SCIENCE CHINA
Information Sciences

• **Supplementary File** •

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*

Appendix A Datasets

Table S1 lists several major differences of SAR and optical imaging mechanisms and the corresponding distinct phenomena. As a result, the information contents in SAR and optical image are partial overlapped and partial exclusive, which means that only part information is observed by both sensors and each sensor observes other information that is not observable by the other sensor.

Table S1 Differences between SAR images and natural optical images

	Optical images	SAR images	SAR unique phenomena
Wave band	Visible light band	Microwave band	Discontinuity, Scintillation
Focusing mechanism	Real aperture	Coherent synthetic aperture	Speckle noise
Projection scheme	Elevation-Azimuth	Range-Azimuth	Layover, foreshortening, shadowing
Resolution	Proportional to range	Invariant to range	No perspective distortion
Data format	Color, intensity	Phase, amplitude, polarization	Multi-channel, complex

The information of GF-3 SAR and UAVSAR data used in our experiments is listed in the following Table S2. Optical data used is downloaded from Google Map around November, 2018, with pixel resolution 0.51 m for GF-3 SAR data and 1.02 m for UAVSAR data respectively.

Table S2 Information about the two datasets employed for experiments

		UAVSAR	GF-3
SAR	Resolution (m)	6	0.51
	Polarization	Quad-Pol	HH or VV
	Angle of incidence (°)	90	40.6642, 36.0820
	Acquisition mode	PolSAR	SL
	Frequency band (MHz)	80	240
	Day of acquisition	2010-04-09, 2013-05-13, etc.	2017-01-02, 2016-08-15
	Location	California, US	Wuhan and Hefei, China
Optical	Resolution (m)	1.02	0.51
	Day of acquisition	2018-11-25	2018-06-05, 2018-05-28
	Geographic coordinate system	WGS 84	WGS 84

UAVSAR radar system is an L-band polarimetric instrument developed by NASA. As shown in Figure S1, UAVSAR data used here mainly consist of five types of earth surfaces, buildings, vegetation (mountains are usually covered with trees and classified as vegetation here), farmlands, waters and deserts. It has pixel resolution of about 6.2 m× 4.9 m. The samples are 256×256 patches

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

cropped from the original large SAR and optical images without any overlapping, which avoids the direct correlation between the training and test samples. Then we acquire a total of 12394 pairs of co-registered samples.

GF-3 satellite is China's first C-band multi-polarization SAR satellite. Two large scenes of GF-3 images are used in the study with a resolution of 0.51 m. The dataset contains different urban/suburban regions. In Figure S2, the GF-3 SAR image after geocoding is shown on the left. It mainly contains five terrain surfaces, i.e., buildings, roads (highways or overpass), vegetation, waters (lakes, rivers or seas) and farmlands. Buildings can be further divided into low-rise and high-rise buildings.

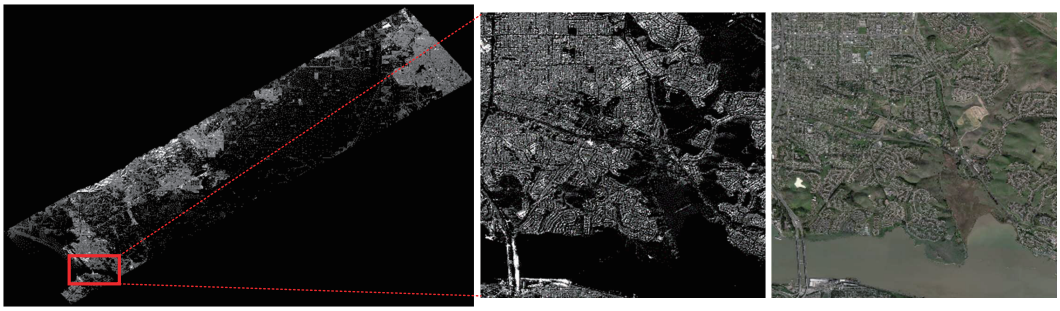

Figure S1 (a) UAVSAR image acquired in California, US. (b) Zoomed region. (c) Corresponding optical image of the zoomed region.

