Exploring the Potential of Conditional Adversarial Networks for Optical and SAR Image Matching

Nina Merkle, Stefan Auer, Rupert Müller and Peter Reinartz, *Member, IEEE*

***Abstract*—Tasks such as the monitoring of natural disasters or the detection of change highly benefit from complementary information about an area or a specific object of interest. The required information is provided by fusing high accurate co-registered and geo-referenced datasets. Aligned high resolution optical and synthetic aperture radar (SAR) data additionally enables an absolute geo-location accuracy improvement of the optical images by extracting accurate and reliable ground control points (GCPs) from the SAR images. In this paper we investigate the applicability of a deep learning based matching concept for the generation of precise and accurate GCPs from SAR satellite images by matching optical and SAR images. To this end, conditional generative adversarial networks (cGANs) are trained to generate SAR-like image patches from optical images. For training and testing, optical and SAR image patches are extracted from TerraSAR-X and PRISM image pairs covering greater urban areas spread over Europe. The artificially generated patches are then used to improve the conditions for three known matching approaches based on normalized cross-correlation (NCC), SIFT and BRISK, which are normally not usable for the matching of optical and SAR images. The results validate that a NCC, SIFT and BRISK based matching greatly benefit, in terms of matching accuracy and precision, from the use of the artificial templates. The comparison with two state-of-the-art optical and SAR matching approaches shows the potential of the proposed method but also revealed some challenges and the necessity for further developments.**

***Index Terms*—Conditional generative adversarial networks (cGANs), multi-sensor image matching, artificial image generation, synthetic aperture radar (SAR), optical satellite images.**

I. Introduction

MULTI-SENSOR image fusion is a prerequisite for the provision of complementary information through the combination of different data. Aligned multi-sensor data enable a more robust interpretation of image scenes or specific objects and is therefore crucial for tasks such as monitoring natural disasters and change detection. In the case of optical and synthetic aperture radar (SAR) satellites, the images acquired by both sensors exhibit quite different characteristics: SAR satellites have an active sensor on board which emits electromagnetic signals and measures the strength and time delay of the returned signal backscattered from ground objects. The visual interpretation of SAR images is a challenging task, due to the specific imaging principle and the presence of speckle in the images. In contrast, optical sensors measure the sun radiation reflected from objects on ground. The interpretation of optical images is easier which makes the development of feature detectors and therefore the detection of features more efficient and robust. An advantage of a SAR sensor (especially TerraSAR-X and TanDEM-X) is that the images exhibit absolute geo-location accuracies within few decimeters, whereas high resolution optical sensors still require ground control points (GCPs) to reach similar accuracies. This can be traced back to the different image acquisition concepts. SAR sensors determine the distance to ground object via the signal traveling time, which can be measured precisely if also atmospheric effects are taken into account and lead to images with high geo-location accuracy. Due to recent developments in SAR geodesy, high resolutions SAR satellites such as TerraSAR-X exhibit an absolute geo-localization accuracy in the range of a few decimeters [1]. Optical sensors in contrast, require the measurement of the attitude angles in space to determine the satellite-viewing direction to ground objects, which often suffers from insufficient accuracy of the measurements and results in images with lower absolute geo-location accuracy.

The main objective of our research is therefore the improvement of the absolute geo-localization accuracy of optical satellite images via automatic extracted GCPs from images acquired by the high resolution radar satellite TerraSAR-X. If GCPs are available, the geo-localization accuracy of the optical images can be enhanced by using these points to correct the underlying sensor model parameters for the geo-referencing process. However, GCPs are commonly measured by tedious in-situ GPS measurements or from very exact maps and are therefore available only in the minority of cases. To overcome this shortage this paper focuses on an automatic procedure to generate GCPs through the matching of optical and SAR images. Note that we will leave out the subsequent step of geo-localization accuracy improvement of the optical images as e.g. performed in [2] and [3].

The process of image matching and registration is of interests for a variety of applications in fields such as medicine, computer vision and remote sensing, and hundreds of different approaches have been developed [4], [5], [6]. Common methods for the matching of optical and SAR images are mostly based on intensity- or feature-based matching concepts. Intensity-based methods often exploit similarity measures such as normalized cross-correlation (NCC) [7], mutual information

Manuscript received September 14, 2017; revised December 4, 2017 and January 23, 2018; accepted January 28, 2018. *(Corresponding author: Nina Merkle)*

The authors are with the Remote Sensing Technology Institute of the German Aerospace Center (DLR), 82234 Wessling, Germany (e-mail: nina.merkle@dlr.de; stefan.auer@dlr.de; rupert.mueller@dlr.de; peter.reinartz@dlr.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2018.2803212

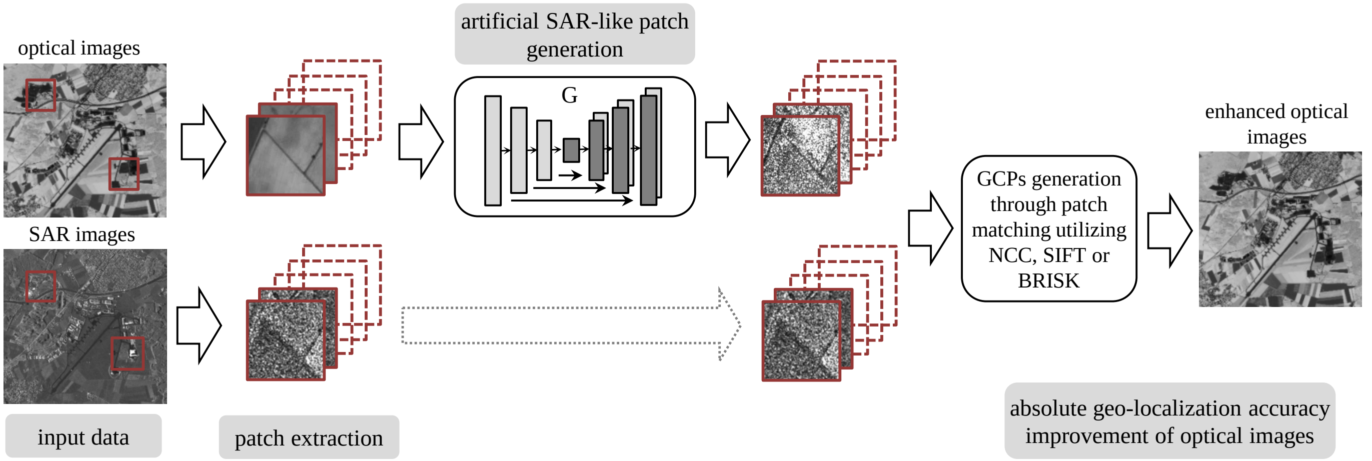
optical images
SAR images
input data
patch extraction
artificial SAR-like patch generation
G
GCPs generation through patch matching utilizing NCC, SIFT or BRISK
enhanced optical images
absolute geo-localization accuracy improvement of optical images

Fig. 1. Graphical overview of the proposed method for the absolute geo-localization improvement of optical images by matching SAR and artificial patches generated by conditional generative adversarial networks (cGANs).

