

Reciprocal Translation between SAR and Optical Remote Sensing Images with Cascaded-Residual Adversarial Networks

Shilei Fu, Feng Xu, and Ya-Qiu Jin

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

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 are trainable with large-volume co-registered SAR and optical images to translate SAR image to optical version for assisted SAR image interpretation. Reciprocal SAR-Optical image translation is a challenging task because its 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 GF3 and airborne UAVSAR images. Results are presented for datasets of different resolutions and polarizations and compared with other state-of-the-art methods. The Frchet inception distance is used to quantitatively evaluate the translation performance. The possibility of unsupervised learning with unpaired 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, Frchet inception distance

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 meters in 2000s (e.g. Radarsat2), and to sub-meter in 2010s (e.g. TerraSAR-X). Higher resolution greatly enhances our capability to separate individual scatterers of one target, which opens up a new way to deterministically interpret the target image on scatterer-by-scatterer level. Multi-dimension includes multi-polarization, multi-temporal, multi-baselines, multi-frequency, multi-static etc. For example, spaceborne bistatic and multistatic SAR as demonstrated in Zhang et al. (2016a,b) is one of the promising multi-dimension SAR technologies in the next decade. New imaging modalities and novel constellation concepts are still under development (Guarnieri and Rocca (2017); Fuster et al. (2017)). 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 is being acquired daily by numerous radar satellites in orbits.

Due 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 1. Humans visual system is adapted to the interpretation of optical images. SAR image is difficult to interpret by ordinary people. Although SAR images contain rich information about target and scene, such as 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.

Experts of SAR imagery interpretation are often trained by comparing the SAR image side-by-side with the corresponding optical image (e.g. Fig. 1). From such SAR-vs-Optical comparison experiences, experts conclude useful rules which translate between features in SAR and optical remote sensing

Table 1
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

images. Thereafter, they are able to direct interpret a new image from similar SAR sensors. Ideally, such training could be done in computers with artificial intelligence (AI) by leveraging the recent progresses in AI and deep learning technologies.

The major objective of this work is to develop deep learning application with large amount of co-registered SAR and optical images where SAR image 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 translation tool, any person without any background knowledge of radar, could be able to understand the primary information contained in SAR image. This could greatly promote the wide application and usage of future and existing archive of SAR remote sensing imagery. Other potential applications include facilitation of data fusion of optical and SAR images, e.g. translating 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.

Deep learning, in particular convolutional neural networks (CNNs), has revolutionized the computer vision regime since the first successful application of CNN in practical image classification task in 2012 Krizhevsky et al. (2017). It utilized stacked convolution and pooling layers to automatically extract features at different scales via supervised learning. Since 2014, CNN-based approaches have been applied in interpretation of SAR images (Zhu

et al. (2017c)), including the typical tasks such as automatic target recognition (ATR) (Song and Xu (2017); Chen et al. (2016)), earth surface classification (Zhang et al. (2017); Zhou et al. (2016)), speckle reduction (Yue et al. (2018)), change detection (Liu et al. (2018)), etc.

Generative adversarial networks (GANs) (Goodfellow et al. (2014)) are a special type of CNNs which are often used to generate synthesized images using actor-critic scheme. It simultaneously trains a generative CNN, the actor, which tries to generate images as realistic as possible, and a discriminative CNN, the critic, which tries to identify the synthesized images from the real images. Some relevant works which employed GAN-like approaches to analyze SAR images are reviewed in subsection 2.1.

On the other hand, image translation itself is an interesting problem in computer vision. Some interesting approaches in this regard are reviewed in subsection 2.2. Most of these image translation studies deal with the problem of style transfer where the objective is to change the style of image, e.g. from optical to cartoon picture, or from semantic map to optical picture.

However, the problem of translation between SAR and optical images is considered slightly more difficult. It involves with two distinct types of raw sensor data. Table 1 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. A successful translation algorithm should be able to: 1) convert the common part of information content from one sensing modality to the other; and ideally, 2) generate the new information content from learned experiences. The former functionality is similar to style-transfer, while the latter is different. We believe that such cross-modality raw data translation requires a specially designed and fine-tuned network scheme and a large volume of co-registered image pairs as training data.

To illustrate the difficulties in translating across the two sensing modalities, Fig. 1 gives four example image pairs, i.e. low-rise buildings, high-rise buildings, waters and roads. The differences between SAR and optical images are marked in red circles. For low-rise buildings, the gap between houses cannot be distinguished in SAR images. For high-rise buildings, the projection directions of buildings differ in the two sensing modalities. And buildings in SAR images scatter heavily and have severe layover effect with nearby buildings. For waters, ripples in optical images may not appear in SAR im-

ages. For roads, small roads are not nearly noticeable in SAR images due to layovers and shadows.

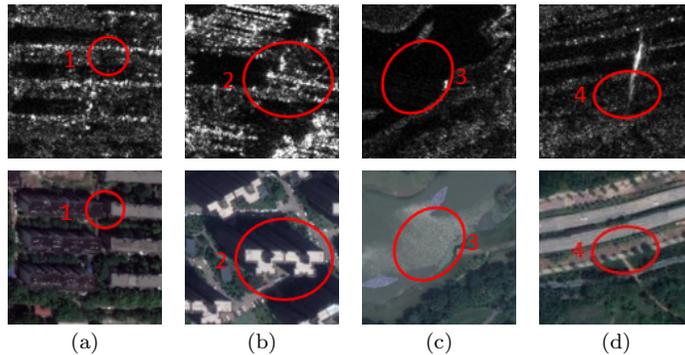


Fig. 1. Distinctions between SAR and optical remote sensing images. The first row lists four example SAR images, and the second row lists the corresponding example optical images. Different qualities of translated optical images are induced by different losses. Each column lists (a) **low-rise buildings**, (b) **high-rise buildings**, (c) **waters** and (d) **roads**.

A SAR-Optical image reciprocal translation GAN architecture is proposed in this paper. It follows the typical image translation GAN architecture (e.g. Isola et al. (2017); Zhu et al. (2017a); Jin et al. (2017); Zhu et al. (2017b)) employing a CNN as the discriminator and a specially-designed network as the generator (translator). The translator uses the multi-scale encoder-decoder CNN as the backbone and incorporates novel multi-scale 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 true sample.

The proposed method is verified on a large volume SAR-Optical image pairs, i.e. 10000 samples of 256256 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. The Frchet inception distance (FID) (Heusel et al. (2017)) is used as the quantitative measure of the similarity between the reconstructed and the true image. For low-resolution (6m, 10m), the reconstructed result ap-

pears to be very similar to ground truths and the FID indicates very high similarity. For high-resolution (0.5m), the reconstructed result appears to be reasonable but not exactly captures the fine geometric features of manmade objects such as buildings and roads. This indicates that the mutual-exclusive part of information content in two sensing modalities becomes significant at higher resolutions. Overall, the translated high resolution optical image can partially serve the purpose of assisted interpretation of SAR images.

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 sensor.
- 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, input image scales etc.
- An extension towards unsupervised learning is tested with the CycleGAN loop (Zhu et al. (2017a)). Results demonstrate that using large volume of unpaired SAR and optical images, the performance can be further improved.

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. Finally in Section 5, conclusions are drawn and the future perspectives of application are discussed.