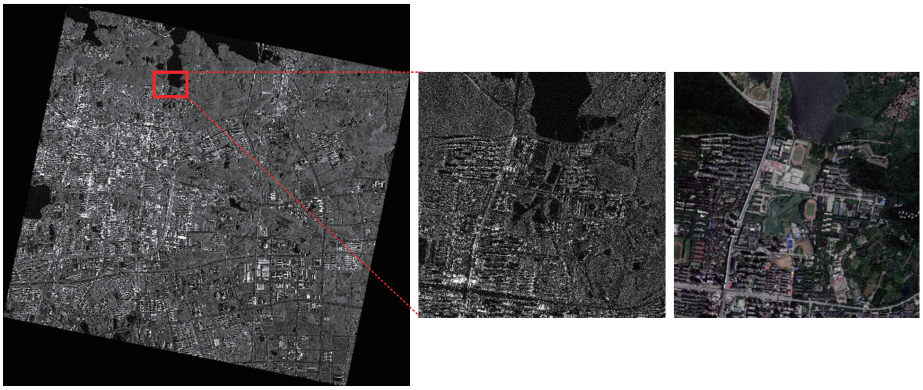

Figure S2 (a) A large SAR image with a 0.51 m resolution in Hongshan District, Wuhan City, Hubei Province, China. (b) Zoomed region. (c) Corresponding optical image of the zoomed region.

Appendix B Experiment

Appendix B.1 Training strategy

Stochastic gradient descent algorithm with adaptive moment estimation (Adam) can be used to train the two translators/discriminators simultaneously. Following GAN training strategy, one iteration consists of the following steps (see Figure S3).

(a) Forward Pass — Weights of translators and discriminators are randomly initialized. A mini-batch of SAR images are then sent to the translator A to synthesize fake optical images, while a mini-batch of optical images are sent to the translator B to synthesize fake SAR images. Next, the fake and real optical images are sequentially sent to the same discriminator A, which generates two probability maps respectively. The fake and real SAR images are sent to the discriminator B and the discriminator B also generates corresponding probability maps.

(b) Backward Pass — The two probability maps of optical images are compared in the loss to optimize the discriminator A, while those of SAR images are compared to optimize the discriminator B. The sigmoid function is selected as the activation function for the discriminator, which functions as a binary classifier. The discriminator is trained to distinguish the fake as 0 and the real as 1. The discriminator classifies patches of the input image separately. This not only limits the receptive field, but also provides more samples for the training. Both of the two losses are also added as the GAN loss for the translators, which have to maximize them. That means the aim of both the translators is to generate sufficiently-realistic images to fool the discriminators. The real and fake ones are also compared directly to ensure the positional mappings of targets are correct. Thus, the joint losses are applied as the final loss function of the two translators. Then the backpropagation is applied to adjust the trainable parameters in the two translators simultaneously.

The forward process alternates with the backward process. The batch size is set as 1. The technique of GPU parallel acceleration with 4 NVIDIA Titan X is employed, which means four pairs of SAR and optical images are used to train the network each iteration simultaneously. After the gradients of the four threads are all calculated, the mean gradients are used to update the optimizers. The backward pass is a single thread. After finishing the back propagation, another four pairs of images are sent in. Traversing all the training dataset is considered as an epoch. Then reshuffle the images and traverse next epoch.

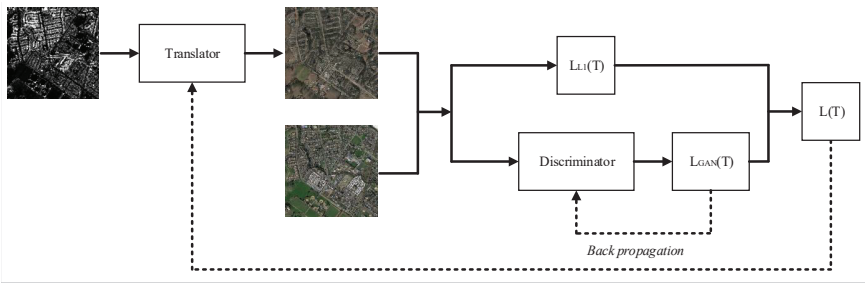

Figure S3 The conceptual process of training the adversarial networks. The left image is the real SAR image, the upper right is the synthesized optical image and the lower right is the real optical image.