(MI) [2], [8] or cross-cumulative residual entropy [9]. On the other hand features like lines [10], contours [11], [12] or regions [13] are widely used for feature-based matching approaches. A modification of the common feature detector SIFT, which usually fails to detect corresponding features in SAR and optical images, has been investigated in [14]. All these approaches suffer from speckle in the SAR images and different geometric and radiometric properties induced by the disparate image acquisition concepts of the two sensors. This leads to the problem of either finding a reliable similarity measure between the images, or extracting reliable features from the image scenes. To circumvent this problem Han et al. [15] proposed an approach, which combines aspects of feature- and intensity-based methods. In our previous work we investigated the applicability of a deep learning based method for the task of optical and SAR images matching [3]. To handle the matching problems arising from optical and SAR data we propose to select specific areas, where only the radiometry is different in both images. Using these areas we successfully trained a Siamese neural network to learn the matching between SAR and optical image patches and achieved better results than state-of-the-art approaches.

Inspired by the high potential and the possibilities provided by new developments in the field of deep learning we continue our investigation and propose a new deep learning based technique for automatic GCPs generation through matching of optical and SAR image patches. Towards this goal we trained a conditional generative adversarial network (cGAN) to generate artificial SAR-like image patches from optical satellite images. In contrast to our previous work in [3], where the matching between optical and SAR patches was directly learned by a Siamese network, the idea here is to use the artificially generated patches to improve the accuracy and precision of common matching approaches, which are usually inapplicable for the matching between optical and SAR images. The evaluation focuses on one intensity-based, NCC [16], and on two feature-based matching approaches, SIFT [17] and BRISK [18]. Optical and SAR image pairs acquired over Europe (from 46 TerraSAR-X and PRISM scenes) and manually aligned, are used for training and evaluating the network. The results are compared with two state-of-the art optical and SAR matching approaches and demonstrate the effectiveness of the proposed method. A visualization of the method is depicted in Figure 1.

This paper represents an extension of our earlier work presented in [19]. Compared to [19] we extended the method by two additional cGAN loss functions and extensively investigated and discussed the influence of the different losses, the different batch sizes, the different training datasets and the influence of a speckle filter on the matching results. Furthermore, we compared the obtained results with two available state-of-the-art optical and SAR matching approaches. The main contributions of our paper are: (i) Providing a new concept to handle the problem of multi-sensor image matching based on cGAN, which (ii) improves the results of common techniques (NCC, SIFT and BRISK) for the matching between optical and SAR images while (iii) achieving comparable results in regard to two state-of-the-art methods.

II. Generative Adversarial Networks

Neural networks, especially convolutional neural networks, proved their high potential in various fields like computer vision, biology, medical imaging and remote sensing. Recently, Goodfellow et al. [20] introduced a new machine learning architecture, GANs, which earned a lot of attention in the field of machine learning and offers new possibilities for several research problems by generating high quality images. In computer vision GANs find application for problems such as semantic segmentation [21] or single image super-resolution [22]. In the field of medicine, GANs are successfully applied for the generation of computed tomography (CT) images from magnetic resonance imaging (MRI) to reduce the radiation exposure to patients during acquisition [23]. In the context of remote sensing, Guo et al. [24] investigated the application of GANs for the synthesis of SAR images.

GANs are generative models with the goal of training a generator network G to map random noise z to output images y. The training is realized through an adversarial process, which is based on the simultaneous training of two networks, the generator G and the discriminator D. The task of D is to distinguish as good as possible between real images and images $G(z)$ generated by G, whereas G tries to produce more and more realistic images to "fool" D as often as possible. The problem can be expressed through a two-player minimax game

$$\min_{G}\max_{D}\mathcal{L}_{\text{GAN}}(G,D) = E_{y\sim p_{\text{real}}(y)}[\log D(y)]+ E_{y\sim p_{\text{real}}(y),z\sim p_z(z)}[\log(1-D(G(z)))], \quad (1)$$

where E denotes the expected value, p_{real} the real data distribution and p_z the noise distribution. D is realized by a binary classification network and outputs the possibility that an input image belongs either to the class 0 ("fake") or to the class 1 ("real"). The aim of D during training is to get $D(G(z))$ close to 0, which means to detect all images generated by G and label them correctly as "fake". In contrast, G aims to get $D(G(z))$ close to 1, which means that D does not identify the artificial images generated by G and wrongly label them as "real". To ensure that the output values of D lie in the range of $[0,1]$ a sigmoid layer can be used as the last layer of D.

In this paper, we investigate the applicability of conditional GANs (cGANs) and we therefore utilize the open source implementations from Isola et al. [25]. In the following section, we will describe the concept of cGANs and how to use them for the matching of optical and SAR images.

III. Multi-Sensor Image Matching

The proposed method deals with the problem of matching SAR and optical image patches in three steps. In the first stage suitable matching areas are selected from optical and SAR images. The second stage is the generation of artificial SAR-like patches from optical image patches through cGANs. The third stage is the matching of artificially generated SAR patches with the real SAR image patches using an intensity-based (NCC) and two feature-based matching approaches (SIFT and BRISK).

A. Matching Area Selection

The pre-selection of suitable matching areas increases the probability to obtain accurate and reliable matching points between SAR and optical images. Candidates for such areas contain almost only planar objects, which exhibit the same (at least to a certain degree) geometric appearance in the optical and in the corresponding SAR image. Furthermore, these areas should contain salient features to increase the probability of a successful matching. In most cases these features are related to man-made infrastructure objects such as streets, street crossings, roundabouts and borders between agricultural fields. The reason for excluding 3D objects are the different geometric distortions induced by the different sensors of optical and SAR satellites. Elevated objects like buildings appear differently in SAR and optical images and get projected to different positions within the image. These features are therefore not suitable for the identification of GCPs. The collection of suitable patches is realized via a semi-manual selection procedure. For obtaining a first indication of areas containing fitting patterns, the CORINE land cover layer [26] was applied. By applying this layer to the images all cities, industrial and forest areas can be excluded from the image search space. This first selection was refined manually to ensure that features within these areas are actually visible in both the optical and the SAR images.

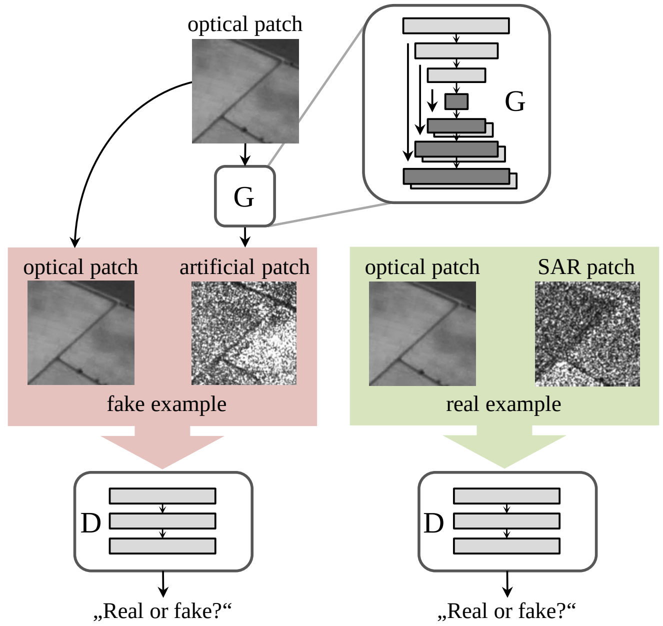

Fig. 2. Overview of cGAN training procedure. On the left side the training setup for "fake" examples (optical and artificial generated patch) as input and on the right side for "real" examples (optical and SAR patch pair) as input.

