SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

February 2021, Vol. 64 122301:1–122301:15
https://doi.org/10.1007/s11432-020-3077-5

Reciprocal translation between SAR and optical remote sensing images with cascaded-residual adversarial networks

Shilei FU, Feng XU* & Ya-Qiu JIN

Key Lab for Information Science of Electromagnetic Waves (MoE), Fudan University, Shanghai 200433*, China*

Received 30 April 2020/Revised 22 July 2020/Accepted 30 September 2020/Published online 21 January 2021

Abstract Despite the advantages of all-weather and all-day high-resolution imaging, synthetic aperture radar (SAR) images are much less viewed and used by general people because human vision is not adapted to microwave scattering phenomenon. However, expert interpreters can be trained by comparing side-by-side SAR and optical images to learn the mapping rules from SAR to optical. This paper attempts to develop machine intelligence that is trainable with large-volume co-registered SAR and optical images to translate SAR images to optical version for assisted SAR image interpretation. Reciprocal SAR-optical image translation is a challenging task because it is a raw data translation between two physically very different sensing modalities. Inspired by recent progresses in image translation studies in computer vision, this paper tackles the problem of SAR-optical reciprocal translation with an adversarial network scheme where cascaded residual connections and hybrid L1-GAN loss are employed. It is trained and tested on both spaceborne Gaofen-3 (GF-3) and airborne Uninhabited Airborne Vehicle Synthetic Aperture Radar (UAVSAR) images. Results are presented for datasets of different resolutions and polarizations and compared with other state-of-the-art methods. The Frechet inception distance (FID) is used to quantitatively evaluate the translation performance. The possibility of unsupervised learning with unpaired/unregistered SAR and optical images is also explored. Results show that the proposed translation network works well under many scenarios and it could potentially be used for assisted SAR interpretation.

Keywords synthetic aperture radar, generative adversarial network (GAN), image translation, cascaded residual connection, Frechet inception distance

Citation Fu S L, Xu F, Jin Y-Q. Reciprocal translation between SAR and optical remote sensing images with cascaded-residual adversarial networks. Sci China Inf Sci, 2021, 64(2): 122301, https://doi.org/10.1007/s11432-020-3077-5

1 Introduction

Synthetic aperture radar (SAR) is capable of imaging at high resolution in all-day and all-weather conditions. As a cutting-edge technology for space remote sensing, it has found wide applications in earth science, weather change, environmental system monitoring, marine resource utilization, planetary exploration, etc. High resolution and multi-dimension are the two major trends of recent development of spaceborne SAR technology. Imaging resolution has been improved from ten-meter in 1990s (e.g., SIR-C/X-SAR), to meter in 2000s (e.g., Radarsat2), and to sub-meter in 2010s (e.g., TerraSAR-X). Despite the rapid progresses in SAR imaging technologies, the bottleneck challenge remains in the interpretation of SAR imagery and it is becoming more and more urgent as a huge volume of SAR data are being acquired daily by numerous radar satellites in orbits.

Owing to its distinct imaging mechanism and the complex electromagnetic (EM) wave scattering process, SAR exhibits very different imaging features from optical images. Some basic differences between SAR images and natural optical images are summarized in Table S1 of Appendix A. Human's visual system is adapted to the interpretation of optical images. SAR images are difficult to be interpreted by ordinary people. Although SAR images contain rich information about targets and scenes, such as

* Corresponding author (email: fengxu@fudan.edu.cn)

geometric structure and material property, they can only be interpreted by well-trained experts. This has now become the major hindrance in utilization of existing SAR archives and further promotion of SAR applications.

The major objective of this work is to develop a deep learning application with large amount of co-registered SAR and optical images where SAR images can be translated to optical images and vice versa. The translated optical image can then be used in assisted interpretation of SAR image by ordinary people. Imagine that, with such a translation tool, any person without any background knowledge of radar, could be able to understand the primary information contained in SAR images. This could greatly promote the wide application and usage of future and existing archives of SAR remote sensing imageries. Other potential applications include facilitation of data fusion of optical and SAR images, e.g., translating an optical image at an earlier date as the reference SAR image for SAR change detection, registering the unpaired SAR and optical images, integrating optical and microwave data into a single image to enhance multi-spectral features, etc.

A SAR-optical image reciprocal translation generative adversarial network (GAN) architecture is proposed in this paper. It follows the typical image translation GAN architecture [1–4] employing a convolutional neural network (CNN) as the discriminator and a specially-designed network as the generator/translator. The translator uses the multiscale encoder-decoder CNN as the backbone and incorporates novel multiscale cascaded-residual connections. To reduce the instability during the training of GAN, a hybrid loss function is used to train the generator which contains two parts: the GAN loss back-propagated from discriminator output, and the L1-distance loss directly applied to the generated sample and the true sample.

The proposed method is verified on a large volume of SAR-optical image pairs, i.e., more than 10000 samples of 256×256 size patch. The dataset covers different urban/suburban regions and mainly contains earth surfaces such as built-up areas, roads, vegetation, waters and farmlands. Appearances of these terrain objects, e.g., buildings, have great diversity which makes the training and testing more generalizable. The algorithm is tested at different resolutions. Frechet inception distance (FID) [5] is used as the quantitative measure of the similarity between the reconstructed and the true image. Overall, the translated high-resolution optical images can partially serve the purpose of assisted interpretation of SAR images. Code is available at the web[1)].

The major contributions of this paper are as follows.

- A modified image translation GAN architecture with multiscale cascaded residual connections is proposed for raw image translation between two very different sensing modalities, SAR and optical sensors.
- Experiment results on large volume of dataset demonstrate good visual quality and variety that can be achieved by the proposed network. Extensive analyses are conducted with quantitative metrics on space-borne and airborne SAR images. Different factors are analyzed including resolutions, targets, polarization, frequency bands, and input image scales.
- An extension towards unsupervised learning is tested with the CycleGAN loop [2]. Results demonstrate that using large volume of unpaired SAR and optical images, the performance can be further improved.
- Several experiments are made to explore the real applications of the network on image segmentation and airplane synthesis.

This paper is organized as follows. Section 2 reviews the relevant recent studies about image translation and SAR generation. Section 3 presents the proposed translation network architecture, loss function and training techniques. Experiments with real SAR images are carried out and results are presented and evaluated in Section 4. Section 5 explores practical applications of CRAN. Finally in Section 6, conclusion is drawn and the future perspectives of application are discussed.

2 Related work

Many studies have been carried out in the multi-sensor image fusion regime [6]. The main objective of data fusion is to integrate complementary information from multi-sensor images of the same region into an enhanced image which appears better than any of the original ones. Hybrid pansharpening method, the weighted combination method (WC method), the integration method based on the magnitude ratio

1) https://github.com/Shilling818/CRAN.

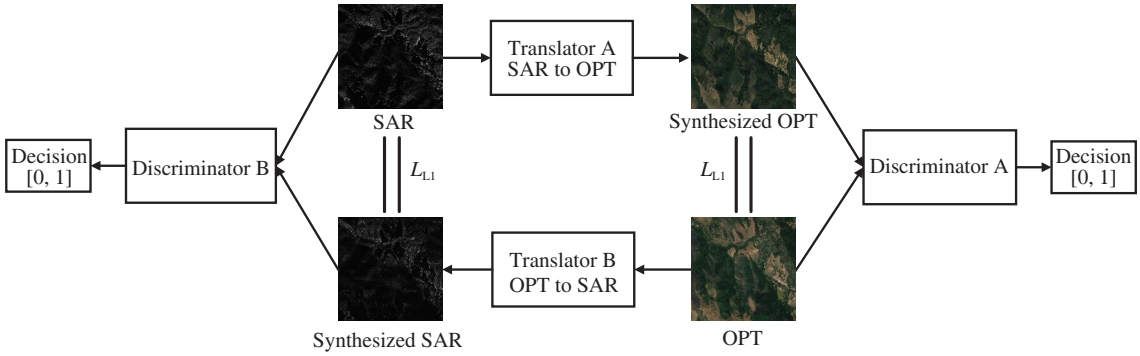

Figure 1 (Color online) Schematic diagram of the translation network inspired by Pix2Pix network [1] during training. A pair of translators are trained together. Each translator consists of an encoder and a decoder. The two discriminators are trained separately. 'SAR', 'OPT', 'Synthesized OPT' and 'Synthesized SAR' respectively represent the true SAR image, the true optical image, the fake optical image and the fake SAR image. The two vertical lines connecting 'SAR' and 'Synthesized SAR' mean that the network should make them equal constrained by a L1 norm loss.

of the two images (MR method) are often employed [6]. The fused images are well-defined and diversely textured. Garzelli [7] leveraged co-registered SAR images to improve the quality of optical images, which extracted specific information from SAR images and complements with optical images so that the targets could appear clearer.