2. Related works

2.1. GANs in SAR image analysis

GANs or generative neural networks have been applied in various tasks in SAR image analysis. For instance, Song and Xu (2017) proposed zero-shot learning scheme for SAR ATR which is able to extract target orientation

and orientation-invariant intrinsic features via training SAR generative networks. Ma et al. (2018) proposed a novel super-resolution method namely dense residual GAN and utilized the memory mechanism to extract hierarchical features for better reconstruction of remote sensing images. Guo et al. (2017) employed GAN to synthesize the desired SAR images according to angle information and its feasibility was validated by comparison with real images and ray-tracing results. Gao et al. (2018) introduced the multi-classifier to the discriminator so that the images labels of the Moving and Stationary Target Acquisition and Recognition (MSTAR) were made the most of to synthesize realistic SAR targets. Hughes et al. (2018) produced hard-negative samples using GAN and validated its applicability to improve training in data sparse applications such as SAR-optical image matching. Ao et al. (2018) proposed a Dialectical GAN based on Spatial Gram matrices and a WGAN-GP framework to transfer low-resolution Sentinel-1 to high-resolution TerraSAR-X images.

2.2. Image translation

Image translation is conversion between two types of images. It constructs an intermediate latent space, which maps the two image fields X_1 , X_2 to the latent space Z through two encoders and then reconstructs X_2 , X_1 respectively through decoders (Liu et al. (2017)). Currently, the mainstream image translation methods are based on GANs. Compared with traditional loss functions such as L2 and L1 norms, GANs can generate sharper and more realistic images. Slightly different from conventional GANs, image translation networks are conditional GANs (Mirza and Osindero (2014)) and require the original image as input. Existing studies of image translation can be mainly divided into two categories.

The first category employs end-to-end image mapping strategy which directly converts the original image to the translated domain. Taigman et al. (2016) translated optical images to cartoon images using the domain transfer network. Isola et al. (2017) proposed the Pix2Pix network to translate optical aerial image to maps. It is based on U-Net architecture (Ronneberger et al. (2015)). Zhu et al. (2017a) proposed CycleGAN for unpaired image-to-image translation and changed the properties of targets in the images, such as the animal categories and the seasons of the scenes. Liu et al. (2017) made a shared-latent space assumption based on CycleGAN. Chen and Koltun (2017) synthesized photographic urban scenes with semantic segmentation maps using a cascaded refinement structure.

The second category employs the style transfer strategy. It assumes that the images in two different domains share the same large-scale features, but have distinct small-scale features. Thus, it first separates the style from the contents of images and then replaces the style of the original image with the translated style. It assumes that the contents of an image are conveyed in the low-frequency large-scale edges while the styles are represented by the high-frequency small-scale textures. Gatys et al. (2016) proposed the Gram Matrix, the correlation coefficient matrix between each feature maps as extracted by CNNs, to represent the style of images. Johnson et al. (2016) trained a feed-forward network to solve the optimization problem proposed by Gatys in real-time rather than adjusting the input according to the target. FaderNets, proposed by Lample et al. (2017), trained the discriminator to disentangle some specific style attributes and invariant intrinsic features from the encoded representation.

2.3. Fusion of SAR and optical images

Many works have been carried out in the multi-sensor image fusion regime (Byun et al. (2013)). 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 (MR method) based on the magnitude ratio of the two images are often employed (Byun et al. (2013)). The fused images are well-defined and diversely textured. Garzelli (2002) leveraged co-registered SAR image to improve the quality of optical images, which extracts specific information from SAR image and complements with optical image so that the targets could appear clearer.

Another type of fusion work is SAR and optical image registration. The focus is to explore the consistent features between the two sensing modalities. Fan et al. (2018) 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. (2018) 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. (2018) synthesized artificial SAR-like patches from optical images and matched them with the true SAR patches utilizing NCC, SIFT or BRISK. Especially, Lius 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 image to generate images of different observation time. He and Yokoya (2018) used a conventional conditional GAN to generate the optical images from optical images at a different date with the aid of a SAR image acquired at the both dates. It also tried to directly generate optical image from a single SAR image but found that the existing GANs failed to do so. Schmitt et al. (2018) trained the network Pix2Pix on a large number of SEN1-2 patch pairs and got good predicted optical image patches. Some earlier attempts (Wang and Patel (2018); Enomoto et al. (2018)) try to converted coarse-resolution SAR or simulated SAR image to visible images using conditional GANs but results show that terrain objects such as buildings cannot be translated.

3. SAR-Optical Reciprocal Translation Network

3.1. Translation framework

The proposed framework is shown in Fig. 2. 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 multi-scale convolutional encoder-and-decoder network as the translator (generator) vs. a convolutional network as the discriminator. The translator takes in 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 translated image.

The discriminator is a conventional CNN for binary classification task. The translator has multi-scale 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 the following subsections.

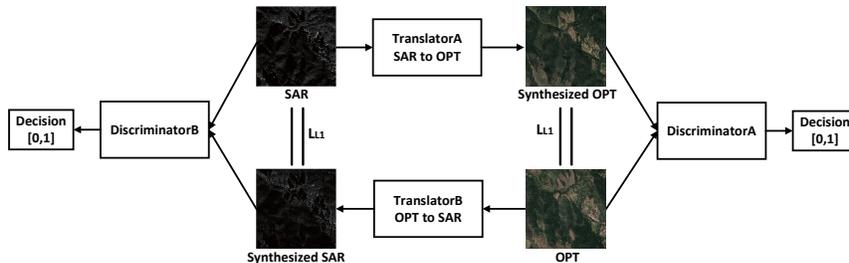


Fig. 2. Schematic diagram of the translation network (Isola et al. (2017)) during training. A pair of translators are trained together. Each translator consists of an encoder and a decoder. The two discriminators are trained separately. SAR, OPT, Synthesized OPT and Synthesized SAR respectively represent 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.

3.2. Network architecture

Fig. 3 shows the architecture and parameters of the translator network, which is named as CRAN. It follows the main structure of the U-net (Ronneberger et al. (2015)) and the Pix2Pix (Isola et al. (2017)) 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 encoder to 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 Zhu et al. (2017b). On the other hand, the network employed in Jin et al. (2017) contains a single skip connection from input to the last stage of U-net output which, according to our experiences, has low capability in generating image details than the cascaded pattern as used in this paper. We believe that such multi-scale 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 encoders 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 decoders 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.

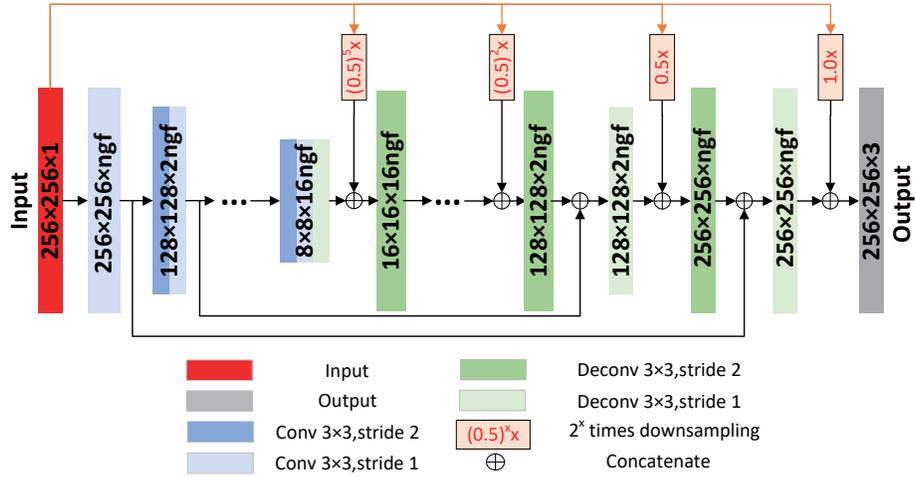


Fig. 3. Translator network architecture with cascaded-residual connections. The input data size is 2562561 and the output data size is 2562563. The first two numbers represent the size of the feature maps and the third number represents the channel of the feature map. The concatenation from the encoder and the input to the decoder is signified by lines with arrows.

