ON THE POSSIBILITY OF CONDITIONAL ADVERSARIAL NETWORKS FOR MULTI-SENSOR IMAGE MATCHING

N. Merkle, P. Fischer, S. Auer, R. Müller

German Aerospace Center (DLR) , Remote Sensing Technology Institute
Oberpfaffenhofen, 82234 Weßling, Germany

ABSTRACT

A major research area in remote sensing is the problem of multi-sensor data fusion. Especially the combination of images acquired by different sensor types, e.g. active and passive, is a difficult task. Over the last years deep learning methods have proven their high potential for remote sensing applications. In this paper we will show how a deep learning method can be valuable for the problem of optical and SAR image matching. We investigate the possible of conditional generative adversarial networks (cGANs) for the generation of artificial templates. Contrary to common template generation approaches for image matching, the generation of templates using cGANs does not require the extraction of features. Our results show the possibility of realistic SAR-like template generation from optical images through cGANs and the potential of these templates for enhancing the matching of optical and SAR images by means of reliability and accuracy.

Index Terms— conditional GANs, deep learning, image matching, multi-sensor, artificial template generation

1. INTRODUCTION

More and more research studies successfully apply deep learning methods to remote sensing problems, like classification of hyperspectral [1] or SAR images [2], the enhancement of road maps [3], or the usage of a deep matching network for the task of aerial image matching [4]. These studies show the opportunities given by the development of deep learning for remote sensing problems. One major topic in remote sensing is the matching and fusion of multi-sensor data. Finding corresponding and reliable features in different data sources is a difficult task. Previous works like [5], show promising results concerning the problem of optical and SAR image matching through templates. The drawback of such template based approaches is the need for extracting a suitable number of features, for instance the geometry of man-made infrastructure like intersection or a roundabout. In this paper, we investigate an approach for a general generation of templates. More precisely, we investigate the generation of SAR-like templates from optical image through conditional adversarial networks.

Generative adversarial networks (GANs) were introduced by Goodfellow et al. [6]. The original idea of GANs was the estimation of generative models through an adversarial process. An interesting development of GANs is shown by Isola et al. [7]. They present a solution for image-to-image translation problems based on conditional GANs (cGANs). Our template generation method is based on this idea and the neural network introduced in [7]. Our aim is to generate templates with geometric properties from an input image and having the simulated style/radiometric properties from a reference image. A possible application for these artificial templates is the problem of SAR and optical image matching.

2. METHODOLOGY

Our proposed strategy for the application of cGANs to SAR and optical image matching is: (1) find a patch I in the optical image, which contains salient features, (2) pre-process I, (3) generate the template T from I with a cGAN, (4) use a similarity measure like normalized cross-correlation (NCC) or a feature detection approach like the scale-invariant feature transform (SIFT) for the matching of the template T with the reference image R.

2.1. Patch Extraction

Due to different radiometric and geometric properties of SAR and optical sensors, the detection of reliable features and the establishment of a suitable transfer function for comparing features between both image modalities is a non-trivial task. Features with a certain height, like buildings, might have a different appearance in SAR and optical images. For the SAR-like template generation, the optical patches should contain planar objects, which have at least to a certain degree the same geometric appearance as in the corresponding SAR patches. Suitable features are in most cases man-made infrastructure objects, e.g. streets and street crossings, roundabouts or runways. For limiting the search space in the image, the optical image patches which contain such features are preselected from areas where the CORINE land cover layer [8] indicates the existence of fitting patterns. To ensure that only patches are selected, which contain features visible in the optical and SAR patches and to exclude patches containing small villages, the pre-selection was refined manually.