Another type of fusion work is registration between SAR and optical images. The focus is to explore the consistent features between the two sensing modalities. Fan et al. [8] designed a uniform nonlinear diffusion-based Harris feature extraction method to explore many more well-distributed feature points with potential of being correctly matched. Liu et al. [9] transformed the two types of images into a feature space where their feature representations became more consistent using a deep convolutional coupling network. Merkle et al. [10] synthesized artificial SAR-like patches from optical images and matched them with the true SAR patches utilizing some known matching approaches, like scale-invariant feature transform (SIFT). Especially, Liu's method is very instructive for the content consistence of the feature space in the SAR and optical image translation.

A third type of fusion is to combine multi-temporal SAR and optical images to generate images at different observation times. He and Yokoya [11] used a conventional conditional GAN to generate the optical image from a SAR image at a close date with the aid of another temporal SAR-optical image pair at the same site. They also tried to directly generate an optical image from a single SAR image but found that the existing GANs failed to do so. Schmitt et al. [12] trained the network Pix2Pix on a large number of SEN1-2 patch pairs and got good predicted optical image patches. Some earlier attempts [13, 14] tried to convert coarse-resolution or simulated SAR images to visible images using conditional GANs but results show that terrain objects such as buildings cannot be translated. Wang et al. [15] adopted CycleGAN to translate SEN-1 images to SEN-2 ones and also achieved the goal of cloud removal. Li et al. [16] proposed a modified conditional GAN trained by the hybrid loss functions, structure similarity index measure (SSIM) and L1 norm, and translated targets are mainly rivers and vegetation. Fuentes et al. [17] added several residual layers to the internal architecture of CycleGAN, and they adopted two types of SAR imageries, Urban Atlas and SEN1-2, and translated them to the grayscale optical images.

3 SAR-optical reciprocal translation network

3.1 Translation framework

The proposed framework is shown in Figure 1. It has two reciprocal directions of translation, i.e., SAR to optical and optical to SAR. Each direction consists of two adversarial deep networks, i.e., a multiscale convolutional encoder-and-decoder network as the translator vs. a convolutional network as the discriminator. The translator takes in a SAR image, maps it to the latent space via the encoder, and then remaps it to a translated optical image. The discriminator takes in both the translated optical image and the true optical image which is co-registered with the original SAR image, and outputs the classification results. The discriminator learns to identify the translated optical images from the true optical images, while the translator network learns to convert the SAR image to an optical image as

realistic as possible to fool the discriminator. On the other direction, the network is constructed exactly in the same manner with the only difference being optical as input and SAR as target.

The discriminator is a conventional CNN for a binary classification task. The translator has multiscale convolutional layers for encoder and decoder where direct paths are connected from the encoder to the decoder at different scales. Besides the direct paths in the latent space, residual connections in the input image space are further incorporated at each scales. A conventional binary classification loss is employed to train the discriminator, while its opposite loss, together with a L1 norm loss, is used to train the translator. These are explained in detail in Subsection 3.3.

3.2 Network architecture

Figure 2 shows the architecture and parameters of the translator network, which is named as cascaded-residual adversarial network, abbreviated CRAN. It follows the main structure of U-Net [18] and Pix2Pix [1] with certain modifications. On the encoder side, the input image is convolved at one scale and downsampled to the next scale repeatedly for 6 times. On the decoder side, the latent feature map is deconvolved and upsampled back to the original scales. Notably, we include direct links from the encoder to the decoder. In addition, the network structure of CRAN contains multiscale cascaded residual connections from input to the multiple decoder stages. This is different from conventional ResNet connections such as the one employed in the encoder part in [4]. On the other hand, the network employed in [3] contains a single skip connection from input to the last stage of U-Net output which, according to our experiences, has lower capability in generating image details than the cascaded pattern as used in this paper. We believe that such multiscale cascaded residual connections are effective in generating vivid high-resolution images. In order to increase the depth of the network, at each time upsampling the feature maps, it first concatenates the encoder's feature maps to the current ones and deconvolves. Then concatenate the residual block to the former output feature maps and deconvolve again. This results in the increase of the decoder's receptive field. Thus the receptive field of the encoder and that of the decoder will be asymmetrical, which may degrade performance. The solution is to convolve feature maps of each scale twice in the encoder.

Regarding the hyperparameters, in the translator, the convolutional kernel is 3×3, the encoder and the decoder each have 12 layers, and the receptive field per pixel of the input image is 191×191. In the discriminator, the kernel is 4×4, and it has 5 layers and the receptive field is 70×70. The benchmark number of feature maps in the generator is set to 50. The number of feature maps doubles at each downsampling and diminishes double after each upsampling. Those in the discriminator are respectively 64, 128, 256, 512 and 1. That means in the discriminator, the feature maps extracted from the input image are finally mapped into a 32×32 matrix and every value corresponds to a 70×70 patch of the input. By contrast, the difference between the discriminative matrix of the true image and that of the reconstructed image could determine how similar the spatial structures of the two images are. The total number of the generator's weights is approximately 53.75 million and that of the discriminator is 2.76 million. The network architectures are depicted in detail in Figures 2 and 3.

3.3 Loss functions

Loss functions are critical for training of the networks. The discriminator is trained with a binary classification log-loss [19], i.e.,

$$L(D) = -E_{x\sim p_{\text{data}}(i)}[\log D(x)] - E_{z\sim p_{\text{data}}(j)}[\log(1 - D(T(z)))], \tag{1}$$

where $i = 0, 1$ in $p_{\text{data}}(i)$ demonstrate the distributions of the true optical and SAR images respectively. i and j are mutually exclusive, that is, $j = 0$ when $i = 1$, and vice versa. $E_{x\sim p_{\text{data}}(i)}$ denotes that x obeys the distribution $p_{\text{data}}(i)$, and $E_{z\sim p_{\text{data}}(j)}$ denotes that z obeys the distribution $p_{\text{data}}(j)$. When z denotes the original input SAR (or optical) image, $T(z)$ denotes the translated optical (or SAR) image and x denotes the corresponding true optical (or SAR) image. $D(\cdot)$ denotes the output probability map of the discriminator. For the discriminator, minimizing $L(D)$ is equivalent to classifying x as 1 and $T(z)$ as 0.