Regarding the hyperparameters, in the translator, the convolutional kernel is 33, the encoder and the decoder each have 12 layers, and the receptive field per pixel of the input image is 191191. In the discriminator, the kernel is 44, and it has 5 layers and the receptive field is 7070. 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 3232 matrix and every value corresponds to a 7070 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 generators weights is approximately 53.75 million and that of the discriminator is 2.76 million. The network architectures are depicted in detail in Fig. 3 and Fig. 4.

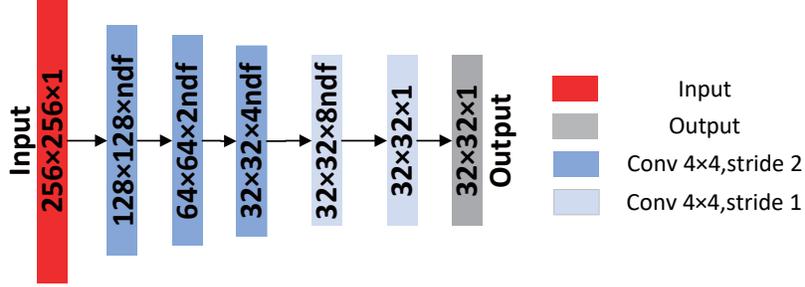


Fig. 4. Discriminator network architecture. The input data size is 2562561 and the output probability map size is 32321. The first two numbers represent the size of the feature maps and the third number represents the channel of the feature map.

3.3. Loss functions

Loss functions are critical for training of the networks. The discriminator is trained with a binary classification log-loss (Goodfellow et al. (2014)), i.e.

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

where $i = 0, 1$ in $p_{data}(i)$ demonstrate the distributions of the true optical and SAR images respectively. $E_{x \sim p_{data}(i)}$ denotes that x obeys the distribution $p_{data}(i)$, and $E_{z \sim p_{data}(j)}$ denotes that z obeys the distribution $p_{data}(j)$. 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 (Goodfellow et al. (2014)), the loss function of the translator is

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

where $L_{GAN}(T)$ 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. (2017) found that the adversary loss function is better to be hybrid with traditional loss, such as L1 or L2 loss. Therefore, L1 norm loss is used to hybrid with the GAN loss, i.e. L1 distance between the translated

image $T(z)$ and the true image x :

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

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

$$L(T) = L_{GAN}(T) + \beta L_{L1}(T) \quad (4)$$

$L(T)$ 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 proposed loss function. Different losses contribute differently to the qualities of reconstructed results (Isola et al. (2017)). In Fig. 5, 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 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 (Arjovsky et al. (2017)). 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.

3.4. 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 step (see Fig. 6):

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

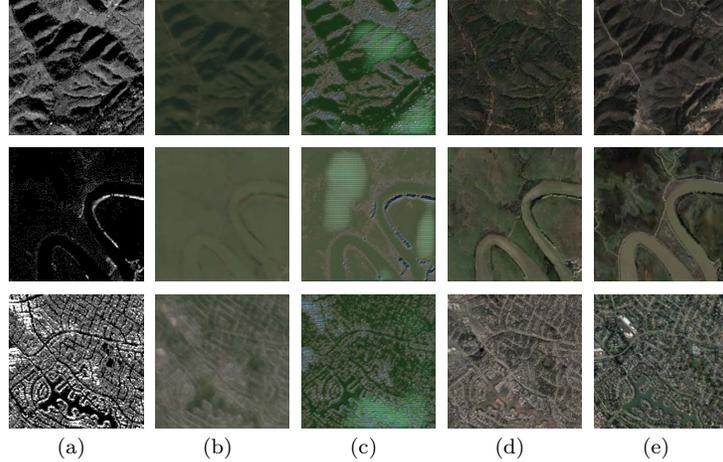


Fig. 5. Different qualities of translated optical images are induced by different losses. The first column lists the **(a) input SAR images**; the intermediate three columns are respectively **(b) translated optical images with L1**, **(c) translated optical images with GAN** and **(d) translated optical images with L1+GAN**; the last column are the corresponding **(e) optical ground truths**.

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 the image patches separately. This not only limits the receptive field, but also provides more samples for the training (Shrivastava et al. (2017)). 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 time 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 images is considered as an epoch. Then reshuffle the images and traverse next epoch.

We made several comparison experiments and set the parameter $\beta = 20$ in Eq. 4, with which the initial values of $L_{GAN}(T)$ and $L_{LI}(T)$ are approximately equal and the model is more stable and generates better results. We set the learning rate to 0.0002. Adam optimizer with $\beta_1 = 0.5$ is used. The input images are linearly mapped to the interval $[-1, 1]$. Leaky ReLU 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 (Isola et al. (2017)). 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 training loss does not decrease for four epochs in a row, the training is forced to stop.

In terms of training time, it takes only about 0.4 second per batch and 18 minutes to run through an epoch (10284 pairs of samples). Another 3 minutes are needed to test the test samples and save the test results and the checkpoint of the model. So it costs around 21 minutes to finish training and testing for an epoch.

3.5. 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 available, 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 (Zhu et al. (2017a)) proposes a cyclic loop which could be leveraged for this purpose. As shown in Fig. 7, the SAR image is first fed to the translator A and synthesizes fake optical

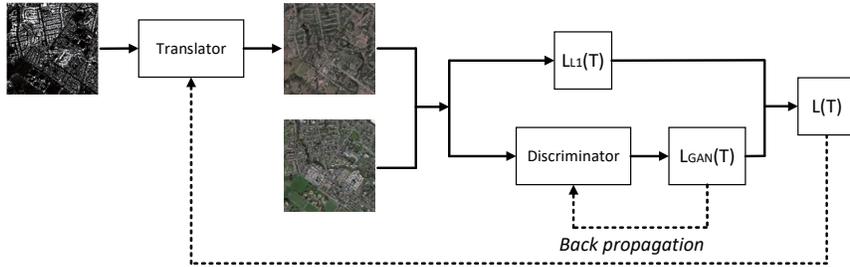


Fig. 6. The conceptual process of training the adversarial networks. The left image is the real SAR image, the upper right is the synthesized optical image and the lower right is the real optical image.