Appendix B.2 Frechet inception distance

IS and FID are usually used to quantitatively evaluate the quality and variety of images generated by GANs. Both of them encode the input image to a feature vector by using the inception network, shown in Figure S4, which functions as the human visual perception. If the two images are identical, their encoded feature vectors should be the same. Different from IS, FID uses the statistics of real world samples and compares them to the statistics of synthetic samples. FID between the Gaussian distribution with mean and covariance (m_1, C_1) and the Gaussian distribution with (m_2, C_2) is defined as $||m_1 - m_2||_2^2 + \mathrm{Tr}(C_1 + C_2 - 2(C_1C_2)^{1/2})$.

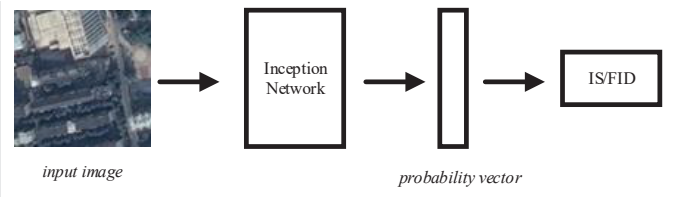

Figure S4 The conceptual process of calculating IS/FID.

Appendix B.3 Polarimetric SAR

The basic form of polarimetric SAR data is the Sinclair scattering matrix [**?**] with horizontal and vertical polarisations. It can be expressed as a 2×2 matrix containing four components S_{HH}, S_{HV}, S_{VH} and S_{VV}, where H, V respectively denotes the horizontal and vertical polarisations. S_{HH} and S_{VV} are co-polarized components; S_{HV} and S_{VH} are cross-polarized components. Different polarimetric channels contain partial electromagnetic information. Some targets may be imaged more clearly in the cross-polarized channels than those in the co-polarized channels, and vice versa [**?**].

Then we convert the full polarimetric information into pseudo-color coded images via Pauli decomposition. Pauli decomposition is to decompose the scattering matrix S into different scattering components, i.e., a is the single-bounce surface scattering intensity; b is the dihedral scattering intensity with incidence angle 0°; c is the volumetric scattering; d is all the antisymmetric components of the scattering matrix:

$$a = \frac{S_{HH} + S_{VV}}{\sqrt{2}}, \quad b = \frac{S_{HH} - S_{VV}}{\sqrt{2}}, \quad c = \frac{S_{HV} + S_{VH}}{\sqrt{2}}, \quad d = j\frac{S_{HV} - S_{VH}}{\sqrt{2}}. \tag{B1}$$

The Pauli image is a pseudo-color image coded using the intensities of these three components, i.e.,

$$\mathrm{I} = \left[|S_{HH} - S_{VV}|^2, 4\,|S_{HV}|^2, |S_{HH} + S_{VV}|^2\right]^{\mathrm{T}}/2. \tag{B2}$$

Appendix B.4 Comparison between Single-pol and Full-pol SAR images

Figure S5 further investigates into a few interesting cases. In each case, one building in single-pol image and the corresponding full-pol image is marked correspondingly in each row.