Following the adversary scheme [19], the loss function of the translator is

$$L_{\text{GAN}}(T) = -\sum_i E_{z\sim p_{\text{data}}(i)}[\log(D(T(z)))], \tag{2}$$

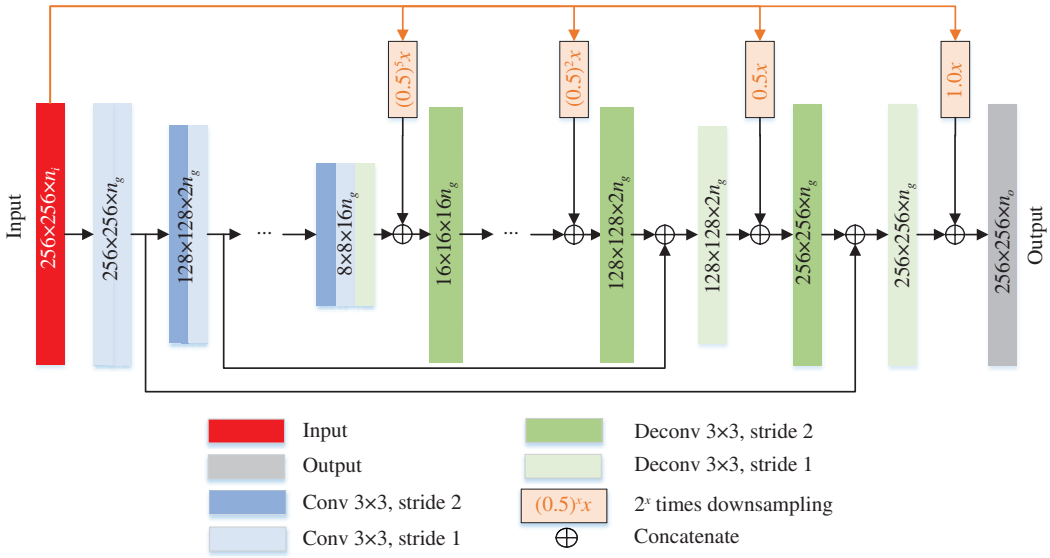

Figure 2 (Color online) Translator network architecture with cascaded-residual connections. The input data size is $256 \times 256 \times n_i$ and the output data size is $256 \times 256 \times n_o$. The first two numbers represent the size of the feature maps and the third number represents the channel of the feature map. The symbols n_i and n_o denote the channel number of the input and output image respectively, set to 1 for SAR images and 3 for optical images. The symbol n_g denotes the benchmark number of feature maps in the generator, that is, the number of feature maps of first layer. The concatenation from the encoder and the input to the decoder is signified by lines with arrows.

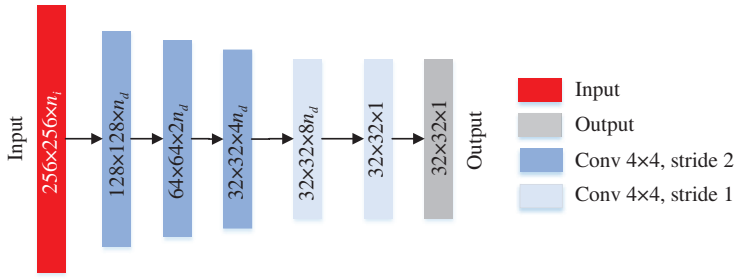

Figure 3 (Color online) Discriminator network architecture. The input data size is $256 \times 256 \times n_i$ and the output probability map size is $32 \times 32 \times 1$. The first two numbers represent the size of the feature maps and the third number represents the channel of the feature map. The symbol n_i denotes the channel number of the input image, and n_d denotes the benchmark number of feature maps in the discriminator, set to 64 here.

where $L_{\mathrm{GAN}}(T)$ is the sum loss of the two translated networks. Opposite to the goal of the discriminator, the translator is aimed at synthesizing realistic images to fool the discriminator to classify them as 1.

Isola et al. [1] found that the adversary loss function is better to be hybrid with traditional loss, such as L1 or L2 loss, which requires training by co-registered image pairs. Therefore here, L1 norm loss is used to hybridize with the GAN loss, i.e., L1 distance between the translated image $T(z)$ and the true image x:

$$L_{\mathrm{L1}}(T) = \sum_{i,j} E_{x\sim p_{\mathrm{data}}(i), z\sim p_{\mathrm{data}}(j)}[||x - T(z)||_1^1]. \tag{3}$$

Combine the above two equations together with appropriate weights and derive the final loss function $L(T)$ of the translators:

$$L(T) = L_{\mathrm{GAN}}(T) + \beta L_{\mathrm{L1}}(T). \tag{4}$$

$L(T)$ is the objective function for two translators, whose parameters are simultaneously updated. The two discriminators are allocated with the independent loss function $L(D)$ and trained separately.

A quick experiment is conducted to show the efficacy of the proposed loss function. Different losses contribute differently to the qualities of reconstructed results. In Figure 4, it is found that reconstructed optical images trained under L1-only loss are blurred and low-frequency features such as contours can be learned while high-frequency fine textures are missing. The model trained under GAN-only loss can

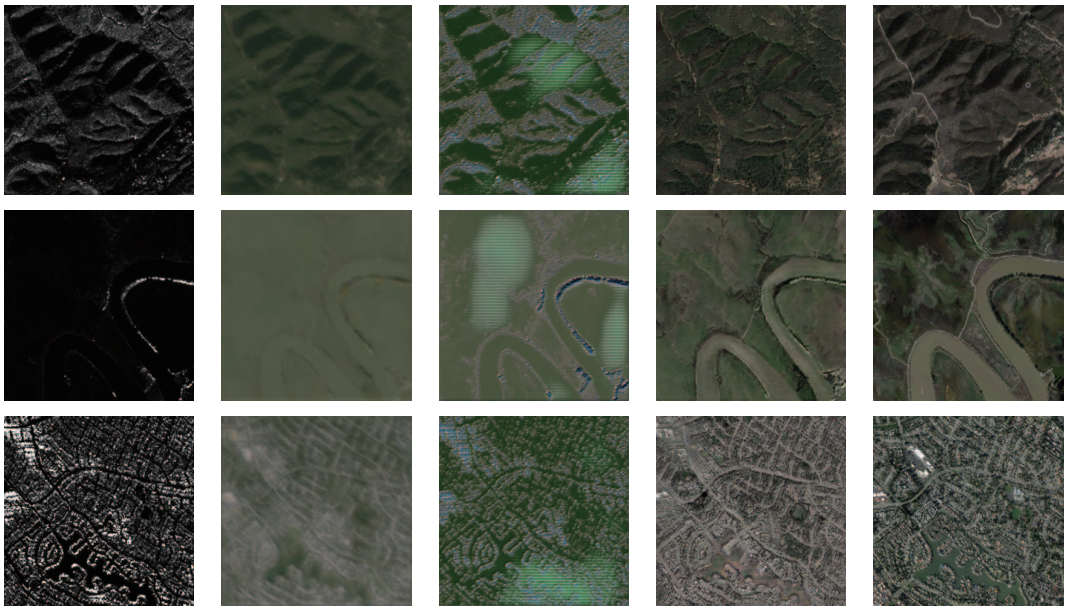

Figure 4 (Color online) Different qualities of translated optical images are induced by different losses. The first column lists (a) the input SAR images; the intermediate three columns are respectively (b) translated optical images with L1-only loss, (c) translated optical images with GAN-only loss and (d) translated optical images with L1+GAN loss; the last column are (e) the corresponding optical ground truths.

learn the details and the targets are more prominent, but large-scale smooth features and their spatial distributions are not well reconstructed. We also notice that some artifacts appear in homogenous areas such as water. Besides, training with GAN-only loss often encounters the well-known 'mode collapse' problem [20]. Mode collapse is a fatal training problem of GAN where $T(z)$ is collapsed to a fixed sample to maintain low loss but sacrificing the diversity. These issues can be alleviated by using the hybrid L1 and GAN loss, in which case, the synthesized images can have both low-frequency and high-frequency characteristics. Moreover, the network can also be trained more stably with the hybrid loss.

Stochastic gradient descent algorithm with adaptive moment estimation (Adam) can be used to train the two translators/discriminators simultaneously. The training strategy is provided in Appendix B.1. We made several comparison experiments and set the parameter $\beta = 20$ in (4), with which the initial values of $L_{\mathrm{GAN}}(T)$ and $L_{\mathrm{L1}}(T)$ are approximately equal and the model is more stable and generates better results. We set the learning rate to 0.0002. Adam optimizer with $\beta 1 = 0.5$, the exponential decay rate of the first moment estimation, is used. The input images are linearly mapped to the interval $[-1, 1]$. LeakyReLU is selected as the activation function. Batch normalization is used before the activation function except the first or last layer. All the trainable parameters are initialized as the truncated normal distribution with mean 0 and standard deviation 0.02. These hyperparameters are mainly selected based on the implementation of Pix2Pix [1]. When the batch size is set too small, due to the difference between training samples, a slight oscillation occurs in the gradient decent and the curve of the loss convergence appears steady declining with oscillations. Early stop is adopted during training. When the test loss does not decrease for four epochs in a row, the training is forced to stop.