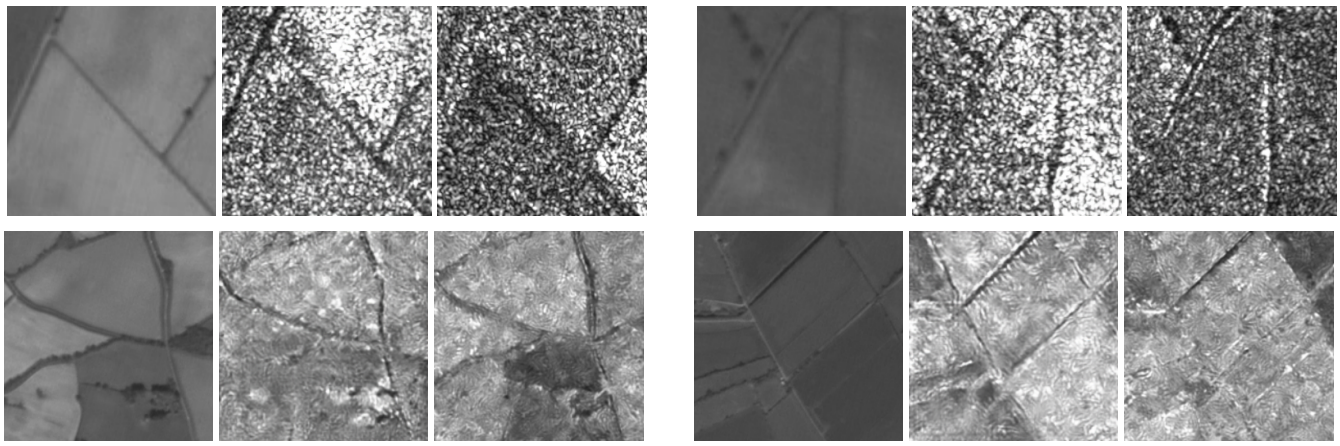

Fig. 1. Side by side comparison between optical, artificial SAR and original (despeckled) SAR images in two columns.

2.2. Patch Preprocessing

The appearance of speckle in SAR images has a strong influence on common matching approaches between SAR and optical images. Therefore, we investigate the generation of despeckled SAR templates T. For the despeckling, the non-local SAR filter proposed in [9] was applied.

2.3. Artificial Template Generation

A generative adversarial network (GANs) consists of a generator network G and, its counterpart, a discriminator network D. The generator G is trained to generate images from random noise. More specific, G is trained to learn the mapping from a random vector z to an output image y. The discriminator D is a classification network and is trained to distinguish between real images and images generated by G. G tries to generate as realistic images as possible to fool D. Conditional GANs use, next to the input z, an observed image x. The overall aim is to optimize the cGAN loss

$$\begin{aligned}\min_G \max_D \mathcal{L}_{\text{cGAN}}(G,D) = E_{x,y\sim p_{\text{data}}(x,y)}[\log D(x,y)]+ \\ E_{x,y\sim p_{\text{data}}(x,y),z\sim p_z(z)}[\log(1-D(x,G(x,z)))],\end{aligned} \tag{1}$$

where E denotes the expected value. In our case x is an optical image patch, y the corresponding SAR patch (the ground truth image) and $G(x,z)$ a SAR-like template. G tries to minimize the loss $\mathcal{L}_{\text{cGAN}}$ and fool D as often as possible, whereby D tries to maximize the loss and detect as many fake images (generated by G) as possible. Isola et al. [7] suggest to include an additional term to the loss $\mathcal{L}_{\text{cGAN}}$, to force D to produce output images, which are close to the ground truth images y (in sense of the L_1 distance). Adding this term, the final objective is

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{cGAN}}(G,D) + \lambda \mathcal{L}_{L_1}(G), \tag{2}$$

where the term $\mathcal{L}_{L_1}$ is defined as

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

For a detailed description of GANs see [6] and for a detailed overview of the network architecture see [7].

The network is trained on pairs of optical and SAR image patches with a size of 201×201. As in [7] we use mini-batch stochastic gradient descent with an alternated and separated training of G and D. Particularly, one gradient descent step of G is followed by one gradient descent step of D. Through the optimization of the objective function during training, the network is able to learn how to generate SAR-like templates from optical image patches. The network implementation is realized through the open source code to the paper [7].