B. Artificial Image Generation

In contrast to the common GAN setup, where new images are generated only from noise, we want to generate artificial images based on a specific input image (an optical image patch). The aim is to generate an artificial SAR-like image with geometric properties of an optical image and with radiometric properties of a SAR image (the impact of geometric distortion is reduced by the pre-selection of patches as described in the previous Section III-A). Therefore, we utilize cGANs, which rely, next to noise z, on observed images x. The cGAN loss can be stated as

$$\mathcal{L}_{\text{cGAN}}(G,D) = E_{x,y\sim p_{\text{real}}(x,y)}[\log D(x,y)]+ E_{x,y\sim p_{\text{real}}(x,y),z\sim p_z(z)}[\log(1-D(x,G(x,z)))], \quad (2)$$

where x denotes an optical patch, y the corresponding SAR patch (the ground truth image patch) and $G(x,z)$ the artificially generated SAR-like patch. As in [25] we extend equations (2) by an additional term

$$\mathcal{L}_{L_1}(G) = E_{x,y\sim p_{\text{data}}(x,y),z\sim p_z(z)}[\|y-G(x,z)\|_1]. \quad (3)$$

This term forces G to produce output images, which are close to the ground truth SAR patches y (in sense of the L_1 distance). Adding this term lead to the final objective

$$G^{*} = \arg\min_{G}\max_{D}\mathcal{L}_{\text{cGAN}}(G,D)+\lambda\mathcal{L}_{L_1}(G). \quad (4)$$

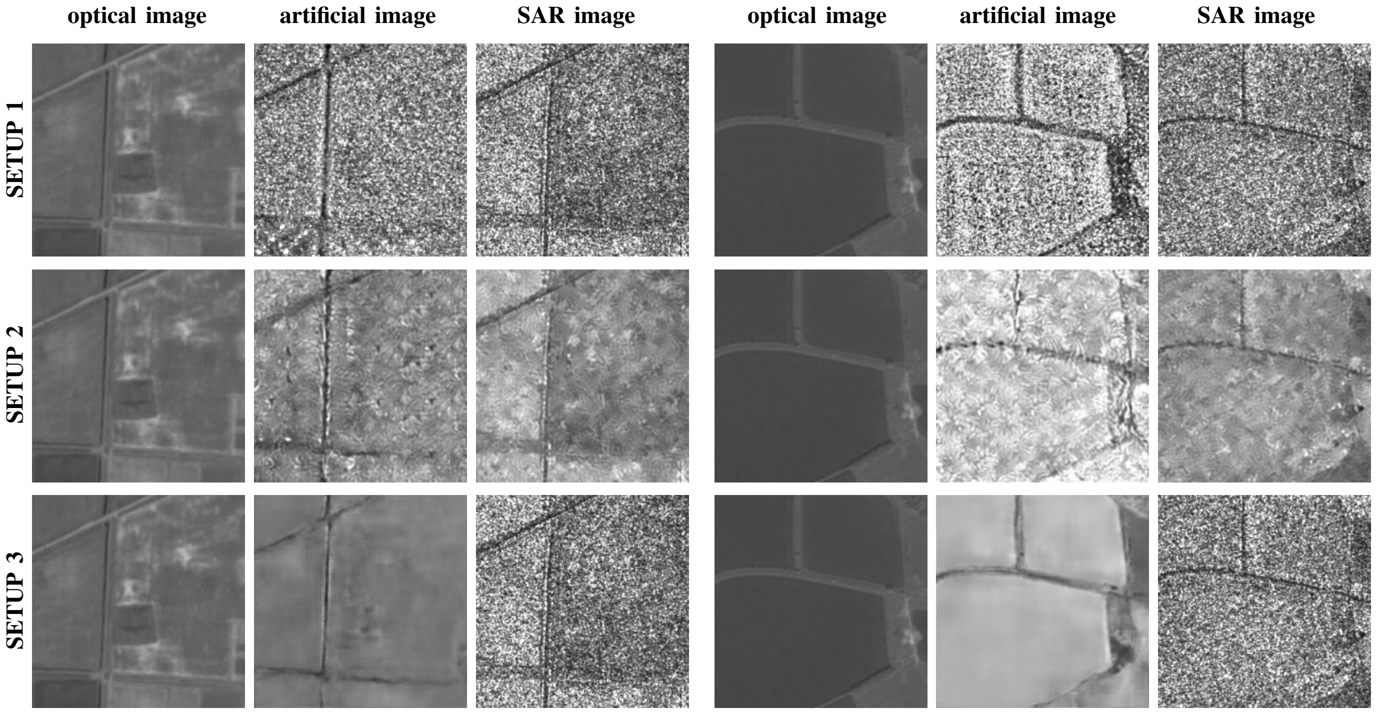

Fig. 3. Side by side comparison between optical, artificial SAR and original (despeckled) SAR image patches with pixel size of 2.5m for three different setups in two columns. Setup 1: SAR-like patch generation utilizing the cWGAN loss with a batch size of 1; Setup 2: despeckled SAR-like patch generation utilizing the cGAN loss with a batch size of 40; Setup 3: SAR-like patch generation utilizing the cLSGAN loss with a batch size of 4.

A common problem of (conditional) GANs with an objective based on the negative log-likelihood (see Equation (2)), is the unstable training. Recent research studies like [27], [28] try to overcome this problem by describing more stable training procedures. We therefore investigate the influence of two alternative training procedures on our matching results. The first was proposed in [27] and only requires a change in the loss function $\mathcal{L}_{\text{cGAN}}$. The idea is to replace the cGAN loss from Equation 2 by a least square loss

$$\begin{aligned}\mathcal{L}_{\text{cLSGAN}}(G,D) = E_{x,y\sim p_{\text{real}}(x,y)}[(D(x,y)-1)^2]+ \\ E_{x,y\sim p_{\text{real}}(x,y),z\sim p_z(z)}[D(x,G(x,z)))^2].\end{aligned} \tag{5}$$

We denote the new cGAN setup, where the least square loss is utilized, with cLSGAN. The second approach was proposed in [28]. Here, the idea is to restate the problem with the aim of minimizing the Wasserstein distance instead of the Jensen-Shannon divergence, which is the case for the common GAN problem. This can be achieved by employing the conditional Wasserstein GAN (cWGAN) loss

$$\begin{aligned}\mathcal{L}_{\text{cWGAN}}(G,D) = E_{x,y\sim p_{\text{real}}(x,y)}[D(x,y)]+ \\ E_{x,y\sim p_{\text{real}}(x,y),z\sim p_z(z)}[D(x,G(x,z))].\end{aligned} \tag{6}$$

Applying cWGANs also requires to clip the weights of the discriminator network D to be in the interval from -0.01 to 0.01. In the following this type of cGAN will be called the cWGAN setup and a detailed theoretical overview of it can be found in [28].