3.4 Towards unsupervised learning

Supervised learning with well co-registered optical and SAR image pairs produces good results. However, such dataset is not always available and even if available, it would require a significant amount of effort for image registration. Thus, this paper also explores the possibility of unsupervised learning with unpaired SAR and optical images. CycleGAN [2] proposes a cyclic loop which could be leveraged for this purpose. As shown in Figure 5, the SAR image is first fed to the translator A and synthesizes a fake optical image. Then the fake optical image is used to synthesize the cyclic fake SAR images by the translator B. On the other hand, the optical image is used to synthesize a fake SAR image which is then further used to synthesize the cyclic fake optical images. The cyclic images are compared with the corresponding true images in a pixel-by-pixel fashion, while the synthesized fake images are fed into the 'critic' discriminator

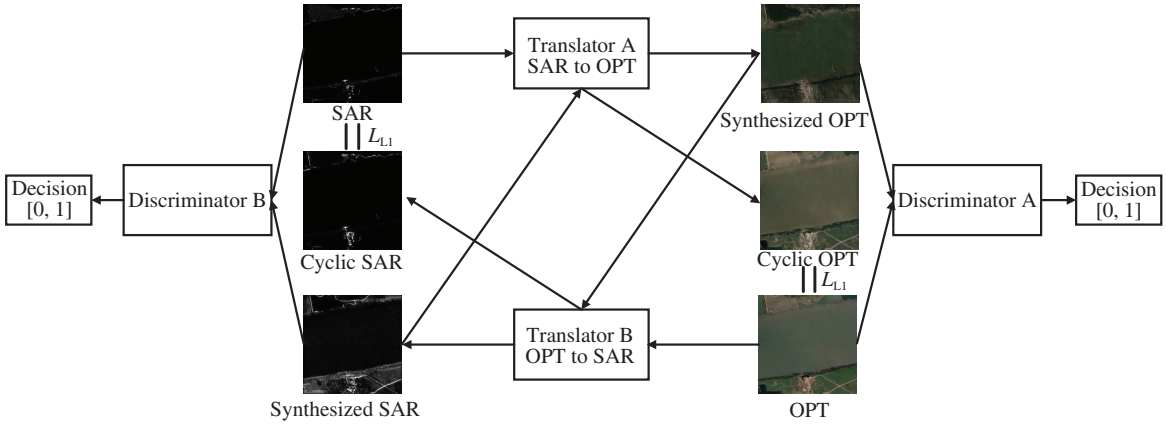

Figure 5 (Color online) Modified network scheme for unsupervised learning with CycleGAN loops [2].

networks. The translators A and B networks are trained alternatively during these two loops together with the discriminator networks. Later in Subsection 4.4, we demonstrate how such unsupervised learning could further improve the performance of translators initially trained with a small number of co-registered image pairs.

4 Experiment

4.1 Datasets

SAR data used in this study mainly come from the spaceborne GF-3 SAR from China[2)] and the airborne UAVSAR system from NASA[3)]. The information of those registered SAR and optical data used in our experiments is listed in Table S2 of Appendix A. It is found that the registration error of UAVSAR image pairs is about 3 pixels, and that of GF-3 image pairs is about 10 pixels. The more blurred contours of GF-3 SAR data increase the difficulty of selecting calibration points, leading to larger registration errors.

A simple preprocessing step is to normalize the pixel values of the SAR images to $[-1, 1]$. Owing to the considerably large range of pixel values, we have to determine a suitable threshold value to normalize the SAR image without changing the contrast. The normalized pixel value of the SAR image is defined as

$$\hat{x} = \begin{cases} 1, & \text{if } x > \bar{x}; \\ 2x/\bar{x} - 1, & \text{if } 0 \leqslant x \leqslant \bar{x}, \end{cases} \tag{5}$$

where x and $\hat{x}$ represent the pixel values of SAR images before and after normalization. $\bar{x}$ is λ times the mean value of the image x, defined as

$$\bar{x} = \lambda \left(\sum\nolimits_{i=1}^{N} x_i \right) / (N - n), \tag{6}$$

where x_i is the i-th pixel of the image x, N is the total number of pixels and n is the total number of pixels in the element 0. Here set $\lambda = 2000$.

Another preprocessing step is to perform speckle filtering on the GF-3 SAR images using a fast nonlocal despeckling filter [21]. We found that speckle filtering can improve the quality of the final synthesize images. We have a total of 12854 pairs of co-registered samples, 20% of which are randomly selected as test samples while the rest as training samples. During the preparation of the dataset, it is found that, owing to the difference of acquisition time of SAR and optical images, some new buildings shown in the recent optical images were not captured in the SAR image. This may adversely affect the final results.

4.2 Quantitative evaluation

Here we design experiments to test the performance of our model for SAR images of different resolutions and different polarization modes. The experiment for different resolutions adopts medium-resolution

2) China Centre for Resources Satellite Data and Application. Gaofen3. 2017. http://www.cresda.com/CN/.
3) NASA. UAVSAR. 2018. https://vertex.daac.asf.alaska.edu/.

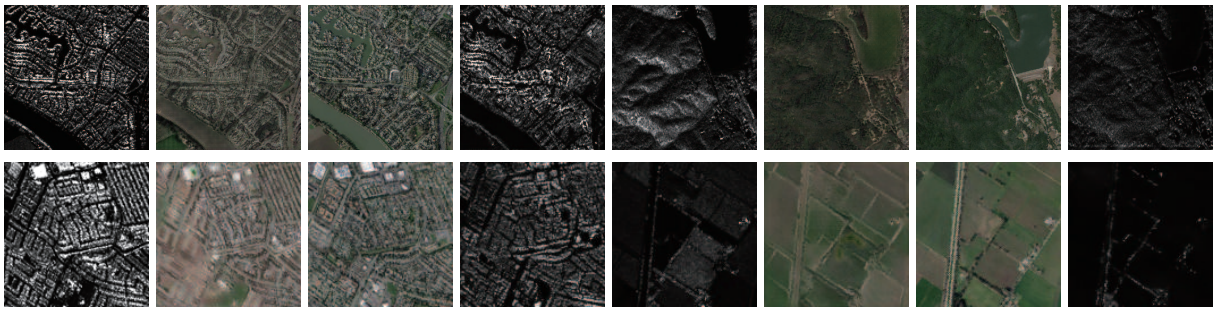

Figure 6 (Color online) Example translation images with UAVSAR (test samples). Images in each row from left to right are the real SAR image ((a1), (a2)) and its translated optical image ((b1), (b2)), the real optical image ((c1), (c2)) and its translated SAR image ((d1), (d2)). The first row is chosen from 6 m UAVSAR dataset, and the latter is from 10 m UAVSAR dataset.

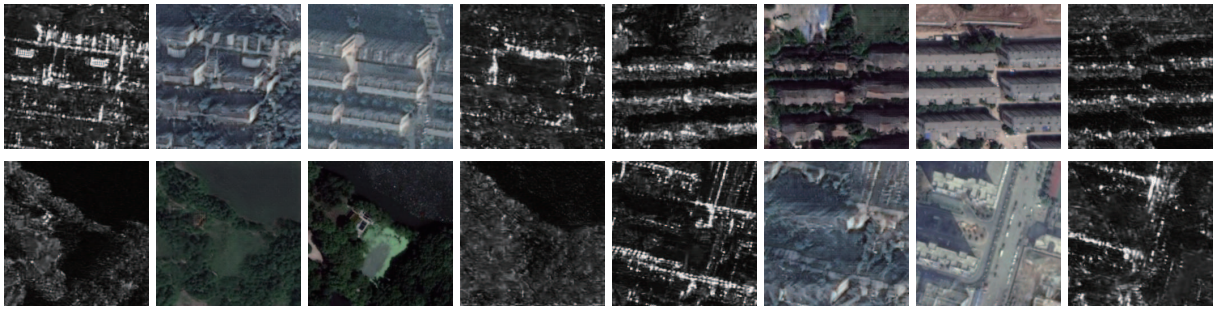

Figure 7 (Color online) Example translation images with GF-3 data. Images in each row from left to right are the real SAR image ((a1), (a2)) and its translated optical image ((b1), (b2)), the real optical image ((c1), (c2)) and its translated SAR image ((d1), (d2)).