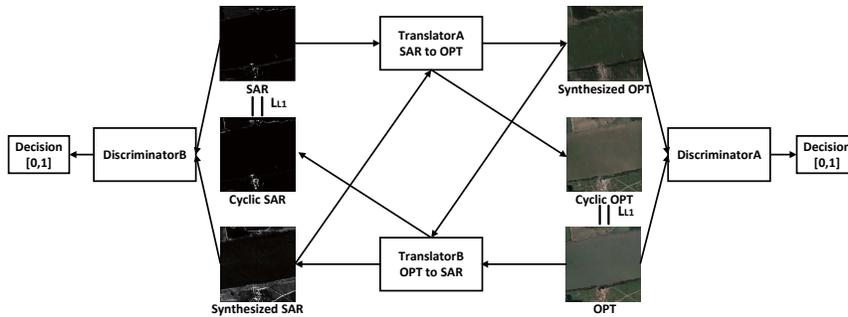


Fig. 7. Modified network scheme for unsupervised learning with CycleGAN loops (Zhu et al. (2017a)).

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 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 networks. The translator A and translator B networks are trained alternately during these two loops together with the discriminator networks. Later in subsection 4.5, we demonstrate how such unsupervised learning could further improve the performance of a translator initially trained with a small number of co-registered image pairs.

4. Experiments and Analyses

4.1. Datasets

SAR data used in this study mainly comes from the spaceborne GF3 SAR from China (China Centre for Resources Satellite Data and Application

(Jul. 2017)) and the airborne UAVSAR system from NASA (NASA 2018 (Dec.2018)). The information of those data used in our experiments is listed in the following Table 2. Optical data used is downloaded from Google Map around November, 2018, with pixel resolution 0.51m for GF3 SAR data and 1.02m for UAVSAR data respectively.

Table 2

Information about the two datasets employed for experiments.

		UAVSAR	GF3
SAR	Resolution	6m	0.51m
	Polarization	Quad-Pol	HH or VV
	Angle of Incidence	90°	40.6642° and 36.0820°
	Acquisition Mode	PolSAR	SL
	Frequency Band	80MHz	240MHz
	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	1.02m	0.51m
	Day of Acquisition	2018-11-25	2018-06-05, 2018-05-28
	Geographic Coordinate System	WGS 84	WGS 84

UAVSAR (Uninhabited Airborne Vehicle Synthetic Aperture Radar) radar system is an L-band polarimetric instrument developed by NASA. As shown in Fig. 8, UAVSAR data used here mainly consists 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 6.24.9m. The samples are 256256 patches 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.

GF3 satellite is Chinas first C-band multi-polarization SAR satellite. Two large scenes of GF3 images are used in the study with a resolution of 0.51m. The dataset contains different urban/suburban regions. In Fig. 9, the GF3 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.

A simple preprocessing step is to normalize the pixel value of the SAR images to $[-1, 1]$. Due to the considerably large range of pixel values, we

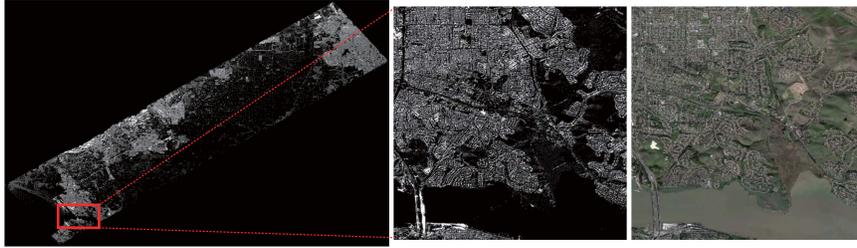


Fig. 8. (left) UAVSAR image acquired in California, US. (center) Zoomed region. (right) Corresponding optical image of the zoomed region.



Fig. 9. (left) A large SAR image with a 0.51m resolution in Hongshan District, Wuhan City, Hubei Province, China. (center) Zoomed region. (right) Corresponding optical image of the zoomed region.

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 the following Eq. 5.

$$\hat{x} = \begin{cases} -1, & \text{if } x \leq 0; \\ 1, & \text{if } x \geq \bar{x}; \\ 2x/\bar{x} - 1, & \text{otherwise.} \end{cases} \quad (5)$$

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

$$\bar{x} = \lambda(\sum_{i=1}^N x_i)/(N - n) \quad (6)$$

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 GF3 SAR images using a fast nonlocal despeckling filter (Cozzolino et al. (2014)). We found that speckle filtering can improve the quality of the final synthesized 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, due 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 UAVSAR and high resolution GF3 datasets respectively; the experiment for different polarization modes uses single-polarized and full-polarized UAVSAR data.

4.2.1. Resolution

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

Fig. 11 shows examples of high-resolution GF3 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

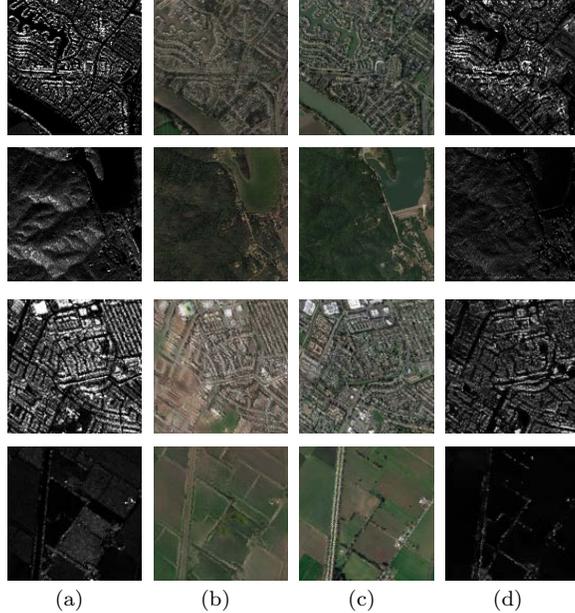


Fig. 10. Example translation images with UAVSAR (test samples). Images in each row from left to right are the (a) **real SAR image** and its (b) **translated optical image**, the (c) **real optical image** and its (d) **translated SAR image**. The first two rows are chosen from **6m UAVSAR**, and the last two are from **10m UAVSAR**.

can be easily reconstructed in both training and test cases. 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, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) (Wang et al. (2004)), 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) (Salimans et al.

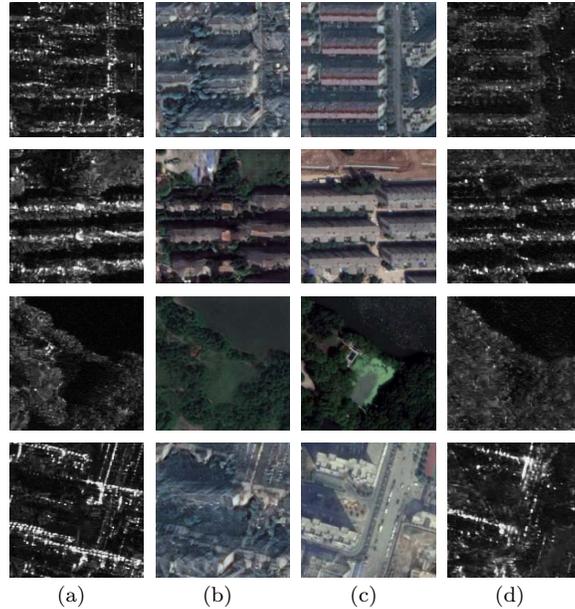


Fig. 11. Example translation images with GF3 data. Images in each row from left to right are the (a) **real SAR image** and its (b) **translated optical image**, the (c) **real optical image** and its (d) **translated SAR image**.