Network Architecture: The generator G is realized via a U-net, which is an encoder-decoder type of network with skip connections between layer i and layer $n-i$ (n is the total number of layers). A skip connection between the layers i and $n-i$ means to concatenate all channels of layer i with those of layer $n-i$. An example of this network type is shown in Figure 2. The discriminator is realized via several convolutional layers and the sigmoid function as the last layer. For a detailed overview of the network architecture we refer to [25].

Network Training: The training dataset consists of optical and SAR image patches, where we determined to train on patches with a size of 201×201 pixels (large enough to ensure the existence of salient features within the patches but not too large to run into problems caused by memory limits of our available GPUs). Before extracting the patches all images must be geometrically aligned. The discriminator network D is alternately trained on two different kinds of training pairs. Half of the training pairs are ”fake” examples and are composed of optical and artificial generated SAR-like patch pairs. The other half are ”real” examples and are composed optical and SAR patch pairs. An illustration of the two different training setups are shown in Figure 2. The networks are trained with stochastic gradient descent with the ADAM optimizer [29] and an initial learning rate of 0.01 for the cGAN and cLSGAN setups and with the RMSProp optimizer [30] and an initial learning rate of 0.0002 for the cWGAN setup. For all setups the two networks are trained at the same time by alternating the training of D and G (one gradient descent step of D is followed by one gradient descent step of G in the cGAN and cLSGAN setups and five gradient descent steps of D are followed by one gradient descent step of G in the cWGAN setup). To improve the

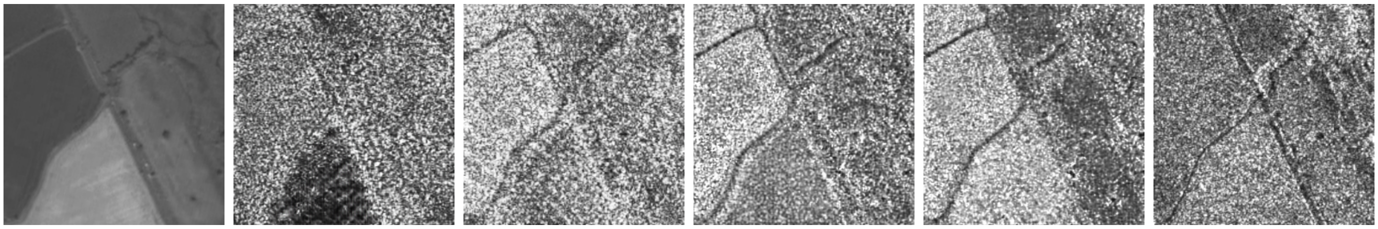

Fig. 4. Development of the generator over training: the optical input patch, the artificial patches at epoch 1,10, 50, 200 and the SAR target patch (from left to right).

quality of the learning while utilizing the $\mathcal{L}_{\mathrm{cGAN}}$ loss we follow the common practice for the training of G, which is to maximize $\log(D(x, G(x, z)))$ instead of minimizing $\log(1 - D(x, G(x, z)))$.

Network Testing: The networks are tested by comparing the quality improvement of the results between the matching of artificial SAR-like (generated by the trained cGAN, cLSGAN and cWGAN setups) with SAR patches and the matching of optical with SAR patches. Different techniques are utilized for the patch matching and are introduced in the following Subsection III-C. Note that in this phase of the process only the trained generator network G is required (as illustrated in Figure 1) and the weights of G are not modified during the test phase (one input patches will always lead to the same artificial output patch). To guarantee a fair evaluation we only utilize artificial patches for the matching, which are generated from a test set. The test set contains optical patches, which were never shown to the networks during training.

C. Artificial Image Matching

Several approaches exist to realize the matching between a template T and a corresponding reference image R. In our investigations we focus on one intensity-based and two feature-based approaches, which usually lead to inaccurate results for the matching of optical and SAR images. For the later evaluation, the template T will either be a patch cropped from the optical image or the generated artificial SAR-like patch and R a patch cropped from the SAR image.

Intensity-based approaches measure the similarity between T and a larger reference image R at all locations within the search space. We use a sliding window technique to compute the NCC [16] value for every location of T within R. The correct matching position is given by the highest NCC value within the search space. Since we are only interested in reliable and accurate matching points, we use the NCC value as a quality measure to detected outliers in the set of matching points. More precisely, we remove all matching points with a NCC value of less than 0.4.

In contrast, feature-based approaches are based on the detection of features in both images, called key points, and the measurement of their similarity in the feature space. The two feature detectors utilized in this paper are the scale-invariant feature transform (SIFT) [17] and the binary robust invariant scalable key points (BRISK) [18]. The idea of both algorithms is to find key points in T and R and to return a descriptor for every key point. The descriptors of two images are then matched by utilizing the Euclidean distance for SIFT and the Hamming distance for BRISK in combination with a nearest neighbor search. To increase the quality and reliability of the detected matching points we remove outliers through RANSAC [31] with an underlying affine model and with a distance threshold of 5 pixels.

IV. Experiments and Results

For the network training and for the evaluation of the results, training and test datasets are generated out of 46 orthorectified and aligned optical (PRISM) and SAR (TerraSAR-X acquired in spotlight mode) satellite image pairs. The manual alignment was realized within the Urban Atlas project [32] with an overall alignment error in the range of 3m. The images cover greater urban zones including suburban, industrial and rural areas of 13 cities in Europe. The pixel spacing of the PRISM images is 2.5m and of the TerraSAR-X images 1.25m. To obtain larger training datasets the TerraSAR-X images are resampled to 2.5m and 3.75m and the PRISM images to 3.75m through bilinear interpolation. To investigate the influence of speckle on the matching results we despeckled all SAR images applying the probabilistic patch-based filter introduced in [33].