2.4. Template Matching

One possibility to match the generated template with the corresponding reference image is to use a similarity measure. In this paper the applicability of NCC is investigated, where the similarity between a template T and the corresponding reference image R at location (m,n) is defined as

$$\rho(m,n) = \frac{\sum_{x,y}(R(m+x,n+y)-\overline{R})(T(x,y)-\overline{T})}{\sqrt{\sum_{x,y}(R(m+x,n+y)-\overline{R})^2(T(x,y)-\overline{T})^2}}. \tag{4}$$

Here, $R(m+x,n+y)$ and $T(x,y)$ are the intensity values at position (x,y) and $\overline{R}$ and $\overline{T}$ are the mean intensity values of the image patches R and the T. In our case, T is a generated SAR-like template with size $N \times N$ and R the reference patch cropped from the SAR image with size $(N+2*\Delta_x)\times(N+2*\Delta_y)$, where Δ_x and Δ_y is the search space in x- and y-direction. The NCC is calculated for all possible positions of T within the larger patch R. The position with the highest NCC value is the position with the best match between T and R. To improve the result of the described matching procedure, the score of Equation (4) can be use to remove outliers, e.g consider only points as valid matching points with a score of 0.5 or higher. Furthermore, the usage of a hybrid evolutionary algorithm enables to lower the computational cost and, hence, speed up the template matching without loss in accuracy.

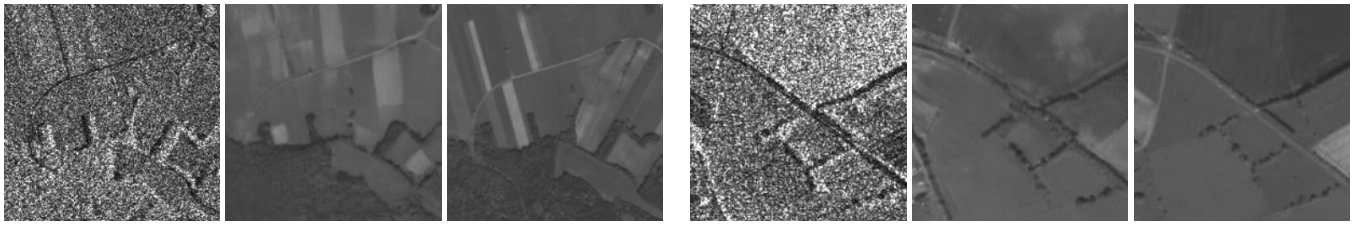

Fig. 2. Side by side comparison between SAR, artificial optical and original optical images in two columns.

Feature-based approaches provide a further possibility to match the artificial templates with the reference images. The key stage of a feature-based approach is the detection of reliable and corresponding features in the artificial template and reference SAR image. In section 4 we will provide some first results, where SIFT and BRISK was utilized to detect and match features from T and R. At the end, RANSAC was applied to increase the quality of the obtained matching points and to estimate the shift between T and R.

3. DATASET

To train and test the cGANs, we generated a dataset out of 46 orthorectified optical (PRISM) and radar (TerraSAR-X) satellite image pairs. The images have been acquired over 13 cities in Europe and cover greater urban zones including suburban, industrial and rural areas. The pixel spacing of the PRISM images is 2.5m. To be consistent with the optical images, we used bilinear interpolation to resample the spatial resolution of the TerraSAR-X images from 1.25m to 2.5m. The optical and the SAR image pairs were manually aligned within the Urban Atlas project [10] and have an overall alignment error of around 3m. The training dataset consists of $69,990$ and the test dataset of $5,171$ pairs of optical and SAR patches. All training and test patches are semi-manually extracted from specific areas in the images (as described in Subsection 2.1). The applied CORINE land cover layer has a pixel size of 100m and is from the year 2012.