(2016)) and Frchet inception distance (FID) (Heusel et al. (2017)) 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 Fig. 12, 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

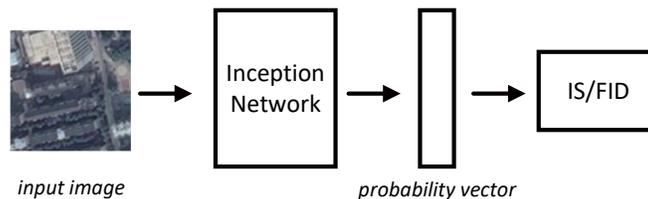


Fig. 12. The conceptual process of calculating IS/FID.

Table 3
FIDs of different datasets.

	0.51m single-pol GF3	6m single-pol UAVSAR	10m single-pol UAVSAR	6m full-pol UAVSAR
Optical	154.8	106.4	138.4	85.6
SAR	53.0	56.0	64.7	52.8

distribution with mean and covariance (m_1, C_1) and the Gaussian distribution with (m_2, C_2) is defined as $\|m_1 - m_2\|_2^2 + Tr(C_1 + C_2 - 2(C_1 C_2)^{1/2})$. Lower FID is better, corresponding to more similar real and generated samples. It is found that the FIDs of the generated 0.51m 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 are very 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. 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 (Heusel et al. (2017)). 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 3 lists the tested FID values for the different resolution datasets 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 train this kind of dataset. It indicates that 6m data performs better than 0.51m and 10m data. The 10m results are not ideal due to their small features hard to extract.

Note that FID could be further reduced by increasing the number of samples. As shown in Table 4 for the case of 6m full-pol UAVSAR, its FID could be reduced to 72 for optical and 42 for SAR if given 10000 samples.

Table 4

Decreasing FID with increasing the number of samples (for the case of 6m full-pol UAVSAR in Table 3).

Num	500	1000	2048	3000	4000	5000	6000	7000	8000	9000	10000
Opt.	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

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 to investigate how the performance might improve if full-pol SAR is used. For simplicity, the pauli color-coded image is used as an proxy of full-pol SAR data.

The basic form of polarimetric SAR data is the Sinclair scattering matrix (Lee and Pottier (2009)) 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 that in the co-polarized channels, and vice versa (Jin and Xu (2013)).

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.

$$a = \frac{S_{HH} + S_{VV}}{\sqrt{2}}, b = \frac{S_{HH} - S_{VV}}{\sqrt{2}}, c = \frac{S_{HV} + S_{VH}}{\sqrt{2}}, d = j \frac{S_{HV} - S_{VH}}{\sqrt{2}} \quad (7)$$

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

$$I = [|S_{HH} - S_{VV}|^2, 4|S_{HV}|^2, |S_{HH} + S_{VV}|^2]^T / 2 \quad (8)$$

Finally, we carry out an experiment to train our model with full-pol images and single-pol images in the same region respectively, and compare

the translation performance. As shown in Table 3, it indicates that 6m full-polarized data has the best performance, especially the reconstructed optical images are much better than those from 6m single-pol data. Four examples of different kinds of earth surfaces, waters, vegetation, farmlands and buildings are shown Fig. 13, respectively. Apparently, the optical images translated from full-pol SAR images are more vivid and realistic than those from single-pol images.

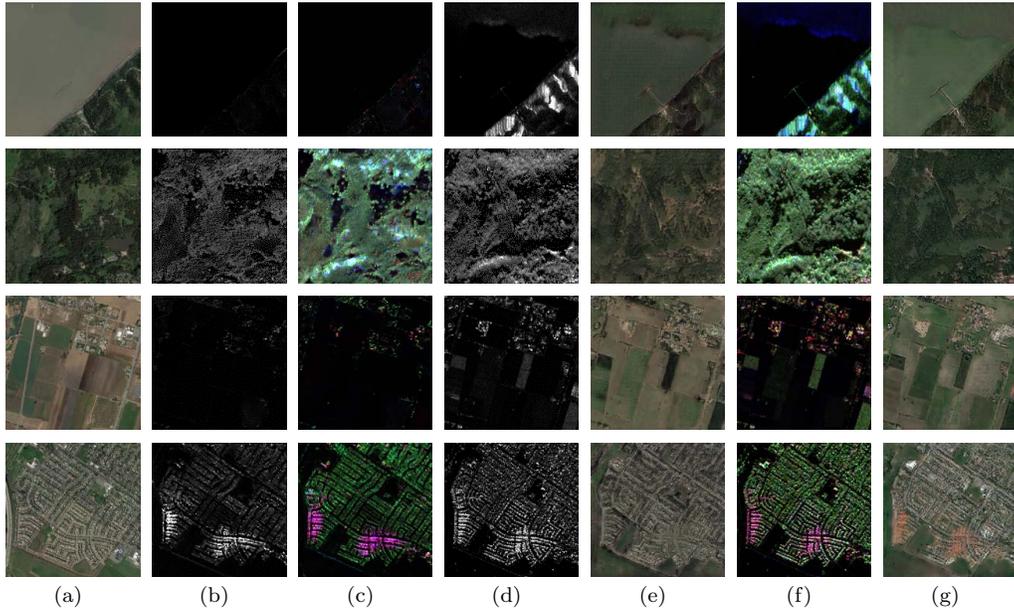


Fig. 13. Images listed above in each row are (a) the optical ground truth and its (b) translated single-polarized SAR image and (c) translated full-polarized SAR image, the (d) single-polarized SAR ground truth and its (e) translated optical image, the (f) full-polarized SAR ground truth and its (g) translated optical image in order. Each row lists a kind of earth surfaces: waters, vegetation, farmlands and buildings.

Fig. 14 further investigates into few interesting cases. In each case, one building in single-pol image and the corresponding full-pol image is marked correspondingly in each row. Note that these buildings are all easily observable in the full-pol image but not in the single-pol image. 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 appear 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.

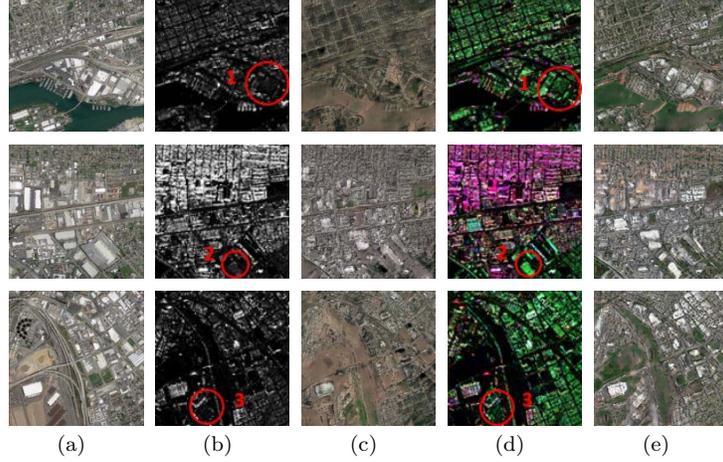


Fig. 14. Images in each row from left to right are the **(a) real optical image**, the real **(b) single-pol SAR image** and its **(c) translated optical image**, the real **(d) full-pol SAR image** and its **(e) translated optical image**.

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 CycleGAN (Zhu et al. (2017a)) and Pix2Pix (Isola et al. (2017)) using the 0.51m GF3 images. 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 initialization, each network is repeatedly trained for 3 times with the same data and then the best result is chosen.

In Fig. 15, 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, waters, roads and vegetation reconstructed by CycleGAN are similar to the proposed method.