A. Matching Area Selection

For the selection of suitable regions within the images (as described in Subsection III-A) we utilized the CORINE land cover layer [26] from the year 2012 and with a pixel spacing of 100m. The following classes are chosen as suitable regions: airports, non-irrigated arable land, permanently-irrigated land, annual crops associated with permanent crops and complex cultivation patterns, land principally occupied by agriculture, with significant areas of natural vegetation. After a manual refinement we generated two different training datasets and one test dataset. The first training dataset contains $69,900$ optical and SAR patch pairs with a resolution of 2.5m. The second training dataset contains all patch pairs from the first training dataset, but with a resolution of 2.5m and 3.75m. The patches with 3.75m resolution are centered around the same location as the 2.5m resolution patches but contain bigger areas and only exists in the dataset if the patches do not exceed the image boundaries. This led to a total number of $137,450$ patch pairs. The second training dataset is deployed to enlarge the number of training samples and to investigate the influence of different image resolutions on the quality of the patch generation and, hence, of the later matching. Since

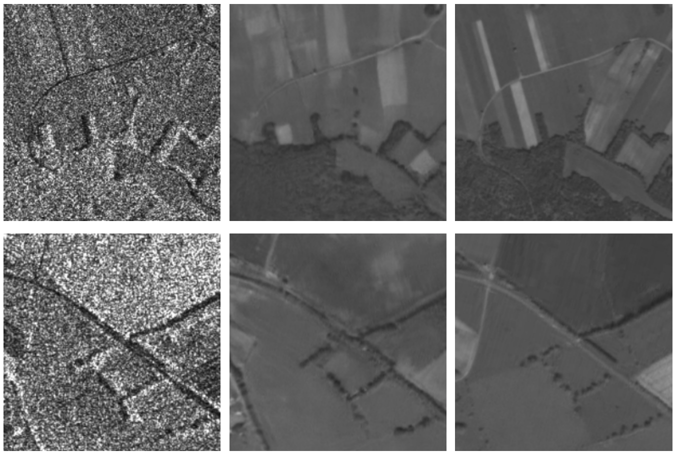

Fig. 5. Side by side comparison between SAR, artificial optical and original optical example patches with a pixel size of 3.75m in two rows.

the matching should be as precise as possible, the test dataset contains only patches with a resolution of 2.5m, which are in total $14,400$ patch pairs. Note that patches extracted from one image are either used for the training or the test dataset.

B. Artificial Image Generation

We investigated several setups for the generation of artificial SAR-like patches. This includes the generation of (despeckled) SAR-like patches at varying scales (pixel size: 2.5m and 3.75m), the training of G through different losses (cGAN, cLSGAN and cWGAN), the training with different batch sizes $(1, 4$ and $40)$ and the training with despeckled and original SAR images as reference. Here, the batch size refers to the number of training instances used in one iteration of the training. For every setup the cGANs are trained over 200 epochs (one epoch refers to one whole cycle through the entire training set) on a single NVIDIA GeForce GTX Titan X GPU. The training time varied from several days to several weeks depending on the batch size, the size of the training dataset and the chosen cGAN setup. An overview of the different training setups can be seen in Table I. Note that all artificially generated patches shown in this paper are generated from test set patches.

Figure 3 shows examples of artificial (despeckled) SAR patch with a pixel size of 2.5m generated by utilizing three different setups: The first setup utilizes the cWGAN, a batch size of 40 and the 2.5m dataset. The second setup utilizes the cGAN, a batch size of 1 and the 2.5m dataset. In contrast to the other two setups, here the filtered SAR images were used for the training procedure. The third setup utilizes the cLSGAN loss, a batch size of 4 and the 2.5m dataset. These examples illustrate that the geometric structures of streets from optical images are preserved in the generated templates, while the radiometric properties are adapted to SAR or despeckled SAR images. The generator learned that, in contrast to optical images, streets normally appear with a lower intensity in SAR images. Furthermore, G tries to represent the characteristics of speckle or the resulting pattern from the speckle filter in case of the first and second setup. A characteristic of the third setup is the blurry appearance of objects such as fields and hence the absence of speckle in the generated patches, which is caused by the utilization of the L_2 loss. The development of the learning process of the generator G of the first (cWGAN) setup over the training time is exemplified by Figure 4.

TABLE I
OVERVIEW OF THE DIFFERENT TRAINING SETUPS.

Setup	dataset	batch size	filter
cGAN	2.5m / 2.5m+3.75m	$1/4/40$	yes / no
cLSGAN	2.5m / 2.5m+3.75m	$1/4/40$	yes / no
cWGAN	2.5m / 2.5m+3.75m	$1/4/40$	yes / no

TABLE II
INFLUENCE OF THE ARTIFICIAL GENERATED TEMPLATES ON THE MATCHING ACCURACY AND PRECISION OF NCC[16], SIFT[17], BRISK[18] AND A COMPARISON WITH TWO BASELINE METHODS. THE MATCHING ACCURACY IS MEASURED AS THE PERCENTAGE OF MATCHING POINTS HAVING A L_2 DISTANCE TO THE GROUND TRUTH LOCATION SMALLER THAN 3 PIXELS, AND AS THE AVERAGE OVER THE L_2 DISTANCES BETWEEN THE PREDICTED MATCHING POINTS AND THE GROUND TRUTH LOCATIONS (MEASURED IN PIXEL UNITS). THE MATCHING PRECISION IS REPRESENTED BY THE STANDARD DEVIATION σ (MEASURED IN PIXEL UNITS).

	matching accuracy		**matching precision**
Methods	< 3 pixels	avg L_2	σ
NCC[16]	35.55%	5.50	4.76
SIFT[17]	31.10%	5.61	1.64
BRISK[18]	39.58%	3.61	1.70
$\text{NCC}_{\text{cLSGAN}}$	75.48%	2.94	5.79
$\text{SIFT}_{\text{cLSGAN}}$	68.85%	2.40	**1.05**
$\text{BRISK}_{\text{cLSGAN}}$	75.21%	2.22	1.10
CAMRI[2]	57.06%	2.80	2.86
DeepMatch[3]	**82.80%**	**1.91**	1.14

Future Prospects: We further considered to reverse the whole process and to generate artificial optical images out of SAR images. An example of such artificial optical images is shown in Figure 5. Despite the reasonable visual appearance the artificial optical images could not improve the later image matching and partly led to a deterioration of the matching results. We attribute this to the fact that optical images reveal a higher level of detail as SAR images and that the extraction and generation of features from SAR images is more difficult as from optical images. Therefore it is more difficult to preserve image features, which is important for a reliable and accurate matching. Nevertheless, this direction provides a possibility for a better interpretation or visual understanding of SAR images for non-experts.

C. Artificial Image Matching

To investigate the influence of the artificial generated patches on the NCC, SIFT and BRISK based matching we evaluated the matching accuracy and precision between SAR and optical image patches and we compared the results with two state-of-the-art methods CAMRI [2] and DeepMatch [3]. Table II gives an overview of the different methods and the corresponding matching accuracies and precisions, which are all evaluated over the same test set. The matching accuracy is measured as the percentage of matching points having a L_2 distance with less than 3 pixels to the ground truth location, and as the average over the L_2 distances between the computed matching points and the ground truth locations (measured in pixel units). The matching precision is represented by the standard deviation σ (measured in pixel units).

