, Vancouver, Milan, and Mexico City. We varied the number of Map classes from 1, 3, 5, 7, 9 to 12 and drew the performance trend chart, calculated by three evaluation metrics. The RMSE value is inversely proportional to image quality for grading schemes, whereas the SSIM and ACC values are directly proportional to the image quality. We found that the curves of scores on five testing datasets follow the same trend, which gives evidence that the performance of RSMT improves as the number of input Map classes increases. Although the performance of RSMT improves with the increase in the number of input Map classes in the source class, obtaining suitable quality map datasets is labor consuming, and a lot of time will be spent on data preprocessing. At the same time, as the amount of datasets grows, the training cost will also aggravate, considering time and GPU processing capacity.

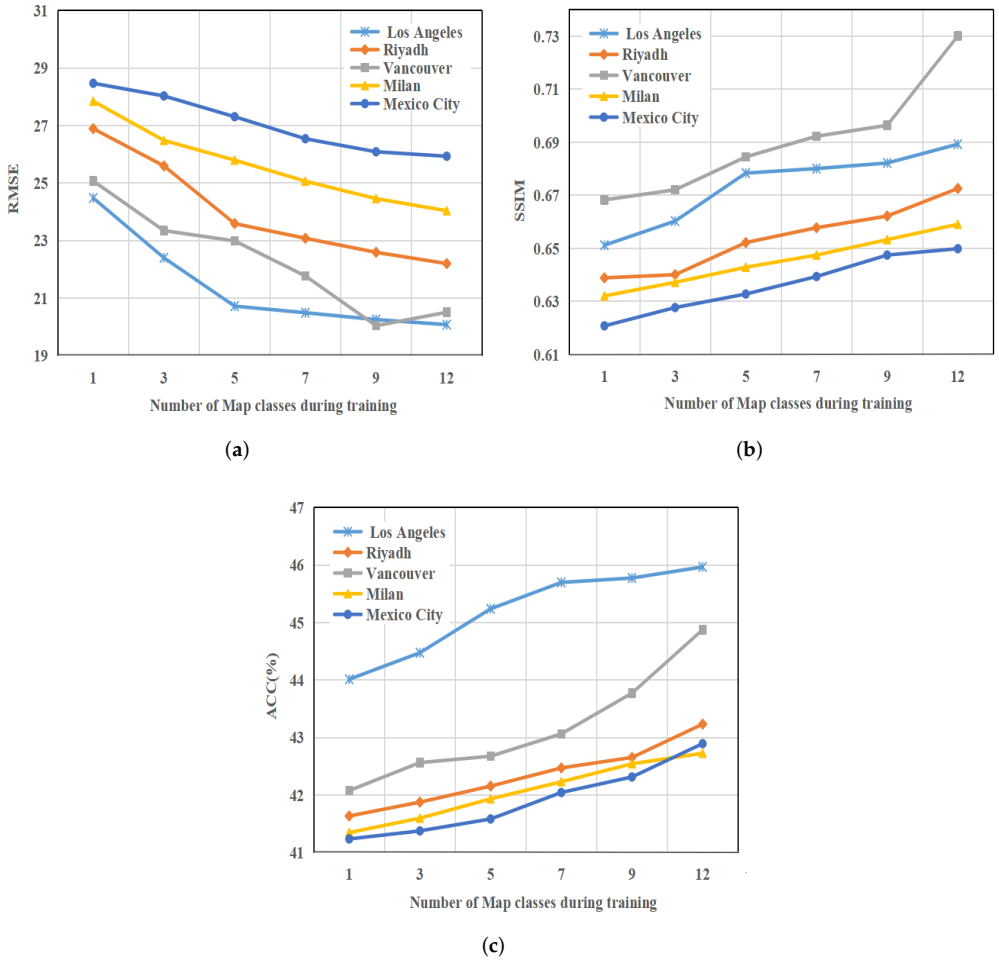


Figure 6. Remote sensing image-to-map translation performance of RSMT by deploying different numbers of input Map classes during training on five testing datasets. During training, the performance of RSMT is positively related tos the number of Map classes. (**a**) Trend chart of RMSE scores for RSMT-K (K = 1, 3, 5, 7, 9, 12) on five testing datasets. (**b**) Trend chart of SSIM scores for RSMT-K (K = 1, 3, 5, 7, 9, 12) on five testing datasets. (**c**) Trend chart of ACC scores for RSMT-K (K = 1, 3, 5, 7, 9, 12) on five testing datasets.