4. EXPERIMENTAL EVALUATION

4.1. Artificial Template Generation

Our results of the template generation through cGANs can be seen in Figures 1-2. All figures show examples of test patches, which are never shown to the network during training. Figure 1 shows a side by side comparison of optical patches, the generated SAR-like templates and the original (despeckled) SAR patches. The examples prove that the geometric structure of streets (crossings) from the optical images is preserved in the generated templates. Furthermore, the templates show radiometric properties of SAR or despeckled SAR images. The cGAN learned that in contrast to optical images, streets normally appear with a lower intensity in SAR images. The network also tries to represent the characteristics of speckle or the resulting pattern from the speckle filter.

We also trained the network to generate optical templates from SAR images. The results can be seen in Figure 2. Since optical images reveal a higher level of detail as SAR images, the generation of artificial optical templates, is more difficult. Furthermore, the extraction of features from SAR images is more difficult. Particularly with regard to the aim of matching optical and SAR images, where it is important to preserved image features as good as possible, our focus is on the SAR-like template generation.

4.2. Template Matching

We investigated the applicability of the generated templates for optical and SAR image matching. A qualitative comparison of the matching between optical and SAR patches, and generated templates and SAR patches through NCC is illustrated in Figure 3. The patch size is 201×201 and the search space is $\Delta_x = \Delta_y = 20$ pixels in each direction. The used templates are generated by using only the L_1 loss from Equation (3) to train the network. The correct matching positions are in the center of the SAR patches. A bright color in the score map is related to a high NCC value. The examples emphasize that the generated SAR-like templates can improve the matching between SAR and optical images through NCC. To confirm this assumption a quantitative evaluation of the matching results is depicted in Table 1. Here, we investigated the influence of the templates in the matching accuracy and precision. The numbers are computed from the obtained matching points with a similarity score higher than 0.5 (122 matches with original images vs. 193 matches with templates). The matching accuracy is measured through the percentage of matching points, where the L_2 distance to the ground truth location is less than 3 pixels and the average L_2 distance. The matching precision is measured through the standard deviation σ.

We further evaluated our first results of a featured-based matching approach using SIFT and BRISK in combination with RANSAC on one image from our test data. We used RANSAC to remove outliers and to get a predicted shift between the SAR and optical patches. By applying SIFT and RANSAC we obtained 24 matches out of 1710 patches and an estimated shift (in pixel units) of 2.63 in x- and 16.84 in y-direction between the optical and SAR patches, and 15 matches and an estimated shift of 3.48 in x- and -1.78 in y-direction between the templates and the SAR images (correct shift: $x = 0$ and $y = 0$). Applying BRISK and RANSAC led

