DEEP LEARNING FOR SAR-OPTICAL IMAGE MATCHING

Lloyd Haydn Hughes[1], Nina Merkle[2], Tatjana Bürgmann[3], Stefan Auer[2], Michael Schmitt[1]

[1] Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany
[2] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany
[3] Airbus Defence and Space GmbH, Immenstaad, Germany

ABSTRACT

The automatic matching of corresponding regions in remote sensing imagery acquired by synthetic aperture radar (SAR) and optical sensors is a crucial pre-requesite for many data fusion endeavours such as target recognition, image registration, or 3D-reconstruction by stereogrammetry. Driven by the success of deep learning in conventional optical image matching, we have carried out extensive research with regard to deep matching for SAR-optical multi-sensor image pairs in the recent past. In this paper, we summarize the achieved findings, including different concepts based on (pseudo-)siamese convolutional neural network architectures, hard negative mining, alternative formulations of the underlying loss function, and creation of artificial images by generative adversarial networks. Based on data from state-of-the-art remote sensing missions such as TerraSAR-X, Prism, Worldview-2, and Sentinel-1/2, we show what is already possible today, while highlighting challenges to be tackled by future research endeavors.

Index Terms— Deep Learning, Image Matching, Optical Images, SAR Images, Data Fusion

1. INTRODUCTION

Synthetic aperture radar (SAR) and optical sensors comprise the two most important modalities for spaceborne Earth observation, as they provide complementary information about observed scenes. While SAR measures physical surface characteristics such as roughness or water content, optical images encode information about the nature of the surface materials. For that reason, SAR-optical data fusion has become a highly relevant research topic [1]. However, an important prerequesite for any fusion undertaking is to first match corresponding image parts. Due to the severe differences in the two sensor modalities' imaging geometries, this is a non-trivial problem. Driven by the success of deep learning for image matching in classic computer vision, we have been working on deep matching solutions for the SAR-optical multi-sensor case for quite some years. This paper summarizes the findings achieved so far and produces perspectives for ongoing challenges and future research directions. To this end, deep learning approaches used in our group are introduced and matching results presented, followed by a discussion related to the matching task.

2. CNN-BASED SAR-OPTICAL IMAGE MATCHING

The first approaches in SAR-optical deep matching were inspired by the classic siamese and pseudo-siamese convolutional neural network (CNN) architectures that have shown great predictive power for conventional optical image pairs (cf., for example, [2]). Of course, due to the much stronger differences in the two image types, SAR-optical image matching requires domain-specific adaptions.

2.1. (Pseudo-)Siamese Architectures

Both in [3] and [4], (pseudo-)siamese network architectures were proposed that were supposed to predict patch correspondence for SAR-optical patch pairs based on a model learned from a sufficient number of examples. The principle of the networks is depicted in Fig. 1. The SAR and the optical input patches are processed by (modality-specific) CNNs which extract representative features, these are then compared by the final layers of the network.

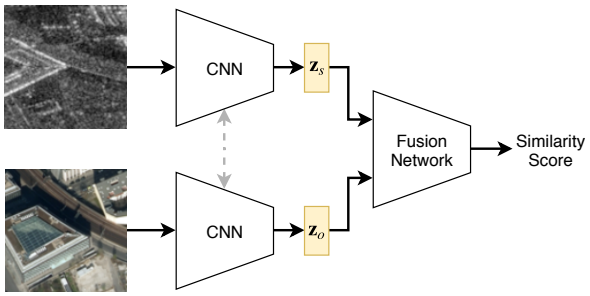


Fig. 1: An overview of the (pseudo-)siamese architecture proposed for SAR-optical image matching and registration. The weights between the two parallel CNN streams are shared in the case of a Siamese architecture, but not in the case of the pseudo-Siamese architecture.

Despite the similar network structure, both approaches focus on the matching of different image scenes and pursue different applications. In [3] the matching of rural and suburban

images, for the generation of tie points in order to improve the geo-localization of optical images, formed the basis of the investigations. Therefore, a siamese network was trained to directly learn the shift between optical and SAR patches. More specifically, the dot product was utilized to measure the similarity between the extracted features. The problem was formulated as a multi-class classification task, whereby each class represented a possible shift of the optical patch within the larger SAR patch. This formulation was then trained using a smooth cross entropy loss. The geo-localization accuracy improvement of optical images was ultimately achieved by adjusting the corresponding optical sensor model parameters through a set of tie points generated by the network.