The test patches are extracted from 6 different optical and SAR image pairs. Note that we applied the SIFT and BRISK based matching in combination with RANSAC (with an affine model) on the patches of each image scene separately. All artificial patches used to obtain the results from Table II are generated by utilizing the same cLSGAN setup, which is trained on the larger test set (2.5m+3.75m) and with a batch size of 4. This setup let to the best overall results (see later discussion about the influence of the loss function and Table III). For the six image scenes and the application of SIFT and RANSAC we obtained $84, 7, 10, 9, 55, 110$ matching points between the optical and the SAR patches and $235, 120, 70, 25, 363, 286$ matching points between the artificial generated and the SAR patches. For the combination of BRISK and RANSAC we obtained $460, 52, 592, 101, 1409, 687$ points for the matching between the optical and the SAR patches and $697, 393, 520, 164, 3834, 1052$ matching points between the artificial generated and the SAR patches. For the NCC based matching we only considered points with an NCC value of 0.4 or higher as valid matching points and obtained in total 346 points (for all 6 image pairs) for the optical and SAR patch matching and 155 points for artificial and SAR patch matching.

In the case of the SIFT and BRISK batch matching the use of artificial templates increased the number of obtained matching points and in all cases it significantly improved the matching accuracy and precision of the NCC, SIFT and BRISK based matching (see Table II). This is an important requirement for the intended application of the proposed method for the geo-location accuracy improvement of optical images. For this application only few matching points are required for every image scene, but these points have to exhibit a high accuracy and precision.

Influence of the Loss Function: To identify the best setup for our application we investigated the influence of the three different loss functions introduced in Subsection III-B and their dependency on the batch size and the dataset size. An overview of the results of the tested setups is shown in Table I. We achieved the best results for the cGAN and cWGAN setup by training on the smaller dataset and with a batch size of 40 and 1, respectively, and for the cLSGAN setup by training on the larger dataset with a batch size of 4. Table III shows a comparison of the obtained results by applying the three loss function of the cGAN, cLSGAN or cWGAN setup. The setup that generated the best matching results (with respect to the matching accuracy and precision) is the cLSGAN, which utilizes the least square loss. As stated in Section IV-B the utilization of the least square loss causes the absence of artificial speckle in the generated patches. Therefore, the better matching performance of the cLSGAN (compared to the cGAN and cWGAN setups) can be traced back to the fact that the applied matching methods (NCC, SIFT and BRISK) normally suffer from speckle in the image patches. Moreover, since the "real" speckle structure of the SAR patches cannot be derived from the optical patches, it cannot be learned by the generator. As a consequence, the generator network will produce patches, which contain random speckle that looks real enough to "fool" the discriminator network. Overall, the occurrence of artificially speckle in the generated patches makes the matching more difficult.

TABLE III
Influence of loss function on the matching accuracy and precision of NCC[16], SIFT[17], BRISK[18]. The matching accuracy is measured as the percentage of matching points having a L_2 distance to the ground truth location smaller than 3 pixels, and as the average over the L_2 distances between the predicted matching points and the ground truth locations (measured in pixel units). The matching precision is represented by the standard deviation σ (measured in pixel units).

	matching accuracy		**matching precision**
Methods	< 3 pixels	avg L_2	σ
NCC_{cGAN}	30.64%	4.76	4.40
$\text{SIFT}_{\text{cGAN}}$	54.55%	2.84	1.19
$\text{BRISK}_{\text{cGAN}}$	36.48%	4.50	1.63
$\text{NCC}_{\text{cLSGAN}}$	**75.48%**	2.94	5.79
$\text{SIFT}_{\text{cLSGAN}}$	68.85%	2.40	**1.05**
$\text{BRISK}_{\text{cLSGAN}}$	75.21%	**2.22**	1.10
$\text{NCC}_{\text{cWGAN}}$	24.00%	6.51	4.08
$\text{SIFT}_{\text{cWGAN}}$	56.51%	2.89	1.39
$\text{BRISK}_{\text{cWGAN}}$	58.06%	3.08	1.30

Influence of the Speckle Filter: The application of a speckle filter is an important pre-processing step for many matching methods and is used to improve the results of CAMRI [2] and DeepMatch [3]. Therefore we exploited two application cases of the speckle filter. First, we investigated the influence of the despeckled SAR patches on the NCC, SIFT and BRISK based matching (without the use of cGANs). Only in the case of the BRISK based matching the usage of the speckle filter led to an improvement of the matching results (# matching points < 3 pixels = 52.21%, avg $L_2 = 3.00$, $\sigma = 1.37$). Second, we investigated the generation of SAR-like despeckled patches via cGANs and their influence on the NCC, SIFT and BRISK based matching. Utilizing these patches led in none of the matching setups to better matching results compared to the matching using SAR-like artificial patches. We trace this back to the fact that even if the texture of the speckle filter is well imitated (as illustrated in the second row of Figure 3) it is randomly generated and independent from the real image objects or their properties and therefore led to unreliable matching results.

Comparison with Baselines: For a better assessment of the quality of the results a comparison with two state-of-the-art approaches is carried out. By applying the SIFT and BRISK based matching we can achieve better results than the first baseline called CAMRI [2]. CAMRI is a mutual information based method and is tailored to the problem of optical and SAR images matching. The second baseline, called DeepMatch [3], is a deep learning based matching approach,

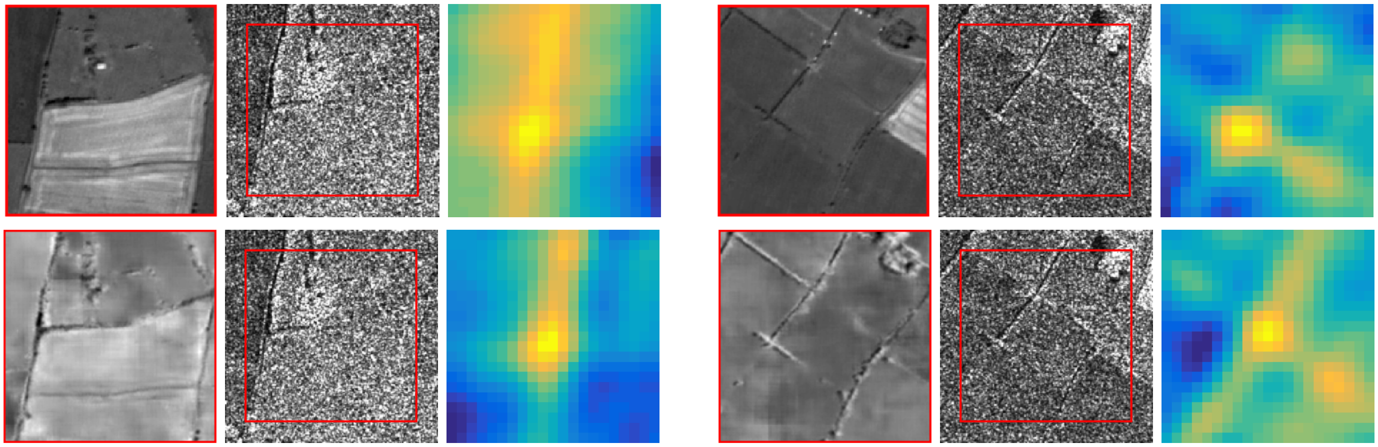

Fig. 6. Comparison of the score maps between the NCC based matching of the optical image with the SAR image and the generated template (from the optical image) with the despeckled SAR image (from top down and in two columns).

where the matching between optical and SAR images is realized through a trained Siamese network. Regarding the accuracy of the matching points DeepMatch achieves better results than applying a SIFT and BRISK based matching between the artificial patches and the SAR patches and in regard to the matching precision our proposed approach achieves slightly better results compared to DeepMatch.