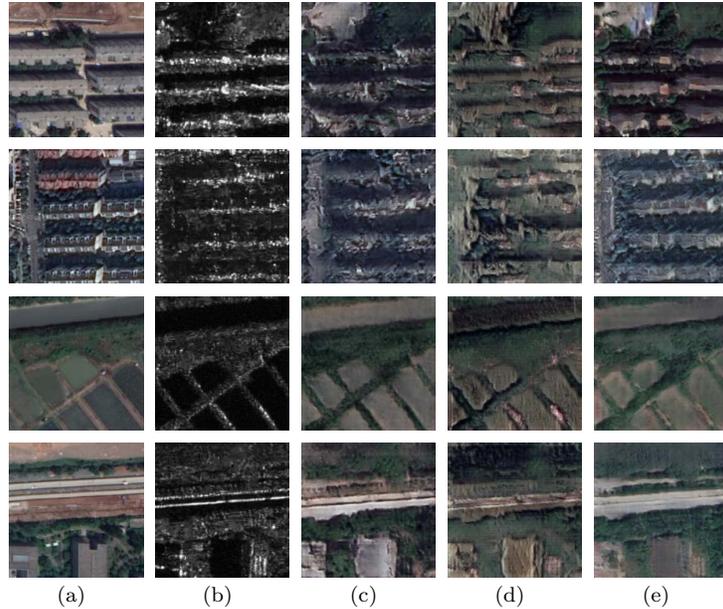


Fig. 15. Comparison of SAR-Optical translation by different methods. Images in each row from left to right are the (a) **real optical image**, the (b) **input SAR image**, its (c) **translated optical image by CycleGAN**, the (d) **translated optical image by Pix2Pix** and the (e) **translated optical image by CRAN**. Each row lists a kind of earth surfaces: buildings, buildings, farmlands and roads.

In Table 5, the three metrics PSNR, SSIM and FID are all employed. 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 generated samples to real world samples, and the smaller FID, the more analogous the two datasets are. On the 0.51m and 6m single-pol datasets, our proposed method outperforms CycleGAN and Pix2Pix in the other three indicators, especially the FID score improved greatly. Note that for the cases of 10m UAVSAR and 6m 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 Fig. 16 below. The metrics of these cases

Table 5

Result comparisons of different methods with different datasets using different evaluation methods.

Dataset	Method	SSIM		PSNR		FID	
0.51m single-pol GF3	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
6m 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
10m 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
6m 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

are also provided above the corresponding images. As we can see from the result that some cases appears better visually but measured with slightly lower metrics. For example, for the images a , b and c selected from 6m 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 10m single-pol UAVSAR data, similarly, e generated by CRAN appears visually better than f , but the latter has a larger PSNR value.

4.4. 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 image from another platform.

An experiment is conducted where the model trained using UAVSAR images is used to translate SAR images from UAVSAR, GF3 and ALOS2

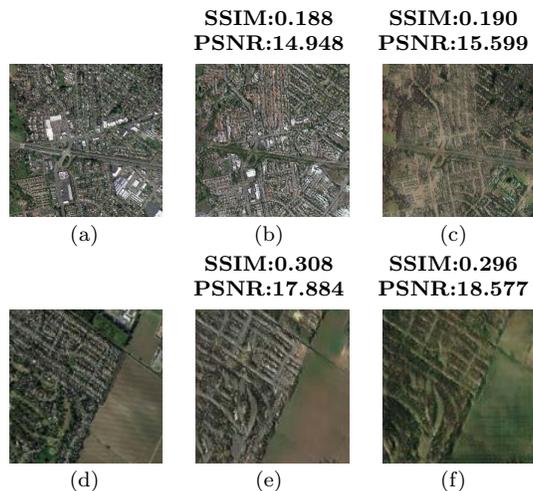


Fig. 16. Comparison of SAR-Optical translation by different methods to verify the not exactly correctness of evaluation metrics. The first row is chosen from 6m full-pol UAVSAR dataset: **(a)Ground Truth**, **(b)Trans. Opt. by CRAN**, **(c)Trans. Opt. by CycleGAN**; the second row is chosen from 10m single-pol UAVSAR dataset: **(d)Ground Truth**, **(e)Trans. Opt. by CRAN**, **(f)Trans. Opt. by Pix2Pix**.

acquired at different regions (Fig. 17). Compared to the ground truth, the performance of translation is largely degraded in the case of GF3 and ALOS2. 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 256256 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.

4.5. 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 trained with co-registered images. Then we can feed the SAR or optical images to be tested to the network and train them using large volume of unpaired optical or SAR images. Compared with supervised learning, in which only the prior knowledge from pretraining can be utilized, the model

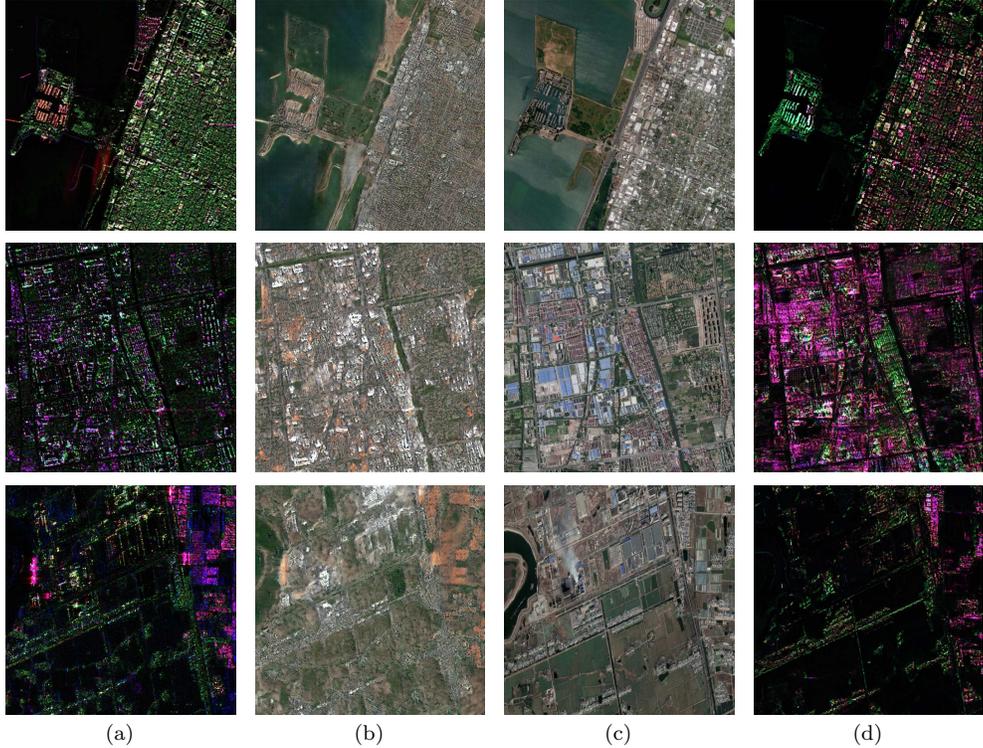


Fig. 17. Images across scenes and across sensors reconstructed by the model pre-trained with 6m UAVSAR images. Images in each row from left to right are the **(a) real SAR image** and its **(b) translated optical image**, the **(c) real optical image** and its **(a) translated SAR image**. Each row lists a kind of dataset: **UAVSAR**, **GF3** and **ALOS2**.