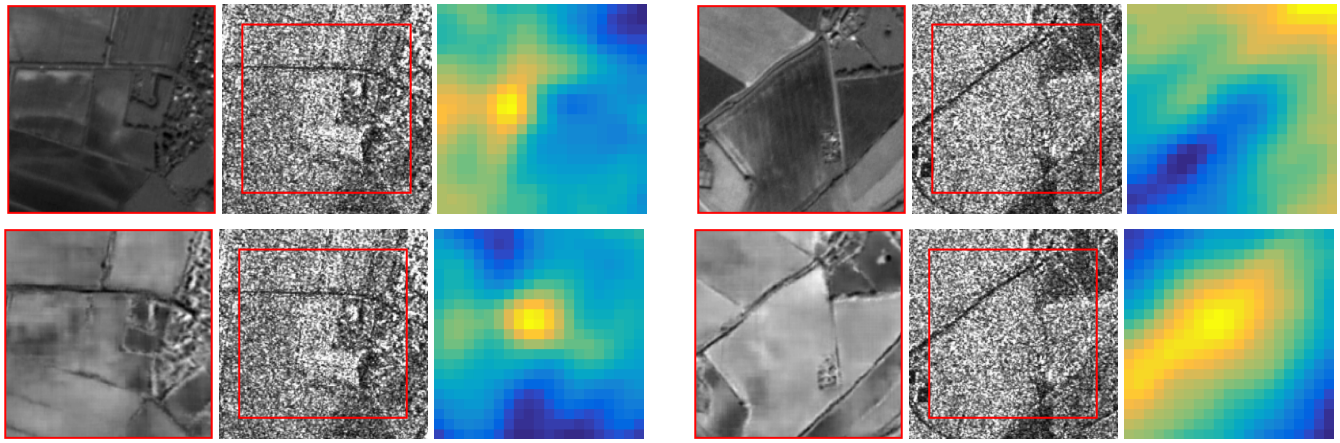

Fig. 3. Comparison of the score maps between the NCC based matching of the optical image with the SAR image and the generated template (from the optical image) with the despeckled SAR image (from top down and in two columns).

to 93 matches and an estimated shift of 1.82 in x- and -2.6 in y-direction between the optical and SAR images and 120 matches and an estimated shift of -0.42 in $x-$ and -0.65 in y-direction between the templates and the SAR images. Using the artificial templates significant improved the quality of the results in both cases.

	matching accuracy		**matching precision**
Methods	<3 pixels	avg L_2	σ
Original	44.26%	3.88	2.59
Template	84.46%	2.20	2.53

Table 1. Influence of the artificial template on the matching accuracy and precision utilizing the similarity measure NCC. The matching accuracy is measured through the percentage of matching points, where the L_2 distance to the ground truth location is less than 3 pixels. The average L_2 distances and the standard deviation σ are measured in pixel.

5. CONCLUSION

In this paper the applicability of a deep learning method for the generation of artificial templates from optical and SAR images is presented. Furthermore, the possibility of using such templates for the problem of optical and SAR image matching is evaluated. The essential part of the method is the artificial template generation from optical images by applying a conditional adversarial network. This network enables the generation of templates without any need of a prior extraction of features. A possible application of the template generation through cGANs is the problem of optical and SAR image matching. We reveal the successful application of artificial SAR-like templates (generated from optical images) to improve the matching accuracy between optical and SAR images. The matching accuracy enhancement is shown for similarity- and feature-based approaches. In the future, the applicability to feature-based matching approaches like SIFT or BRISK will be further investigated on a larger test dataset.

6. REFERENCES

[1] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep Learning-Based Classification of Hyperspectral Data," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 7, no. 6, pp. 2094–2107, June 2014.

[2] J. Geng, J. Fan, H. Wang, X. Ma, B. Li, and F. Chen, "High-Resolution SAR Image Classification via Deep Convolutional Autoencoders," *Geoscience and Remote Sensing Letters, IEEE*, vol. 12, no. 11, pp. 2351–2355, Nov 2015.

[3] G. Matthyus, S. Wang, S. Fidler, and R. Urtasun, "Enhancing World Maps by Parsing Aerial Images," *IEEE International Conference on Computer Vision*, pp. 1689–1697, Dec 2015.

[4] H. Altwaijry, J. Trulls, E.and Hays, P. Fua, and S. Belongie, "Learning to Match Aerial Images with Deep Attentive Architectures," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] N. Merkle, R. Müller, and P. Reinartz, "Registration of Optical and SAR Satellite Images Based on Geometric Feature Templates," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XL-1-W5, pp. 447–452, Nov 2015.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B.g Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *NIPS*, 2014.

[7] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," *arxiv*, 2016.

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

[9] C. Deledalle, L. Denis, F. Tupin, A. Reigber, and M. Jäger, "NL-SAR: A Unified Nonlocal Framework for Resolution-Preserving (Pol)(In)SAR Denoising," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2021–2038, 2015.

[10] 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," *Internet Proceedings: 3rd ALOS PI Symposium*, pp. 1–6, Jan 2010.