In contrast, in [4] matching was performed on very high resolution (VHR) imagery of urban areas, in order to determine a point-wise similarity score which could be used in key point matching scenarios. Instead of a weight-sharing siamese network, a pseudo-siamese architecture was selected to allow each stream to learn modality specific features. The final feature maps of these two streams were then concatenated and used as input to a so-called fusion network to determine if the SAR and optical image patches correspond. The full network (pseudo-siamese streams and fusion network) were trained in an end-to-end manner as a binary classification task using a binary cross-entropy loss. The soft-max activation of the final layer was then adopted as a measure of similarity for patch matching applications.

Overall, the utilized (pseudo-)siamese networks exhibit a large number of trainable parameters and therefore require a significant amount of training data. However, the generation of such training data commonly requires a time and cost-intensive manual selection of tie points or corresponding patches. Especially for VHR data of complex urban scenes, the manual selection of tie points is often not feasible at all, because the strongly different appearance of the two image types renders matching impossible even for human experts.

2.2. Enhancing Predictive Power by Triplet Loss and Hard Negatives

Besides relying on imagery of semi-urban and rural scenes for easier generation of layover- and shadow-free training data, we also dealt with the introduction of architectural elements dedicated to supporting more efficient training in situations of training data scarcity. In [5], we investigated an adaption of the HardNet architecture [6], which combines two independent CNN streams for modality-specific feature learning with a triplet loss formulation and the exploitation of hard negative examples. The triplet loss is based on the utilization of triplets of image patches - where for a reference patch, called the anchor, both a positive and negative partner are provided. These are selected through hard negative mining and the loss is calculated as the difference of the positive (between anchor and positive) and negative (between anchor and negative) descriptor distances plus a specified margin. Hard negative mining refers to choosing a negative or non-matching patch that is the most similar to the positive or matching patch and hence minimizing the Euclidean distance between the matching descriptor and the closest non-matching descriptor [6]. Thus, the network is trained to better discriminate between matching and non-matching image patches. In addition, experiments were carried out regarding transfer learning from lower resolution to higher resolution data. Both our HardNet adaption and a classic pseudo-siamese architecture showed significant potential here, which indicates that pre-training networks on large-scale medium-resolution datasets such as SEN1-2 [7] significantly supports the generation of deep SAR-optical matching models for high-resolution data.

3. GANS FOR SAR-OPTICAL IMAGE MATCHING

While achieving promising results, the CNN based matching networks proposed in [3], [4] and [5] highlighted many challenges which still remain unsolved in the application of deep learning to SAR and optical image matching. The challenges originate from the large geometric and radiometric differences between imagery in the SAR and optical domains. These differences are directly responsible for the poor performance of existing, signal based matching techniques such as SIFT and are related to the overfitting and lack of robustness which has been reported in some of the CNN based approaches.

Generative adversarial networks (GANs) have shown great promise in translating imagery between modalities [8], as well in the generation of high resolution and realistic imagery [9]. Thus it was reasoned that the ability for GANs to learn and translate between complex data manifolds could be exploited to improve deep SAR-optical matching pipelines.

3.1. Image-to-Image Translation for Easier Similarity Prediction

Traditional matching approaches such as SIFT or BRISK have proven to yield accurate and reliable results in the case of single-sensor image matching, but commonly fail for SAR-optical matching. Therefore, an intermediate step was proposed in [10] in order to eliminate (to a certain degree) radiometric differences between the images to be registered, and hence enable the utilization of traditional matching approaches. This step was realized through a conditional generative adversarial network (cGAN), which enables the generation of artificial image patches with the texture of SAR reference image patches, while keeping the geometric properties of a given optical image patch. The cGAN architecture proposed in [8] in combination with a least-squares loss was utilized for the artificial SAR image generation process. The results validate that traditional matching techniques (SIFT or BRISK) greatly benefit, in terms of matching accuracy

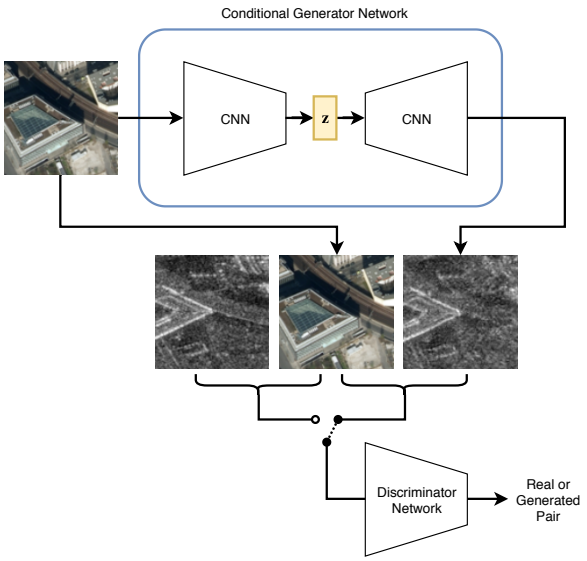


Fig. 2: Conditional GAN architecture: the generator network learns to translate images between domains, while the discriminator learns to distinguish generated and real image pairs.