The evaluation demonstrates that the number of input Map classes during training impacts RSMT's performance, indicating that the remote sensing image-to-map translation model seeing more diverse Map classes during training performs better during testing. In addition, the two encoders work separately: content encoder extracts variant latent

representations of remote sensing images, and style encoder extracts class-specific latent representations of map style. They do not interact with each other.

4.6. Ablation Study

The ablation study was conducted to verify the rationality of remote sensing image-to-map translation framework RSMT further. As explained in Section 3, we inputted one remote sensing image and corresponding K maps of different cities during training. During testing, we used remote sensing images of Los Angeles datasets to validate the importance of core components in RSMT. From the beginning, we kept all components and divided the experiments into two parts; one is setting the number of input maps K = 5 of 5 Map classes (RSMT-5); the other is setting K = 10 (RSMT-10). Then, we removed the spatial attention feature map computed by discriminator but retained the feature map loss (RSMT-no-$\mathcal{M}_{D_m}$), because the feature map loss was calculated between real and generated maps by an attention-guided discriminator. Further, the spatial attention feature map was fed to the generator from the discriminator without adding feature map loss (RSMT-no-$\mathcal{L}_{FM}$). Finally, the map consistency loss was removed (RSMT–no-$\mathcal{L}_{MC}$). We describe these several ablation methods as follows:

- RSMT-no-$\mathcal{M}_{D_m}$: RSMT without spatial attention fed to the generator.
- RSMT-no-$\mathcal{L}_{FM}$: RSMT without feature map loss function constraint.
- RSMT-no-$\mathcal{L}_{MC}$: RSMT without map consistency loss function constraint.

The quantitive scores testing on Los Angeles datasets are reported in Table 3. In general, the performance score of RSMT-10 is slightly higher than RSMT-5 on three evaluation indicators. Under the same number of input Map classes, RSMT outperforms RSMT-no-$\mathcal{M}_{D_m}$ and RSMT-no-$\mathcal{L}_{FM}$, demonstrating that the spatial attention mechanism applied to the discriminator has the ability to drive network learning strong extensible applicability for remote sensing image-to-map translation task. Obviously, the feature map loss restricts the comparability between the distinguished regions of ground truths and the generated maps at the level of abstraction, which can help the model achieve realistic outputs. However, in comparison, the impact of map consistency loss on the model is minimal.

We display the qualitative results of RSMT by removing spatial attention, feature map loss, and map consistency loss testing on Los Angeles datasets in Figure 7. Our model can generate high-quality maps considering translation accuracy and image quality. We outline the most discriminative areas on the output maps in red. From the figure, RSMT can translate more impressive maps from remote sensing images than RSMT-no-$\mathcal{M}_{D_m}$, RSMT-no-$\mathcal{L}_{FM}$, and RSMT-no-$\mathcal{L}_{MC}$. This suggests that calculating feature map loss and applying spatial attention mechanism to the discriminator lead to higher performance, including the generation of roads and buildings.

Table 3. Scores of ablation study testing on Los Angeles datasets by removing core elements of RSMT calculated by three evaluation metrics. (bold: best; underline: runner-up).

Method	$K = 5$			$K = 10$		
	RMSE	**SSIM**	**ACC(%)**	**RMSE**	**SSIM**	**ACC(%)**
RSMT-no-$\mathcal{M}_{D_m}$	25.9228	0.5212	40.2405	24.8307	0.5456	41.7637
RSMT-no-$\mathcal{L}_{FM}$	23.1028	0.5738	43.7245	22.9706	0.5869	44.0631
RSMT-no-$\mathcal{L}_{MC}$	<u>21.2417</u>	<u>0.6446</u>	<u>45.2739</u>	<u>21.0378</u>	<u>0.6798</u>	<u>45.0195</u>
RSMT	**20.7162**	**0.6784**	**45.5392**	**20.2485**	**0.6822**	**45.7756**