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

Table 6

FIDs of results by supervised and unsupervised learning.

	Supervised Learning	+Unsupervised Learning
Optical	107.8	88.9
SAR	58.1	41.2

vised network, train until the early stop and save the translated SAR images;

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

As shown in Table 6, it indicates that the translation results are greatly improved with unsupervised learning. The results further refined with unsupervised training are shown in Fig. 18, where we can see that the refined results are more vivid and realistic. However, waters in SAR images don't differ from those farmlands greatly, which results in the imperfect reconstruction of waters.

4.6. Computational cost

In this section, 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.

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. It should be noted that only the convolutional layers are considered, and the trainable parameters and operations generated by LeakyReLU and BN are ignored. Table 7 indicates that the numbers of parameters are almost same, but the number of operations in CycleGAN is approximately twice that of Pix2Pix and CRAN, due 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 Fig. 19 we can find:

- For the same method, the training speed using 4 GPUs is approximately $2 \sim 3$ times of that using 1 GPU. This is due to the communication

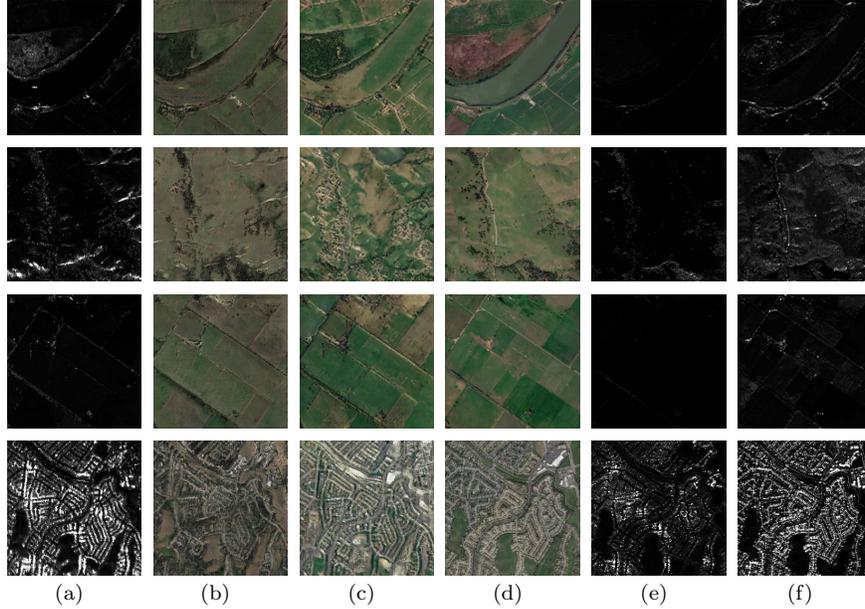


Fig. 18. Translated images further refined with unsupervised learning. Images in each row from left to right are the **(a) input SAR image**, the **(b) translated optical image** and the further **(c) refined optical image** by unsupervised learning, the **(d) input optical image** and its **(e) translated SAR image** and the further **(f) refined optical image** by unsupervised learning. Each row lists a kind of earth surfaces: waters, vegetation, farmlands and buildings.

overhead and some part of computation that cannot be run distribut-
edly.

- The speed of CRAN and Pix2Pix is much faster than CycleGAN, which agrees with the analyses given in Table 7.

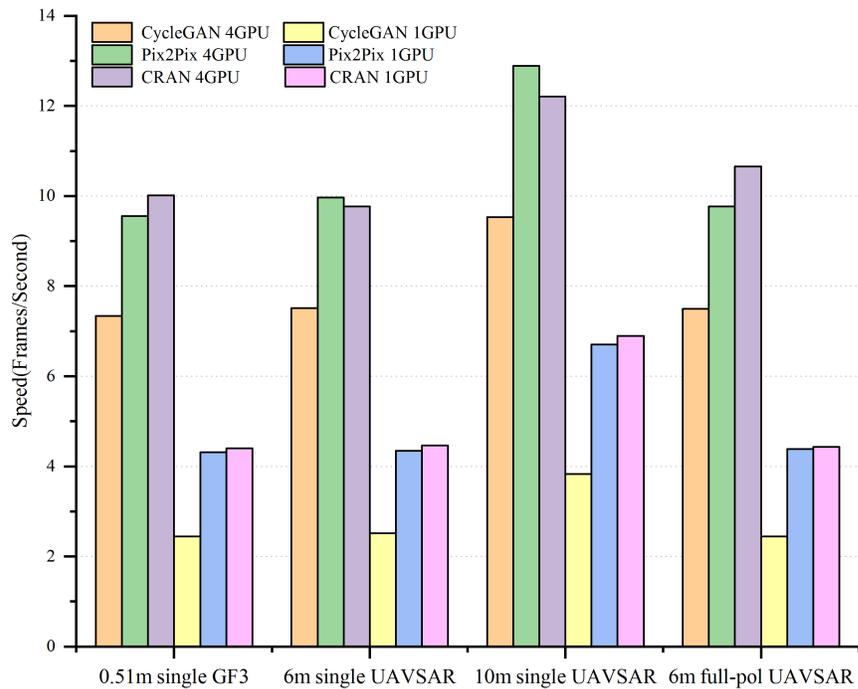
5. Conclusion

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 (6m, 10m) UAVSAR dataset, the reconstructed images appear very similar to the

Table 7

Number of trainable parameters and operations in the three translation networks.

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

**Fig. 19.** Speed comparison between different methods, different methods and different number of GPUs. On each dataset, the three translation networks respectively run on four GPUs and one single GPU.

true data and the corresponding FID is low. For high-resolution (0.51m) GF3 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. Finally, 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.

Acknowledgement

This work was supported in part by National Key *R&D* Program of China no. 2017YFB0502703 and Natural Science Foundation of China no. 61822107, 61571134.

References

- Ao, D., Dumitru, C., Schwarz, G., Datcu, M., 2018. Dialectical gan for sar image translation: From sentinel-1 to terrasars-x. *Remote Sensing* 10, 1597. doi:10.3390/rs10101597.
- Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein gan. arXiv preprint arXiv:1701.07875 .
- Byun, Y., Choi, J., Han, Y., 2013. An area-based image fusion scheme for the integration of sar and optical satellite imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6, 2212–2220. doi:10.1109/JSTARS.2013.2272773.
- Chen, Q., Koltun, V., 2017. Photographic image synthesis with cascaded refinement networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1520–1529. doi:10.1109/ICCV.2017.168.
- Chen, S., Wang, H., Xu, F., Jin, Y.Q., 2016. Target classification using the deep convolutional networks for sar images. *IEEE Transactions on*