and precision, from the use of the artificially generated SAR patches. However, the comparison against the siamese based approach from [3] also revealed the necessity for further developments.

3.2. GAN-based Hard Negative Mining

As mentioned in Section 2.2, hard negative mining is often employed during the training phase of deep matching networks in order to increase the discriminative ability of the network, and thus decrease the false positive rate (FPR). A low FPR is of crucial importance in image matching since the correspondence results are usually used as primary input to more complex data fusion tasks - such as stereogrammetric 3D reconstruction. Therefore, false matches would have a direct negative effect on the accuracy of data fusion products.

Traditional hard negative mining [11] requires a sufficiently large dataset, such that there are sufficient hard negative samples available during training iterations, even at low FPRs. Thus the data sparsity which is present in SAR-optical matching problems does not lend itself to the application of hard negative mining.

For this reason in [12] we proposed a framework for generating hard negative samples from existing data. This framework is based on a variational-GAN which generates a novel, yet similar, sample for every input SAR patch. This generated patch, with the corresponding original optical image is then added to the existing dataset as a negative training pair. By training the pseudo-siamese network of [4] with a combination of generated hard-negatives, and randomly assigned negatives, [12] managed to achieve a significant reduction in FPR without affecting the overall matching accuracy, and without

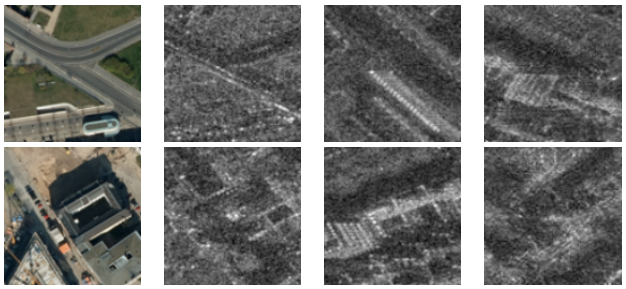


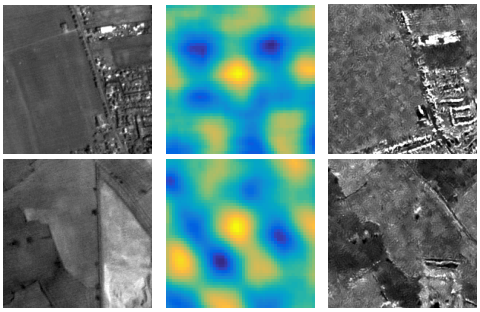
Fig. 3: A comparison between mined and generated hard negative samples. (a) original optical image, (b) corresponding SAR image, (c) hard negative sample mined from the dataset, and (d) generated hard negative.

the need for additional SAR-optical training pairs. Examples of the generated hard negative samples, as well as hard negative samples mined via the traditional approach, as used in [5], are depicted in Fig. 3.

4. RESULTS ACHIEVED SO FAR

In order to evaluate the relative success of each of the previously described approaches, tests were performed on real world data. These datasets vary in resolution, scene density (urban vs rural) and size and thus the experiments are not directly comparable, but do still provide insights into the various approaches. The results achieved so far can be summarized as follows:

- Applying the siamese architecture proposed in [3] to six TerraSAR-X (1.25m GSD) and PRISM (2.5m GSD) image pairs, an overall matching accuracy of 1.91 pixels with a precision of 1.14 pixels was achieved. The resulting tie points were used to improve the overall absolute geolocalization of the optical images. A visual representation of the matching performance can be seen in Fig. 4.
- The pseudo-siamese network described in [4] was able to achieve a matching accuracy of 83% at a FPR of 16% when applied to a patch-based feature matching scenario in an urban environment (0.5m GSD SAR and optical imagery).
- The addition of regularizing elements, such as hard negative mining and triplet loss, to the adapted HardNet architecture of [5], led to a binary patch-based matching accuracy of 93% at a FPR of 5% and a GCP matching precision of 4.4m when applied to semi-urban test scenes.
- Incorporating GAN-based hard negatives into the training of the pseudo-siamese matching network led to a significant reduction in the FPR, of 3% points, while obtaining an accuracy of 86% [12]. Constraining the FPR to 5% the network was able to achieve a matching accuracy of 81%.
- Finally, while SAR-optical image-to-image translation by cGANs only works reliably for semi-urban and rural areas, the results of [10] showed that this approach can be used