UAVSAR and high-resolution GF-3 datasets respectively; the experiment for different polarization modes uses single-polarized (single-pol) and full-polarized (full-pol) UAVSAR data.

4.2.1 *Resolution*

The UAVSAR images are resampled to resolution of 6 m and 10 m and then used to train the proposed network. An example of translated SAR and optical images are shown in Figure 6 where a good visual quality is achieved.

Figure 7 shows examples of high-resolution GF-3 images translated by the proposed networks. The first row is a training sample and the rest three rows are test samples. It is found that earth surfaces like waters and vegetation can be easily reconstructed in both training and test cases. The textures of buildings vary greatly, so the reconstructed buildings are definitely difficult to match the ground truths. In addition, low-rise buildings can be rebuilt into cubes, but their edges are not well-aligned. If the buildings are too close, the open space between them is difficult to distinguish. For high-rise buildings, the ones shown in the training samples appear to be reasonably realistic. However, the ones in the test sample in the bottom row appear to be smeared. It seems like that the network got confused by the viewing angles. Apparently, for tall 3D terrain objects, both SAR and optical images are very sensitive to the view angles. Without incorporation of the projection mechanism, the proposed network is not able to generalize in this dimension.

It is necessary to quantitatively measure the difference between the translated images and true ones. Traditional methods, such as L1 norm, peak signal-to-noise ratio (PSNR) and SSIM, could be used to measure the similarity between two images. However, these methods still compare the similarity in terms of pixel values but rather than in the sense of perceptual similarity. Inception score (IS) and FID [5] are usually used to quantitatively evaluate the quality and variety of images generated by GANs.

FID uses the statistics of real world samples and compares them to the statistics of synthetic samples. Lower FID is better, corresponding to more similar real and generated samples. It is found that the FIDs of the generated 0.51 m optical and SAR images are respectively 154.7532 and 53.0067. The values are quite large, which indicates that the model performs badly. However, the reconstructed images seem very

Table 1 FIDs of different datasets

	0.51 m single-pol GF-3	6 m single-pol UAVSAR	10 m single-pol UAVSAR	6 m full-pol UAVSAR
Optical	154.8	106.4	138.4	85.6
SAR	53.0	56.0	64.7	52.8

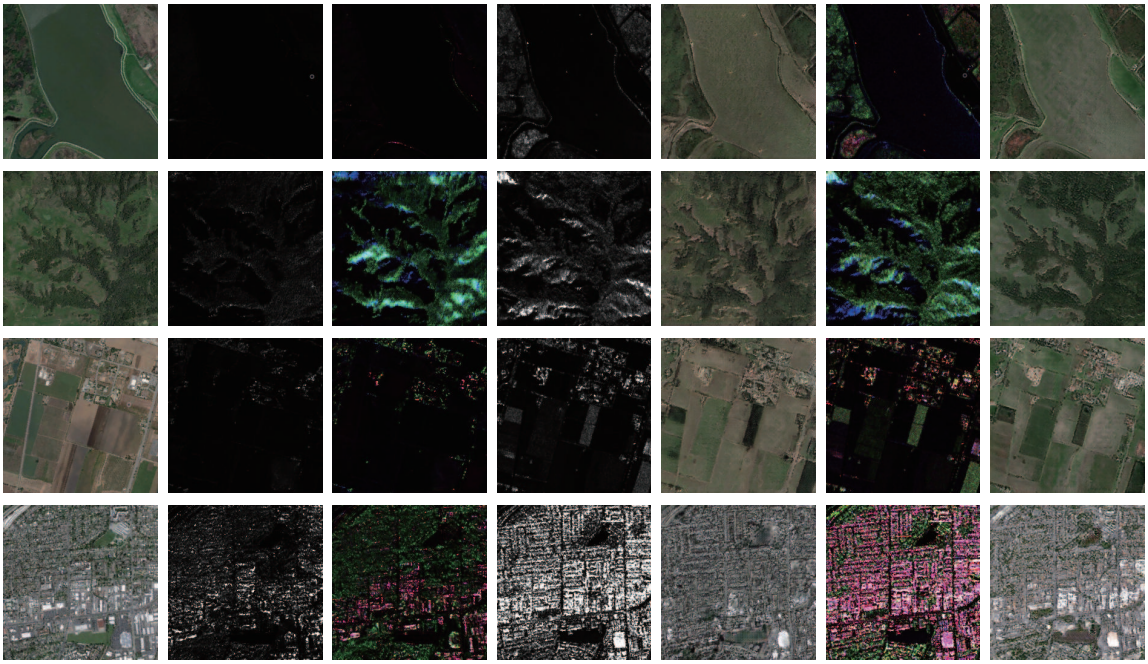

Figure 8 (Color online) Images listed above in each row are (a) the optical ground truth and (b) its translated single-pol SAR image and (c) translated full-pol SAR image, (d) the single-pol SAR ground truth and (e) the optical image translated by single-pol SAR image, (f) the full-pol SAR ground truth and (g) the optical image translated by full-pol SAR image in order. Each row lists a kind of earth surfaces: waters, vegetation, farmlands and buildings.

good from the perspective of human eye. The buildings, farmlands, green areas, etc. in each image are generally well classified. The textures are also allocated properly. Nevertheless, an exception exists. The textures of buildings vary widely and are hard to match one-to-one with ground truths. For large-scale urban scenes, high-frequency parts such as noise and details in ground truths are difficult to learn because those in each sample differ greatly. Our main purpose is to reconstruct their main contours.

The number of samples to calculate the Gaussian statistics (mean and covariance) should be greater than the dimension of the last coding layer, here 2048 for the inception pool 3 layer [5]. Otherwise the covariance is not full rank, which will result in complex numbers and nans by calculating the square root. Here, we use 2048 pairs of test samples to calculate the FID to estimate the capability of the generators.

Table 1 lists the tested FID values for the datasets of different resolutions mentioned above. Randomly select 2048 pairs of samples from each dataset and calculate the corresponding FID value. Repeat three times and use the mean as the ability of our model to learn this kind of dataset. As shown from the middle three columns in Table 1, it indicates that 6 m data performs better than 0.51 m and 10 m data. Generally speaking, the lower the resolution, the less detail the image has and the less difficult it is to be reconstructed. In optical-SAR translation, owing to little texture information of SAR imageries, the FID values differ slightly between different datasets. The 10 m results are not ideal owing to their small features hard to extract.

4.2.2 *Polarization*

The SAR data used in this study so far is all single-pol, i.e., HH or VV single-pol. Full-pol data contains rich polarimetric information. It is worth investigating how the performance might improve if full-pol SAR is used. For simplicity, the Pauli color-coded image is used as a proxy of full-pol SAR data. The derivation of Pauli data is provided in Appendix B.3.

We carry out an experiment to train our model with full-pol and single-pol images in the same region

Table 2 Decreasing FID with increasing the number of samples (for the case of 6 m full-pol UAVSAR in Table 1)

	500	1000	2048	3000	4000	5000	6000	7000	8000	9000	10000
Optical	125.0	102.9	85.6	81.2	77.9	75.9	74.8	74.1	73.4	72.7	72.1
SAR	86.9	68.8	52.8	49.4	46.8	45.9	44.5	43.2	42.5	42.0	41.9

respectively, and compare the translation performance. As shown in Table 1, it indicates that 6 m full-pol data have the best performance, especially the reconstructed optical images are much better than those from 6 m single-pol data. Four examples of different kinds of earth surfaces, waters, vegetation, farmlands and buildings are shown in Figure 8, respectively. Apparently, the optical images translated from full-pol SAR images are more vivid and realistic than those from single-pol images.