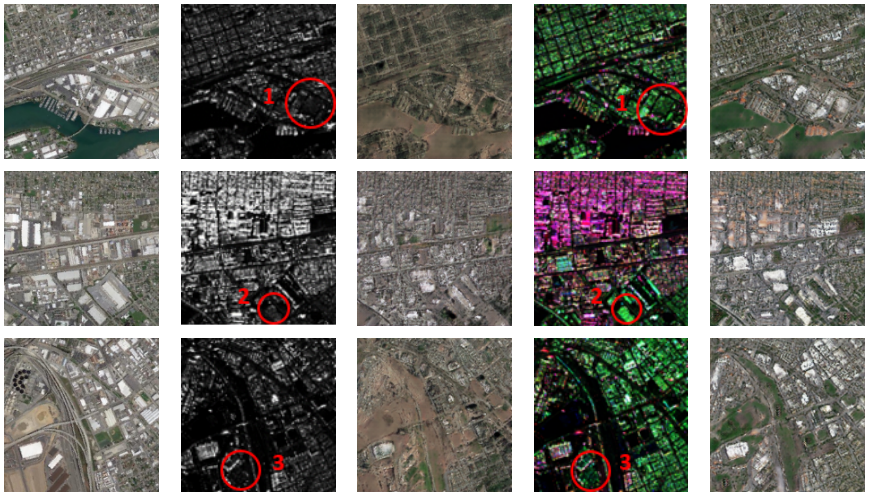

Figure S5 Images in each row from left to right are (a) the real optical image, (b) the real single-pol SAR image and (c) the optical image translated by single-pol SAR image, (d) the real full-pol SAR image and (e) the optical image translated by full-pol SAR image.

Appendix B.5 Generalization to different SAR platforms

Generalization capability is critical to make the proposed method applicable in practical scenarios. One key aspect is generalization to different geographic scenes. From the cases presented in previous subsections, the test samples are acquired from different regions than the training samples, where the low FID has demonstrated that the proposed method can be generalized to different scenes. Another critical test is generalization to different SAR platforms, e.g., a model is trained with data from one SAR platform but used to translate SAR images from another SAR platform.

An experiment is conducted where the model trained using UAVSAR images is used to translate SAR images from UAVSAR, GF-3 and Advanced Land Observing Satellite 2 (ALOS-2) acquired at different regions (Figure S6). Compared to the ground truth, the performance of translation is largely degraded in the case of GF-3 and ALOS-2. The boundaries of different terrain surfaces are smeared. We believe that this is partially attributed to the fact that SAR images from different platforms are not cross-calibrated.

Note that when applied to processing real SAR image, we prefer to process one large image at a time. Although the network is trained and designed to take inputs of 256×256 patches, it is a fully convolutional network and can be directly extended to process larger size images without any modification. Experiments were conducted to verify the performance of the proposed method when used to process large size images.

Appendix B.6 Computational cost

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.

Figure S6 Images across scenes and across sensors reconstructed by the model pre-trained with 6 m UAVSAR images. Images in each row from left to right are (a) the real SAR image and (b) its translated optical image, (c) the real optical image and (a) its translated SAR image. Each row lists a kind of dataset: UAVSAR, GF-3 and ALOS-2.

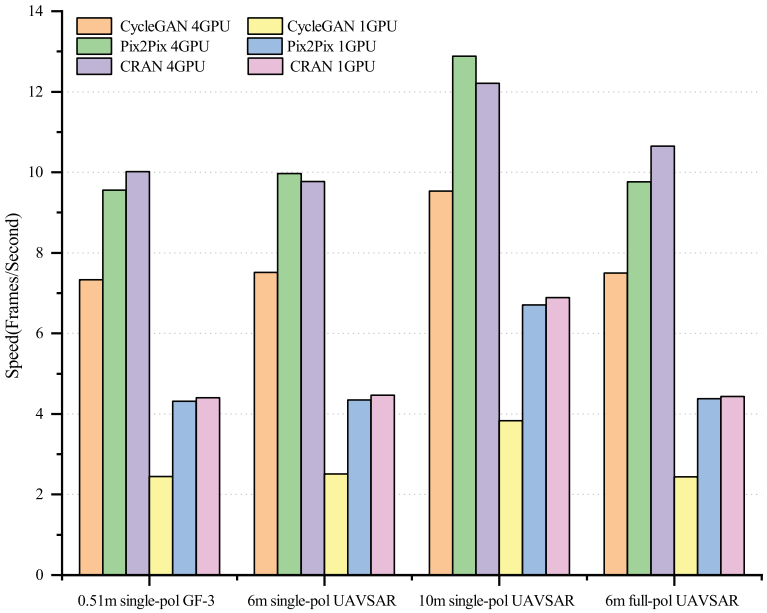

Figure S7 Speed comparison between different methods, different datasets and different numbers of GPUs. On each dataset, the three translation networks respectively run on four GPUs and one single GPU.