(a) (b) (c)

Fig. 4: A comparison between (a) optical patches, (b) the resulting score maps, and (c) despeckled SAR reference patches.

in combination with traditional, hand-crafted similarity descriptors, thus forgoing the need to train SAR-optical deep matching in an end-to-end manner.

5. OPEN RESEARCH DIRECTIONS

While the results achieved so far are promising, it is apparent that only SAR-optical deep matching for meter-resolution imagery of rural and semi-urban areas has reached a somewhat operational stage by now. If the resolution gets higher, i.e. down to the sub-meter domain, or if densely built-up urban areas shall be matched, the predictive power of the (pseudo-) siamese matching networks is not yet strong enough, while SAR-optical image-to-image translation doesn't work in a completely satisfying manner. This is caused by several reasons:

- As mentioned before, a crucial point certainly is the lack of sufficient training data for inner-city matching scenarios. While it is comparably easy to determine useful training data for rural scenarios, training data generation for urban scenes remains a highly challenging task. Future research will have to focus on the development of fully automatic engineering procedures to match corresponding SAR-optical image parts based on available 3D prior knowledge.
- Another direction of future research will be the adaption of unsupervised deep learning approaches that are able to learn powerful representations from large sets of non-corresponding data. Besides, approaches such as transfer learning (e.g. from easier-to-annotate low-resolution to high-resolution datasets) or multi-task learning (e.g. by combining semantic segmentation with image matching) seem to provide promising perspectives.
- While the lack of training data can – to some extent – be tackled by more sophisticated learning strategies, another important problem is the strong influence of different imaging geometries in both sensor modalities. Even if matching networks with strong predictive power can finally be trained, SAR-optical matching procedures will have to involve robust search strategies that take anisotropic geometric distortions and prior knowledge about the imaging parameters into account.

6. SUMMARY AND CONCLUSION

In this paper, we have summarized the last years of research on deep learning approaches for SAR-optical image matching within our TUM/DLR group. By comparing the thus-far achieved results, we were able to identify remaining challenges and future research directions. In conclusion, it can be said that deep learning will help to tackle the SAR-optical matching challenge in the future, but quite some more research and engineering efforts are necessary to achieve this ambitious goal.

Acknowledgements

This work was partially supported by the German Research Foundation (DFG) under grant SCHM 3322/1-1.

7. REFERENCES

[1] M. Schmitt, F. Tupin, and X. X. Zhu, "Fusion of SAR and optical remote sensing data - challenges and recent trends," in *Proc. IGARSS*, 2017, pp. 5458–5461.

[2] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. CVPR*, 2015, pp. 4353–4361.

[3] N. Merkle, L. Wenjie, S. Auer, R. Müller, and R. Urtasun, "Exploiting deep matching and SAR data for the geo-localization accuracy improvement of optical satellite images," *Remote Sens.*, vol. 9, no. 9, pp. 586–603.

[4] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, "Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 784–788.

[5] T. Bürgmann, W. Koppe, and M. Schmitt, "Matching of TerraSAR-X derived ground control points to optical image elements using deep learning," Submitted to ISPRS Journal of Photogrammetry and Remote Sensing.

[6] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, "Working hard to know your neighbor's margins: Local descriptor learning loss," in *Proc. NIPS*, 2017.

[7] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The SEN1-2 dataset for deep learning in SAR-optical data fusion," in *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, 2018, vol. 4-1, pp. 141–146.

[8] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017, pp. 5967–5976.

[9] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. ICLR*, 2018.

[10] N. Merkle, S. Auer, R. Müller, and P. Reinartz, "Exploring the potential of conditional adversarial networks for optical and SAR image matching," *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.*, vol. 11, no. 6, pp. 1811–1820.

[11] K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 39–51, 1998.

[12] L. H. Hughes, M. Schmitt, and X. X. Zhu, "Mining hard negative samples for SAR-optical image matching using generative adversarial networks," *Remote Sens.*, vol. 10, no. 10, art. no. 1552.