Figure B3 in Appendix B further investigates some interesting cases, and note that these buildings are all easily observable in the full-pol image but not in the single-pol one. This is mainly because of the imbalance of scattering power distribution over different polarization channels. The optical image translated by the full-pol SAR image appears to be much more realistic and closer to true image. Apparently, it is benefited from the additional rich information conveyed in the full-pol SAR image.

Note that FID could be further reduced by increasing the number of samples. As shown in Table 2 for the case of 6 m full-pol UAVSAR, its FID could be reduced to 72 for optical and 42 for SAR if given 10000 samples.

4.3 Comparison with existing translation networks

To evaluate the performance of the proposed method in the context of existing image translation approaches, here, we compare it with widely-used CycleGAN [2] and Pix2Pix [1] using the 0.51 m GF-3 dataset. These two networks are often used for image translation in the optical image domain, and as a benchmark for the experiment comparison. Note that the CycleGAN implemented here shares the same network structure with Pix2Pix, but is trained with the cyclic loop strategy. In order to ensure the fairness of comparison, the discriminators and the receptive fields of the generators are the same. The number of the generators' layers and that of the total trainable parameters are the same. The parameters of the network are randomly initialized. To remove the slight dependence of training result on the parameter initialization, each network is repeatedly trained for 3 times with the same data and then the best result is chosen.

In Figure 9, four representative pairs of different earth surfaces are selected. It can be found that in the first two rows, the buildings reconstructed by our method are more natural. In the third and fourth rows, farmlands, roads and vegetation reconstructed by CycleGAN are similar to the proposed method.

In Table 3, the three metrics PSNR, SSIM and FID are all employed respectively for SAR and optical images. PSNR is the inverse of the sum of pixel difference between the reference image and the measured image and SSIM is a metric to evaluate image similarity from brightness, contrast and structure aspects. The larger the value of PSNR or SSIM, the more similar the two images are. FID is used to measure the distance from the distribution of generated samples to that of real world samples, and the smaller FID, the more analogous the two datasets are. On the 0.51 m and 6 m single-pol datasets, our proposed method outperforms CycleGAN and Pix2Pix in the other three indicators, especially the FID score improves greatly. Note that for the cases of 10 m UAVSAR and 6 m full-pol UAVSAR datasets, it is generally better than the other two methods.

Note that the selected quantitative metrics can only be used as a general reference of image generation performance. In some cases, it may not faithfully and precisely reflect the actual visual appearance of the generated image. Two cases are given in Figure 10. The metrics of these cases are also provided above the corresponding images. As we can see from the result, some cases appear visually better but are measured with slightly lower metrics. For example, for the images a, b and c selected from 6 m full-pol UAVSAR dataset, the optical image b translated by CRAN appears to be better than the image c generated by CycleGAN, but the values of SSIM and PSNR are actually smaller than the latter; for the images d, e, f taken from 10 m single-pol UAVSAR data, similarly, e generated by CRAN appears visually better than f, but the latter has a larger PSNR value.

4.4 Enhancement with unsupervised learning

Finally, we explore the possibility of further refining the network with unsupervised learning. Note that the unsupervised training starts from the network pre-trained with co-registered images. Then we can

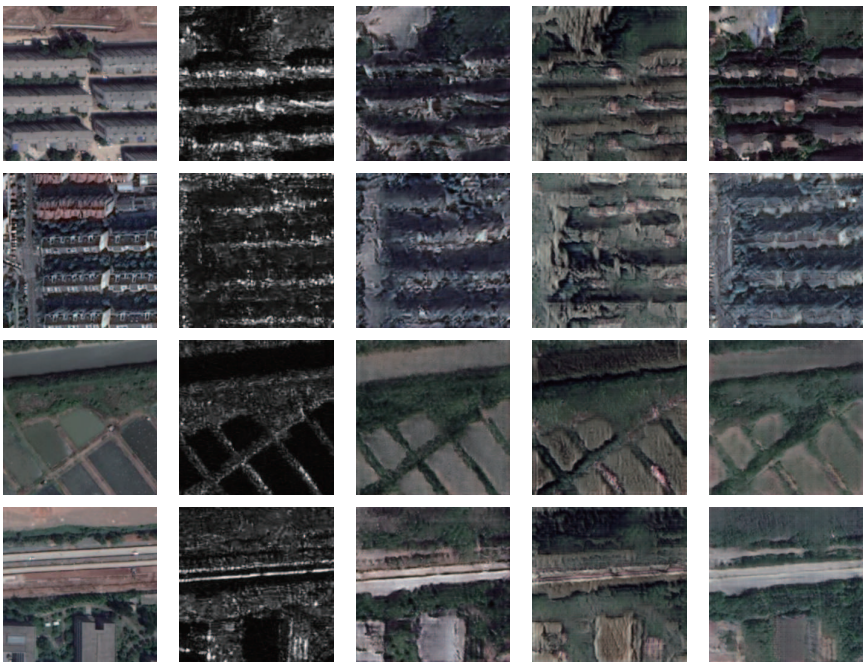

Figure 9 (Color online) Comparison of SAR-optical translation by different methods. Images in each row from left to right are (a) the real optical image, (b) the input SAR image, (c) its translated optical image by CycleGAN, (d) the translated optical image by Pix2Pix and (e) the translated optical image by CRAN. Each row lists a kind of earth surfaces: buildings, buildings, farmlands and roads.

Table 3 Result comparisons of different methods with different datasets using different evaluation methods[a)]

Dataset	Method	SSIM		PSNR		FID	
		SAR	Opt.	SAR	Opt.	SAR	Opt.
0.51 m single-pol GF-3	CycleGAN	0.2535	0.2656	15.7171	14.9675	62.1420	185.3181
	Pix2Pix	0.2194	0.2317	15.4978	14.4686	77.6901	212.5304
	CRAN	**0.2595**	**0.2799**	**15.9172**	**15.5820**	**53.0067**	**154.7532**
6 m single-pol UAVSAR	CycleGAN	0.3585	0.3005	19.5424	16.1030	50.5496	132.1710
	Pix2Pix	0.3407	0.3081	19.6044	15.7463	**48.5541**	**99.7782**
	CRAN	**0.3640**	**0.3092**	**20.2907**	**16.1323**	56.0201	106.3988
10 m single-pol UAVSAR	CycleGAN	0.2879	0.2973	**18.5911**	16.2957	**53.2890**	**113.288**
	Pix2Pix	**0.2917**	0.3072	18.3707	16.0357	63.5519	146.7449
	CRAN	0.2819	**0.3346**	18.3092	**16.4238**	64.7359	138.3651
6 m full-pol UAVSAR	CycleGAN	0.3418	0.3254	18.3431	16.0414	**46.0073**	95.69
	Pix2Pix	0.3716	**0.3308**	**19.5295**	16.0421	65.1980	94.9724
	CRAN	**0.3768**	0.3109	19.2188	**16.1489**	52.7645	**85.5704**

a) 'SAR' column compares SAR ground truths and translated SAR images, while 'Opt.' column compares optical ground truths and corresponding translated optical images.