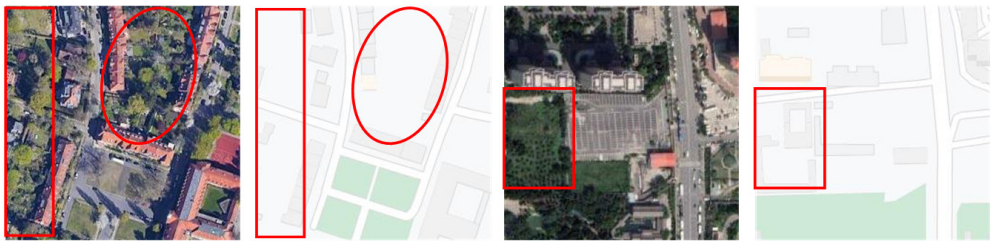
Figure 7. Qualitative comparisons over removing the core elements from the remote sensing image-to-map translation model RSMT. From top to bottom, the lines are input remote sensing images, translation results of RSMT-no-$\mathcal{M}_{D_m}$, RSMT-no-$\mathcal{L}_{FM}$, RSMT-no-$\mathcal{L}_{MC}$ by setting K = 5 and K = 10, translation results of RSMT, and ground truth.

4.7. Limitations

Although RSMT can generate maps much more impressively from remote sensing images of unseen areas than other general one-to-one mapping image-to-image translation methods and multi-class image-to-image translation methods, the results are not uniformly positive, especially for obscure areas with roads and sparse distributions of buildings. Through repetitious attempts and experiments, we drew the conclusion that the model's performance relies on the quality of training datasets. We have noticed two issues on maps that may affect the quality of the training datasets. Firstly, as shown in Figure 8a, buildings are artificially divided into two colors (pale yellow and gray) in maps, but there is no significant difference in remote sensing images. Secondly, as shown in Figure 8b, human map annotators often label vegetation green, while some remote sensing imagery green areas are often artificially ignored on maps. These issues would lead to chaotic learning for the network during training, which decreases the transformation performance and the quality of generated maps. Obtaining suitable quality map datasets is labor consuming, which will spend a lot of time in data preprocessing.

On the other hand, as illustrated in Section 4.5, the generation ability of RSMT is related to the number of input Map classes conditions. Experimental results give evidence that the performance of RSMT improves as the number of input Map classes increases. However, as the amount of datasets grows, the training cost will also aggravate, considering time and GPU processing capacity. In addition, our method is currently only applicable to urban areas and performs poorly on datasets in rural areas. This is possible because the distribution of ground objects in rural areas is sparse, and the features are not specific.

(**a**)



(**b**)

Figure 8. Issues on artificial maps. (**a**) The colors of the buildings in the map are artificially divided into two types (pale yellow and gray), but there is no difference in the remote sensing images. (**b**) Green areas in maps are not marked.

5. Conclusions

We propose a novel remote sensing image-to-map translation model named RSMT. For map generation, RSMT can extensively achieve functional capabilities in multiple regions based on an adversarial deep transfer network, which is designed to extract content and style representations from remote sensing images and maps. With the help of a spatial attention mechanism, RSMT attends to the regions of interest and produces more semblable images by learning features from the discriminator. Moreover, we applied feature map loss and map consistency loss to preserve domain-specific features and improved the transfer capability of the model. After extensive experimental comparisons on different

map datasets, our model can produce more competitive generations than previous remote sensing image-to-map approaches objectively and subjectively.

For future works, we intend to pre-process the datasets comprehensively while improving translation accuracy. Adopting a weak or semi-supervised approach may alleviate the data collecting pressure. Additionally, we will also take the consistency of maps at different levels into consideration.

Author Contributions: Conceptualization, J.S. and H.C.; methodology, J.S. and H.C.; experiment, J.S.; data curation, J.S. and J.W.; writing—original draft preparation, J.S.; writing—review and editing, H.C.; supervision, J.L. and H.C.; funding acquisition, J.L. and H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant U19A2058, Grant 61806211, Grant 41971362, in part by the Natural Science Foundation of Hunan Province China under Grant 2020JJ4103.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Zhang, K.; Tao, D. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2427–2436.
2. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
3. Song, J.; Li, J.; Chen, H.; Wu, J. MapGen-GAN: A Fast Translator for Remote Sensing Image to Map Via Unsupervised Adversarial Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2021**, *14*, 2341–2357. [CrossRef]
4. Liu, M.Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T.; Lehtinen, J.; Kautz, J. Few-shot unsupervised image-to-image translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 10551–10560.
5. Choi, Y.; Uh, Y.; Yoo, J.; Ha, J.W. Stargan v2: Diverse image synthesis for multiple domains. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8188–8197.
6. Kim, J.; Kim, M.; Kang, H.; Lee, K.H. U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
7. Tang, H.; Xu, D.; Sebe, N.; Yan, Y. Attention-guided generative adversarial networks for unsupervised image-to-image translation. In Proceedings of the IEEE 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.
8. Alami Mejjati, Y.; Richardt, C.; Tompkin, J.; Cosker, D.; Kim, K.I. Unsupervised Attention-guided Image-to-Image Translation. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 3693–3703.
9. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Adv. Neural Inf. Process. Syst.* **2014**, *3*, 2672–2680. [CrossRef]
10. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1857–1865.
11. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2849–2857.
12. Park, T.; Efros, A.A.; Zhang, R.; Zhu, J.Y. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 319–345.
13. Wu, P.W.; Lin, Y.J.; Chang, C.H.; Chang, E.Y.; Liao, S.W. Relgan: Multi-domain image-to-image translation via relative attributes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 5914–5922.
14. Alharbi, Y.; Smith, N.; Wonka, P. Latent filter scaling for multimodal unsupervised image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1458–1466.
15. Almahairi, A.; Rajeshwar, S.; Sordoni, A.; Bachman, P.; Courville, A. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 195–204.

16. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8789–8797.
17. Lin, J.; Xia, Y.; Wang, Y.; Qin, T.; Chen, Z. Image-to-image translation with multi-path consistency regularization. *arXiv* **2019**, arXiv:1905.12498.
18. Hui, L.; Li, X.; Chen, J.; He, H.; Yang, J. Unsupervised multi-domain image translation with domain-specific encoders/decoders. In Proceedings of the IEEE 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 2044–2049.
19. Zhao, B.; Chang, B.; Jie, Z.; Sigal, L. Modular generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 150–165.
20. Luong, M.T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1412–1421.
21. Cheng, Y. Agreement-based joint training for bidirectional attention-based neural machine translation. In *Joint Training for Neural Machine Translation*; Springer: Singapore, 2019; pp. 11–23.
22. Bahuleyan, H.; Mou, L.; Vechtomova, O.; Poupart, P. Variational Attention for Sequence-to-Sequence Models. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 21–25 August 2018; pp. 1672–1682.
23. Chiu, C.C.; Sainath, T.N.; Wu, Y.; Prabhavalkar, R.; Nguyen, P.; Chen, Z.; Kannan, A.; Weiss, R.J.; Rao, K.; Gonina, E.; et al. State-of-the-art speech recognition with sequence-to-sequence models. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4774–4778.
24. Chen, L.C.; Yang, Y.; Wang, J.; Xu, W.; Yuille, A.L. Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3640–3649.
25. Rout, L.; Misra, I.; Moorthi, S.M.; Dhar, D. S2a: Wasserstein gan with spatio-spectral laplacian attention for multi-spectral band synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 188–189.
26. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 2048–2057.
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
28. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 7354–7363.
29. Emami, H.; Aliabadi, M.M.; Dong, M.; Chinnam, R.B. Spa-gan: Spatial attention gan for image-to-image translation. *IEEE Trans. Multimed.* **2020**, *23*, 391–401. [CrossRef]
30. Tang, H.; Liu, H.; Xu, D.; Torr, P.H.; Sebe, N. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *arXiv* **2019**, arXiv:1911.11897.
31. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
32. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A survey on deep transfer learning. In *International Conference on Artificial Neural Networks*; Springer: Cham, Switzerland, 2018; pp. 270–279.
33. Ganin, Y.; Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 1180–1189.
34. Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M. Domain-adversarial neural networks. *arXiv* **2014**, arXiv:1412.4446.
35. Tzeng, E.; Hoffman, J.; Darrell, T.; Saenko, K. Simultaneous deep transfer across domains and tasks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4068–4076.
36. Tzeng, E.; Hoffman, J.; Saenko, K.; Darrell, T. Adversarial discriminative domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7167–7176.
37. Luo, Z.; Zou, Y.; Hoffman, J.; Fei-Fei, L. Label efficient learning of transferable representations across domains and tasks. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 164–176.
38. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* **2016**, arXiv:1612.03928.
39. Lee, H.Y.; Tseng, H.Y.; Huang, J.B.; Singh, M.; Yang, M.H. Diverse image-to-image translation via disentangled representations. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 35–51.
40. Huang, X.; Liu, M.Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 172–189.
41. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.

42. Park, T.; Liu, M.Y.; Wang, T.C.; Zhu, J.Y. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2337–2346.
43. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82. [CrossRef]
44. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
45. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved training of wasserstein GANs. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5769–5779.