Qualitative Results of NCC: Figure 6 shows a qualitative comparison of the NCC based matching between optical and SAR patches, and generated templates and SAR patches. The search space is $\Delta_x = \Delta_y = 20$ pixels in each direction around the center position. The used templates are generated by using the best setup (cLSGAN setup with the least square loss). The correct matching positions are for all examples in the center of the SAR patches. The brighter the color of the score map, the higher is the NCC value at the corresponding location. The examples emphasize that the generated SAR-like templates can improve the matching between SAR and optical images through NCC.

Limitations: A problem of (conditional) GANs is the difficult validation of the training success. In contrast to other machine learning architectures, where the loss function or different metrics can be used to evaluate the quality of the training progress over a validation set, GANs require an evaluation mainly via the visual quality of the generated images or (in our case) the evaluation of the matching results. This is time consuming, since every setup has to be trained till the end to find the best one. A further time consuming task is the training of the cGANs, which takes from some days to several weeks. Besides high computational cost of the network training and data quality evaluation, the experiments revealed that it is important to generate patches which retain the geometric structures of the optical patches instead of generating patches which visually look like real SAR images (see Table III and corresponding discussion). Therefore, not every loss and cGAN setup is applicable for the problem of optical and SAR image matching.

Strengths: An advantage of the proposed method is that it enables the application of well know matching techniques (NCC, BRISK and SIFT) on the problem of matching optical and SAR images. These three matching methods proved their high quality for the matching of images acquired from the same sensor (e.g. NCC for SAR to SAR matching [34] and SIFT and BRISK for matching optical images [35]), but normally fail in the case of optical and SAR images. The evaluation of the results and the comparison with two state-of-the-art matching approaches revealed the potential of the proposed method and the possibility to apply it for the problem of absolute geo-location accuracy improvement of optical images. A further benefit is the fast applicability of the proposed method to new image scenes once the generator is trained. For the matching of new images scenes artificial SAR-like patches can be generated within minutes from given optical patches. Furthermore, through the variety in our training dataset, which contains images acquired at different times of the year and over different locations in Europe, our proposed approach is applicable to a wide range of images acquired over different countries.

Future Prospects: For the future the proposed method could be further improved by utilizing the sensor model of the input image for RANSAC instead of an affine model. The affine model works well for relatively flat areas but is not suitable for every image scene. Moreover, the investigation of different generator architectures represents a further interesting investigation. Another possible enhancement for the future is the combination of the proposed technique with DeepMatch[3]. So far, the training of the cGANs is geared to the problem of generating images, which look realistic enough to ”fool” the discriminator. The results reveal that patches, which look more like real SAR images not necessarily lead to better matching results. Therefore, it is more important to preserve features such as edges or corners, which are beneficial for a matching technique, in the artificial patches. By replacing the discriminator with the Siamese matching network proposed in [3] the training of the generator G could be tailored towards the problem of generating artificial patches, which lead to better matching results than using the original optical patches. This combination represents a promising development for the future to further improve the results obtained by the proposed method and by DeepMatch.

V. Conclusion

We proposed a new concept for the problem of multi-sensor image matching based on conditional generative adversarial networks (cGANs). Different cGANs setups are trained for the task of generating SAR-like image patches from optical images. We showed the feasibility to improve the matching accuracy and precision of a NCC, SIFT and BRISK based matching between optical and SAR image patches by artificial generated patches. By applying BRISK for the matching of SAR and artificial SAR-like patches we achieve matching points with an average L_2 distance to the ground truth locations of 2.22 pixels and a precision (standard deviation) of 1.10 pixels. The results further validate the potential of the proposed approach in comparison to two state-of-the-art methods but also revealed the need for further enhancements of the proposed method. Especially, the necessity for a generator network, which reliably and precisely retain the geometric structures of the optical images, should be the main focus of further investigations. Overall, the proposed method opens up new possibilities for future developments towards the goal of matching optical and SAR images. The combination of a generator network with a deep learning based matching approach represents thereby a promising future extension to generate even more suitable artificial images patches and hence, to further improve the quality of the image matching.

References

[1] M. Eineder, C. Minet, P. Steigenberger, X. Cong, and T. Fritz, "Imaging Geodesy - Toward Centimeter-Level Ranging Accuracy With TerraSAR-X," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 2, pp. 661–671, 2011.

[2] S. Suri and P. Reinartz, "Mutual-Information-Based Registration of TerraSAR-X and Ikonos Imagery in Urban Areas," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 2, pp. 939–949, 2010.

[3] N. Merkle, W. Luo, S. Auer, R. Müller, and R. Urtasun, "Exploiting Deep Matching and SAR Data for the Geo-Localization Accuracy Improvement of Optical Satellite Images," *Remote Sensing*, vol. 9, no. 6, 2017.

[4] L. G. Brown, "A Survey of Image Registration Techniques," *ACM Computing Surveys*, vol. 24, no. 4, pp. 325–376, 1992.

[5] B. Zitov and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, pp. 977–1000, 2003.

[6] Z. Xiong and Y. Zhang, "A Critical Review of Image Registration Methods," *International Journal of Image and Data Fusion*, vol. 1, no. 2, pp. 137–158, 2010.

[7] W. Shi, F. Su, R. Wang, and J. Fan, "A Visual Circle Based Image Registration Algorithm for Optical and SAR Imagery," in *IEEE International Geoscience and Remote Sensing Symposium*, Munich, Germany, July 2012, pp. 2109–2112.

[8] M. A. Siddique, M. S. Sarfraz, D. Bornemann, and O. Hellwich, "Automatic Registration of SAR and Optical Images Based on Mutual Information Assisted Monte Carlo," in *IEEE International Geoscience and Remote Sensing Symposium*, Munich, Germany, July 2012, pp. 1813–1816.

[9] M. Hasan, M. R. Pickering, and X. Jia, "Robust Automatic Registration of Multimodal Satellite Images Using CCRE with Partial Volume Interpolation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 10, pp. 4050–4061, 2012.

[10] R. S. T. Hong, "A Robust Technique for Precise Registration of Radar and Optical Satellite Images," *Photogrammetric Engineering & Remote Sensing*, vol. 71, no. 5, pp. 585–593, 2005.

[11] H. Li, B. S. Manjunath, and S. K. Mitra, "A Contour-Based Approach to Multisensor Image Registration," *IEEE Transactions on Image Processing*, vol. 4, no. 3, pp. 320–334, 1995.

[12] C. Pan, Z. Zhang, H. Yan, G. Wu, and S. Ma, "Multisource Data Registration Based on NURBS Description of Contours," *International Journal of Remote Sensing*, vol. 29, no. 2, pp. 569–591, 2008.