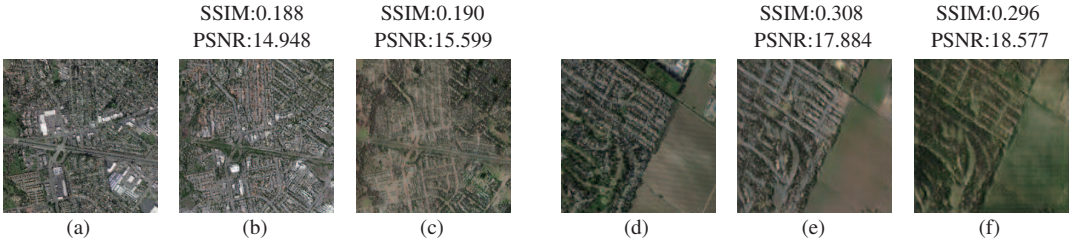

Figure 10 (Color online) Comparison of SAR-optical translation by different methods to verify the not exactly correctness of evaluation metrics. The left three columns are chosen from 6 m full-pol UAVSAR dataset: (a) optical ground truth, (b) translated optical image by CRAN, (c) translated optical image by CycleGAN; the right columns are chosen from 10 m single-pol UAVSAR dataset: (d) optical ground truth, (e) translated optical image by CRAN, (f) translated optical image by Pix2Pix.

Table 4 FIDs of results by supervised and unsupervised learning

	Supervised learning	Unsupervised learning
Optical	107.8	88.9
SAR	58.1	41.2

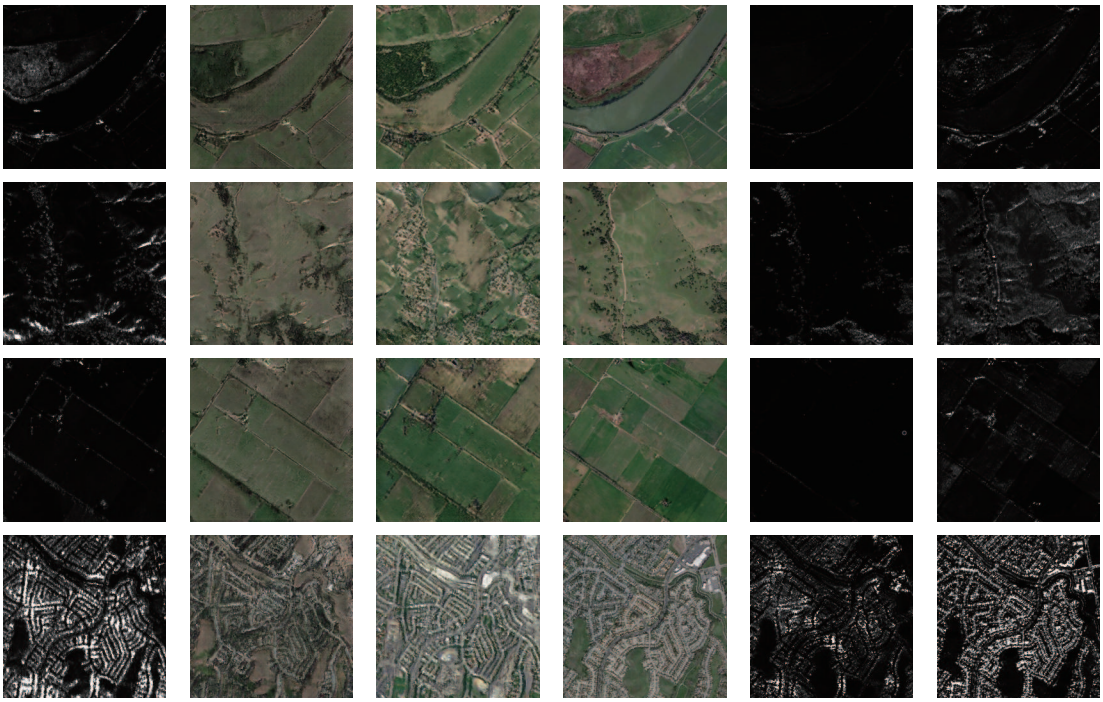

Figure 11 (Color online) Translated images further refined with unsupervised learning. Images in each row from left to right are (a) the input SAR image, (b) the translated optical image and (c) the further refined optical image by unsupervised learning, (d) the input optical image and (e) its translated SAR image and (f) the further refined SAR image by unsupervised learning. Each row lists a kind of earth surfaces: waters, vegetation, farmlands and buildings.

feed the SAR or optical images to be tested to the network and train them using large volume of unpaired optical or SAR images, whose distributions of earth surfaces are similar to those of data used by the pre-trained model. Note that if the distributions of earth surfaces vary greatly or buildings are too few, the ambiguity of reconstructed surfaces can be introduced, and especially buildings can be easily overwhelmed by vegetation. Compared with supervised learning, in which only the prior knowledge from pre-training can be utilized, the model of unsupervised learning can also dynamically learn something new from the extended dataset and refine the results through iterations.

The major experimental procedures are as follows.

• Randomly select n pairs of optical and SAR images outside the dataset to be tested. Ensure that the earth surfaces are evenly distributed (slightly more buildings for their difficulty to be reconstructed);

• Feed the N test SAR (optical) images and the n optical (SAR) images to the unsupervised network, train until the early stop and save the translated optical (SAR) images;

• Check the results and quantitatively evaluate them with those by supervised learning.

As shown in Table 4, it indicates that the translation results are greatly improved with unsupervised learning. The results further refined with unsupervised training are shown in Figure 11, where we can see that the refined results are more vivid and realistic. However, waters in SAR images do not differ from those farmlands greatly, which results in the imperfect reconstruction of waters.

4.5 Computational cost

In this subsection, the computational performances of the three translation networks, including the computational complexity and the speed of processing images per second, are analyzed. Note that for neural networks, the computation cost is about the same for training and inference per image. Thus, only training performance is analyzed here.

Table 5 Number of trainable parameters and operations in the three translation networks

Model	Number of parameters	Number of operations/FLOPs
CycleGAN generator	113.73 M	152.39 G
Pix2Pix Generator	107.16 M	89.50 G
CRAN Generator	107.49 M	79.41 G
Discriminator	5.35 M	6.53 G

Table 6 FCN-scores for the two segmentation scheme

Scheme	Per-pixel accuary	Per-class accuary	Class IOU
SAR-Segmentation	0.5988	0.4927	0.3742
Translated Optical-Segmentation	0.5052	0.4395	0.2923

Deep learning models are intensive in resource consumption, mainly measured by the number of trainable parameters and the number of float operations. All the translation networks used are reciprocal here. It should be noted that only the convolutional layers are considered, and the trainable parameters and operations generated by LeakyReLU and batch normalization are ignored. Table 5 indicates that the numbers of parameters are almost same, but the number of operations in CycleGAN is approximately twice that of Pix2Pix and CRAN, owing to the additional cyclic loop.

The networks are all implemented on TensorFlow and run on Ubuntu server with 4 Titan X. Here we compare how many pictures can be processed per second respectively by the three methods. From Figure B5 in Appendix B we can find the following.

• For the same method, the training speed using 4 GPUs is approximately 2–3 times of that using 1 GPU. This is due to the communication overhead and some part of computation that cannot be run distributedly.

• The speed of CRAN and Pix2Pix is much faster than CycleGAN, which agrees with the analyses given in Table 5.

5 Discussion

This section is mainly about practical applications of CRAN. We make a contrastive experiment: (1) straightforwardly segment SAR images into regions, (2) convert SAR images to optical ones and then segment optical images. In our early work, we registered SAR and optical images, and annotated them. Before the experiment, we divide the whole dataset into three parts, training, validation and test data, and augment each seperately to improve the pixel ratio of buildings. For the scheme 1, we train pairs of SAR images and segmentation maps using the pretrained DeepLabv3+ model [22]. For the scheme 2, we retrain CRAN with new augmented SAR and optical patches and then convert all SAR images to optical ones. Then we train the pretrained DeepLabv3+ model to convert the translated optical images to segmentation maps.

We evaluate the segmentation results and FCN-scores [1] in Table 6 shown that the scheme 1 performs better but not much better. The dense buildings and the mixed areas of buildings and vegetation in SAR images are cluttered, so that they are not strictly spaced apart in the translated optical images, resulting in terrible subsequent segmentation results. As shown in Figure 12, segmentation results of the scheme 2 are less likely to be judged as background than those of the scheme 1, and specifically, the total number of background pixels of the scheme 2 is 2446019 less than that of the scheme 1, which is an increase of 2.4685%.