- Geoscience and Remote Sensing 54, 4806–4817. doi:10.1109/TGRS.2016.2551720.
- China Centre for Resources Satellite Data and Application, Jul. 2017. Dataset: Gaofen3[Online]. URL: <http://www.cresda.com/CN/>.
- Cozzolino, D., Parrilli, S., Scarpa, G., Poggi, G., Verdoliva, L., 2014. Fast adaptive nonlocal sar despeckling. IEEE Geoscience and Remote Sensing Letters 11, 524–528. doi:10.1109/LGRS.2013.2271650.
- Enomoto, K., Sakurada, K., Wang, W., Kawaguchi, N., Matsuoka, M., Nakamura, R., 2018. Image translation between sar and optical imagery with generative adversarial nets, in: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, IEEE. pp. 1752–1755. doi:10.1109/IGARSS.2018.8518719.
- Fan, J., Wu, Y., Li, M., Liang, W., Cao, Y., 2018. Sar and optical image registration using nonlinear diffusion and phase congruency structural descriptor. IEEE Transactions on Geoscience and Remote Sensing 56, 5368–5379. doi:10.1109/TGRS.2018.2815523.
- Fuster, R.M., Usón, M.F., Ibars, A.B., 2017. Interferometric orbit determination for geostationary satellites. Science China Information Sciences 60, 060302. doi:10.1007/s11432-016-9052-y.
- Gao, F., Ma, F., Wang, J., Sun, J., Yang, E., Zhou, H., 2018. Semi-supervised generative adversarial nets with multiple generators for sar image recognition. Sensors 18, 2706. doi:10.3390/s18082706.
- Garzelli, A., 2002. Wavelet-based fusion of optical and sar image data over urban area. International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences 34, 59–62.
- Gatys, L.A., Ecker, A.S., Bethge, M., 2016. Image style transfer using convolutional neural networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2414–2423. doi:10.1109/CVPR.2016.265.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger,

- K.Q. (Eds.), *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., pp. 2672–2680.
- Guarnieri, A.M., Rocca, F., 2017. Options for continuous radar earth observations. *Science China Information Sciences* 60, 060301. doi:10.1007/s11432-016-9067-7.
- Guo, J., Lei, B., Ding, C., Zhang, Y., 2017. Synthetic aperture radar image synthesis by using generative adversarial nets. *IEEE Geoscience and Remote Sensing Letters* 14, 1111–1115. doi:10.1109/LGRS.2017.2699196.
- He, W., Yokoya, N., 2018. Multi-temporal sentinel-1 and -2 data fusion for optical image simulation. *ISPRS International Journal of Geo-Information* 10. doi:10.3390/ijgi7100389.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: *Advances in Neural Information Processing Systems*, pp. 6626–6637.
- Hughes, L., Schmitt, M., Zhu, X., 2018. Mining hard negative samples for sar-optical image matching using generative adversarial networks. *Remote Sensing* 10, 1552. doi:10.3390/rs10101552.
- Isola, P., Zhu, J., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976. doi:10.1109/CVPR.2017.632.
- Jin, K.H., McCann, M.T., Froustey, E., Unser, M., 2017. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing* 26, 4509–4522. doi:10.1109/TIP.2017.2713099.
- Jin, Y.Q., Xu, F., 2013. *Polarimetric scattering and SAR information retrieval*. John Wiley & Sons. doi:10.1002/9781118188149.
- Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution, in: *European conference on computer vision*, Springer. pp. 694–711. doi:10.1007/978-3-319-46475-6_43.

- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. URL: <http://doi.acm.org/10.1145/3065386>, doi:10.1145/3065386.
- Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., et al., 2017. Fader networks: Manipulating images by sliding attributes, in: *Advances in Neural Information Processing Systems*, pp. 5967–5976.
- Lee, J.S., Pottier, E., 2009. *Polarimetric radar imaging: from basics to applications*. CRC press. doi:10.1201/9781420054989.
- Liu, J., Gong, M., Qin, K., Zhang, P., 2018. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. *IEEE transactions on neural networks and learning systems* 29, 545–559. doi:10.1109/TNNLS.2016.2636227.
- Liu, M.Y., Breuel, T., Kautz, J., 2017. Unsupervised image-to-image translation networks, in: *Advances in Neural Information Processing Systems*, pp. 700–708.
- Ma, W., Pan, Z., Guo, J., Lei, B., 2018. Super-resolution of remote sensing images based on transferred generative adversarial network, in: *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, IEEE. doi:10.1109/IGARSS.2018.8517442.
- Merkle, N., Auer, S., Müller, R., Reinartz, P., 2018. Exploring the potential of conditional adversarial networks for optical and sar image matching. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11, 1811–1820. doi:10.1109/JSTARS.2018.2803212.
- Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 .
- NASA 2018, A.D., Dec.2018. Dataset: UAVSAR[Online]. URL: <https://vertex.daac.asf.alaska.edu/>.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: *International Conference on Medical image computing and computer-assisted intervention*, Springer. pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans, in: Advances in neural information processing systems, pp. 2234–2242.
- Schmitt, M., Hughes, L.H., Zhu, X.X., 2018. The sen1-2 dataset for deep learning in sar-optical data fusion. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences IV-1, 141–146. URL: <https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/IV-1/141/2018/>, doi:10.5194/isprs-annals-IV-1-141-2018.
- Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R., 2017. Learning from simulated and unsupervised images through adversarial training, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2242–2251. doi:10.1109/CVPR.2017.241.
- Song, Q., Xu, F., 2017. Zero-shot learning of sar target feature space with deep generative neural networks. IEEE Geoscience and Remote Sensing Letters 14, 2245–2249. doi:10.1109/LGRS.2017.2758900.
- Taigman, Y., Polyak, A., Wolf, L., 2016. Unsupervised cross-domain image generation. arXiv preprint arXiv:1611.02200 .
- Wang, P., Patel, V.M., 2018. Generating high quality visible images from sar images using cnns, in: 2018 IEEE Radar Conference (RadarConf18), IEEE. pp. 0570–0575. doi:10.1109/RADAR.2018.8378622.
- Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al., 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13, 600–612. doi:10.1109/TIP.2003.819861.
- Yue, D.X., Xu, F., Jin, Y.Q., 2018. Sar despeckling neural network with logarithmic convolutional product model. International journal of remote sensing 39, 7483–7505. doi:10.1080/01431161.2018.1471539.
- Zhang, H., Deng, Y., Wang, R., Li, N., Zhao, S., Hong, F., Wu, L., Loffeld, O., 2016a. Spaceborne/stationary bistatic sar imaging with terrasars-x as an illuminator in staring-spotlight mode. IEEE Transactions on Geoscience and Remote Sensing 54, 5203–5216. doi:10.1109/TGRS.2016.2558294.

- Zhang, M., Wang, R., Deng, Y., Wu, L., Zhang, Z., Zhang, H., Li, N., Liu, Y., Luo, X., 2016b. A synchronization algorithm for spaceborne/stationary bistatic imaging based on contrast optimization with direct signal from radar satellite. *IEEE Transactions on Geoscience and Remote Sensing* 54, 1977–1989. doi:10.1109/TGRS.2015.2493078.
- Zhang, Z., Wang, H., Xu, F., Jin, Y.Q., 2017. Complex-valued convolutional neural network and its application in polarimetric sar image classification. *IEEE Transactions on Geoscience and Remote Sensing* 55, 7177–7188. doi:10.1109/TGRS.2017.2743222.
- Zhou, Y., Wang, H., Xu, F., Jin, Y.Q., 2016. Polarimetric sar image classification using deep convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters* 13, 1935–1939. doi:10.1109/LGRS.2016.2618840.
- Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251. doi:10.1109/ICCV.2017.244.
- Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E., 2017b. Toward multimodal image-to-image translation, in: *Advances in Neural Information Processing Systems*, pp. 465–476.
- Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017c. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine* 5, 8–36. doi:10.1109/MGRS.2017.2762307.