[13] P. Dare and I. Dowmanb, "An Improved Model for Automatic Feature-Based Registration of SAR and SPOT Images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 56, no. 1, pp. 13 – 28, 2001.

[14] B. Fan, C. Huo, C. Pan, and Q. Kong, "Registration of Optical and SAR Satellite Images by Exploring the Spatial Relationship of the Improved SIFT," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 4, pp. 657–661, 2013.

[15] Y. Han and Y. Byun, "Automatic and Accurate Registration of VHR Optical and SAR Images Using a Quadtree Structure," *International Journal of Remote Sensing*, vol. 36, no. 9, pp. 2277–2295, 2015.

[16] W. Burger and M. J. Burge, *Principles of Digital Image Processing: Core Algorithms*. Springer Publishing Company, Incorporated, 2009.

[17] D. G. Lowe, "Object Recognition from Local Scale-Invariant Features," in *Proceedings of the International Conference on Computer Vision*, vol. 2. Washington, DC, USA: IEEE Computer Society, 1999, pp. 1150–1158.

[18] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary Robust Invariant Scalable Keypoints," in *Proceedings of the 2011 International Conference on Computer Vision*. Washington, DC, USA: IEEE Computer Society, 2011, pp. 2548–2555.

[19] N. Merkle, P. Fischer, and R. M. S. Auer, "On the Possibility of Conditional Adversarial Networks for Multi-Sensor Image Matching," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Fort Worth, Texas, US, July 2017.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems NIPS*, Montreal, Canada, December 2014, pp. 2672–2680.

[21] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic Segmentation using Adversarial Networks," in *NIPS Workshop on Adversarial Training*, Barcelona, Spain, December 2016, pp. 1–12.

[22] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," pp. 105–114, 2017.

[23] D. Nie, R. Trullo, C. Petitjean, S. Ruan, and D. Shen, "Medical Image Synthesis with Context-Aware Generative Adversarial Networks," in *Medical Image Computing and Computer-Assisted Intervention*. Quebec City, QC, Canada: Springer International Publishing, September 2017, pp. 417–425.

[24] J. Guo, B. Lei, C. Ding, and Y. Zhang, "Synthetic Aperture Radar Image Synthesis by Using Generative Adversarial Nets," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 7, pp. 1111–1115, July 2017.

[25] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, US, July 2017.

[26] M. Bossard, J. Feranec, and J. Otahel, "CORINE Land Cover Technical Guide - Addendum 2000," *European Environmental Agency, Copenhagen, Denmark*, 2000.

[27] X. Mao, Q. Li, H. Xie, R. Lau, Z. Wang, and S. Smolley, "Least Squares Generative Adversarial Networks," in *The IEEE International Conference on Computer Vision*, Venice, Italy, Oct 2017.

[28] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70. International Convention Centre, Sydney, Australia: PMLR, August 2017, pp. 214–223.

[29] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations*, San Diego, California, US, May 2015, pp. 1–13.

[30] T. Tieleman and G. Hinton, " Lecture 6.5-RMSProp: Divide the Gradient by a Running Average of Its Recent Magnitude," *COURSERA: Neural Networks for Machine Learning*, 2012.

[31] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[32] M. Schneider, R. Müller, T. Krauss, P. Reinartz, B. Hörsch, and S. Schmuck, "Urban Atlas - DLR Processing Chain for Orthorectification of Prism and AVNIR-2 Images and TerraSAR-X as possible GCP Source," in *Internet Proceedings: 3rd ALOS PI Symposium*, Kona, Hawaii, USA, Nov 2009, pp. 1–6.

[33] C. Deledalle, L. Denis, and F. Tupin, “Iterative Weighted Maximum Likelihood Denoising with Probabilistic Patch-Based Weights,” *IEEE Transactions on Image Processing*, vol. 18, no. 12, pp. 2661–2672, 2009.

[34] Y. Wang, Q. Yu, and W. Yu, “An Improved Normalized Cross Correlation Algorithm for SAR Image Registration,” in *IEEE International Geoscience and Remote Sensing Symposium*, Munich, Germany, July 2012, pp. 2086–2089.

[35] P. Schwind and P. dAngelo, “Evaluating the Applicability of BRISK for the Geometric Registration of Remote Sensing Images,” *Remote Sensing Letters*, vol. 6, no. 9, pp. 677–686, 2015.

Nina Merkle received the B.Sc. and M.Sc. degrees in Mathematics from the University of Applied Science Regensburg, Regensburg, Germany, in 2011 and 2013, respectively. Since 2014, she has been a research fellow and PhD candidate with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany. In spring and autumn 2016, she was a guest scientist at the Computer Science Department of the University of Toronto, Toronto, Ontario, Canada. Her current research interests include image matching and registration of multi-sensor satellite data and machine learning techniques, especially deep learning and its applications in the field of remote sensing.

Stefan Auer received the Dipl.-Ing.(Univ.) degree in Geodesy in 2005 and the Dr.-Ing. degree in Remote Sensing in 2011 from Technical University of Munich (TUM), Munich, Germany. Since December 2014, he has been a senior researcher and project manager at the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany. Besides managing projects in the fields of multispectral and hyperspectral imaging, he works on the alignment of multi-modal remote sensing data. Beyond, his interest is in 3D simulation / data fusion techniques and the interpretation of high-resolution SAR and optical images. Stefan Auer developed the open-source 3-D SAR simulator RaySAR which can be used to better understand the nature of prominent SAR image signatures based on object models. In the context of his doctoral thesis, Stefan Auer spent three months as guest researcher at the Department of Electronic and Telecommunication Engineering (DIET) at the University of Naples ”Federico II” in Italy.

Rupert Müller received the Dipl.-Phys. degree in physics from the Ludwig Maximilians University of Munich, Germany. He is a Team Leader of the Processors and Traffic Monitoring group, Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling, Germany, and is currently project manager of the Ground Segment of MUSES (Multi-User System for Earth Sensing) with the hyperspectral instrument DESIS (DLR Earth Sensing Imaging Spectrometer) to be installed on the International Space Station (ISS). His main research interests include photogrammetric evaluation of optical data from air- and spaceborne sensors, digital image processing, machine learning and hyperspectral imaging.

Peter Reinartz (M’09) received his Diploma (Dipl.-Phys.) in theoretical physics in 1983 from the University of Munich and his PhD (Dr.-Ing) in civil engineering from the University of Hannover, in 1989. His dissertation is on optimization of classification methods for multispectral image data. Currently he is department head of the department Photogrammetry and Image Analysis at the German Aerospace Centre (DLR), Remote Sensing Technology Institute (IMF) and holds a professorship for computer science at the University of Osnabrueck. He has more than 30 years of experience in image processing and remote sensing and over 400 publications in these fields. His main interests are in machine learning, stereo-photogrammetry and data fusion using space borne and airborne image data, generation of digital elevation models and interpretation of very high resolution data from sensors like WorldView, GeoEye, and Pleiades. He is also engaged in using remote sensing data for disaster management and using high frequency time series of airborne image data for real time image processing and for operational use in case of disasters as well as for traffic monitoring.