In addition to translating images at the terrain surface level, CRAN can also be applied to synthesize specific targets, such as airplanes. We acquire the appropriate amount of TerraSAR-X airport data with good registration and use CRAN for training. The selected scene is located in the aircraft cemetery, and the number and types of airplanes are large. As shown in Figure 13, although the SAR images themselves cannot be reconstructed as optical ones with details as same as optical ground truths, but the translated images still seem to be sufficient for the task of target recognition.

Besides, CRAN can also be used for registering unpaired SAR and optical images. Once the SAR image is translated into an optical image, it is easy to use some matching method to register two optical images in the same image domain. Certainly, we can also get the difference between the original SAR

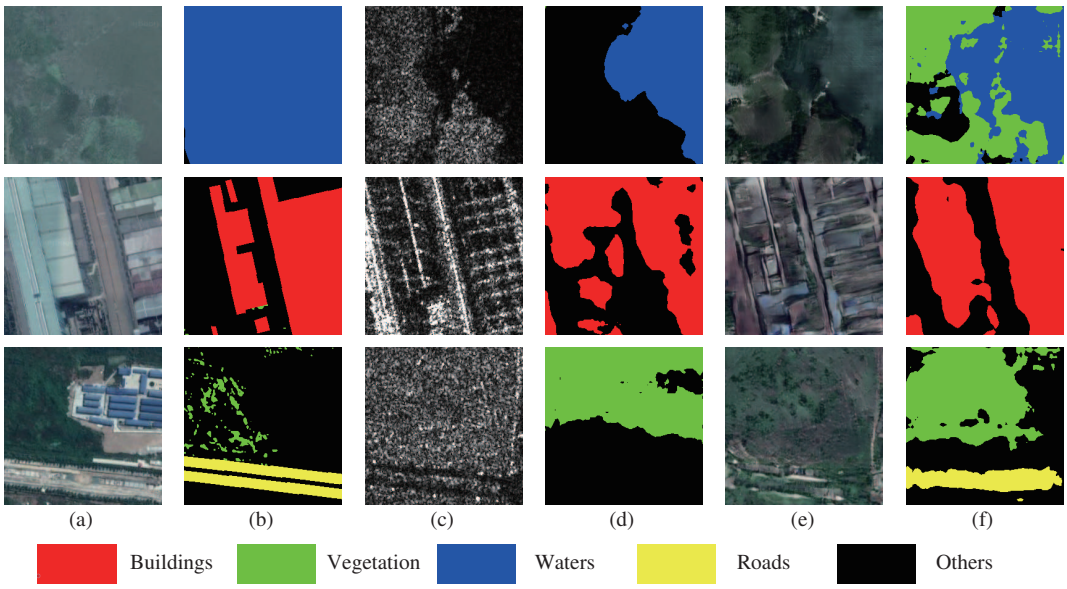

Figure 12 (Color online) A contrastive experiment on SAR image segmentation. Images in each row from left to right are (a) optical ground truth, (b) segmentation ground truth, (c) input SAR image, (d) map segmented from (c), (e) optical image generated from (c) by CRAN and (f) map segmented from (e). For segmentation maps, colors red, green, blue, yellow, and black represent buildings, vegetation, waters, roads, and others respectively.

Figure 13 (Color online) Airplane synthesis by CRAN. Images in each row from left to right are optical ground truth ((a1), (a2)), input SAR image ((b1), (b2)), and translated optical images ((c1), (c2)).

image and the optical image by comparing the two optical images for SAR change detection.

6 Conclusion

For the purpose of assisted interpretation of SAR imagery by ordinary people, this paper proposes an image translation network architecture for reciprocal translation between SAR and optical remote sensing images. In order to evaluate the translated images from the perspective of human visual perception, the quantitative metric FID is employed. For low-resolution (6 m, 10 m) UAVSAR dataset, the reconstructed images appear very similar to the true data and the corresponding FID is low. For high-resolution (0.51 m) GF-3 dataset, the reconstructed results appear reasonable but not exactly capture the geometric features of certain built-up objects such as high-rise buildings. Under the same condition, the proposed network outperforms conventional image translation networks such as CycleGAN and Pix2Pix. Results also show that full-pol SAR image is preferable as input for translation because certain objects are not observable in single-pol SAR images. It is also confirmed that the network does not perform well if generalized across different SAR platforms. Next, we demonstrate that unsupervised learning could further improve the performance of a translator initially trained with a small number of co-registered image pairs which points the right direction towards general application of assisted SAR image interpretation. Finally, we make several experiments to explore the real applications of CRAN on image segmentation and airplane

synthesis and achieve good results.

Acknowledgements This work was supported in part by National Key R&D Program of China (Grant No. 2017YFB0502703) and Natural Science Foundation of China (Grant Nos. 61822107, 61571134).

Supporting information Appendixes A and B. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Isola P, Zhu J, Zhou T, et al. Image-to-image translation with conditional adversarial networks. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 5967–5976
2 Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV), 2017. 2242–2251
3 Jin K H, McCann M T, Froustey E, et al. Deep convolutional neural network for inverse problems in imaging. IEEE Trans Image Process, 2017, 26: 4509–4522
4 Zhu J Y, Zhang R, Pathak D, et al. Toward multimodal image-to-image translation. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 465–476
5 Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 6626–6637
6 Byun Y, Choi J, Han Y. An area-based image fusion scheme for the integration of SAR and optical satellite imagery. IEEE J Sel Top Appl Earth Observ Remote Sens, 2013, 6: 2212–2220
7 Garzelli A. Wavelet-based fusion of optical and sar image data over urban area. Int Arch Photogrammetry Remote Sensing Spatial Inf Sci, 2002, 34: 59–62
8 Fan J, Wu Y, Li M, et al. SAR and optical image registration using nonlinear diffusion and phase congruency structural descriptor. IEEE Trans Geosci Remote Sens, 2018, 56: 5368–5379
9 Liu J, Gong M, Qin K, et al. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. IEEE Trans Neural Netw Learn Syst, 2018, 29: 545–559
10 Merkle N, Auer S, Muller R, et al. Exploring the potential of conditional adversarial networks for optical and SAR image matching. IEEE J Sel Top Appl Earth Observ Remote Sens, 2018, 11: 1811–1820
11 He W, Yokoya N. Multi-temporal sentinel-1 and -2 data fusion for optical image simulation. ISPRS Int J Geo-Inf, 2018, 7: 389
12 Schmitt M, Hughes L H, Zhu X X. The sen1-2 dataset for deep learning in sar-optical data fusion. In: Proceedings of ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences IV-1, 2018. 141–146
13 Wang P, Patel V M. Generating high quality visible images from SAR images using CNNs. In: Proceedings of 2018 IEEE Radar Conference (RadarConf18), 2018. 0570–0575
14 Enomoto K, Sakurada K, Wang W, et al. Image translation between sar and optical imagery with generative adversarial nets. In: Proceedings of IGARSS IEEE International Geoscience and Remote Sensing Symposium, 2018. 1752–1755
15 Wang L, Xu X, Yu Y, et al. SAR-to-optical image translation using supervised cycle-consistent adversarial networks. IEEE Access, 2019, 7: 129136–129149
16 Li Y, Fu R, Meng X, et al. A SAR-to-optical image translation method based on conditional generation adversarial network (cGAN). IEEE Access, 2020, 8: 60338–60343
17 Fuentes R M, Auer S, Merkle N, et al. SAR-to-optical image translation based on conditional generative adversarial networks-optimization, opportunities and limits. Remote Sens, 2019, 11: 2067
18 Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer-assisted Intervention. Berlin: Springer, 2015. 234–241
19 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, 2014. 2672–2680
20 Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. 2017. ArXiv:1701.07875
21 Cozzolino D, Parrilli S, Scarpa G, et al. Fast adaptive nonlocal SAR despeckling. IEEE Geosci Remote Sens Lett, 2014, 11: 524–528
22 Chen L C, Zhu Y, Papandreou G, et al. Encoderdecoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 801–818