A SEMI-SUPERVISED APPROACH TO SAR-OPTICAL IMAGE MATCHING

L. H. Hughes, M. Schmitt

Signal Processing in Earth Observation, Technical University of Munich, Munich, Germany
- (lloyd.hughes, m.schmitt)@tum.de

KEY WORDS: Image Matching, Synthetic Aperture Radar (SAR), Optical Remote Sensing, Deep Learning, Deep Matching, Semi-supervised Learning

ABSTRACT:

Matching synthetic aperture radar (SAR) and optical remote sensing imagery is a key first step towards exploiting the complementary nature of these data in data fusion frameworks. While numerous signal-based approaches to matching have been proposed, they often fail to perform well in multi-sensor situations. In recent years deep learning has become the go-to approach for solving image matching in computer vision applications, and has also been adapted to the case of SAR-optical image matching. However, the hitherto proposed techniques still fail to match SAR and optical imagery in a generalizable manner. These limitations are largely due to the complexities in creating large-scale datasets of corresponding SAR and optical image patches. In this paper we frame the matching problem within semi-supervised learning, and use this as a proxy for investigating the effects of data scarcity on matching. In doing so we make an initial contribution towards the use of semi-supervised learning for matching SAR and optical imagery. We further gain insight into the non-complementary nature of commonly used supervised and unsupervised loss functions, as well as dataset size requirements for semi-supervised matching.

1. INTRODUCTION

The collection and exploitation of complementary information from multi-modal data sources enables a deeper understanding of the world and is critical in many applications across multiple domains. A key first step in any data fusion process is determining correspondences among these data sources in order to align and further exploit the complementary information in each modality (Schmitt and Zhu, 2016). In the case of image-based data fusion this relates to determining corresponding image regions across images which may have been acquired by different sensors, at different viewpoints and at various resolutions.

While the task of determining correspondences in conventional computer vision applications, such as structure from motion and pose estimation, has seen great progress and is solved to the degree of being usable operationally, it is still an open and relevant problem in the field of remote sensing. This is especially true when considering the case of determining correspondences in highly complementary, but vastly different image sources such as between synthetic aperture radar (SAR) and optical imagery (Schmitt et al., 2017).

As can be seen in Figure 1, the vastly different image acquisition schemes of SAR and optical sensors lead to imagery that not only depicts different properties of a scene, but also contains significantly different geometric distortions and imaging artifacts. Synthetic aperture radar imagery captures the physical characteristics of a scene, such as surface roughness or water content, while optical imagery provides details as to the chemical composition of the target area. Furthermore, SAR imagery suffers from imaging artifacts such as speckle, layover and radar shadow - none of which are present in optical imagery. These vast differences make determining correspondences between the data a challenging task.

Although many traditional feature matching methods have been proposed for matching SAR and optical data, e.g. (Ye and Shen, 2016, Ye et al., 2017, Dellinger et al., 2015), many of them still exhibit sub-optimal performance especially in high and very high resolution imagery. The advent and success of deep learning in developing robust solutions to the correspondence problem in traditional computer vision settings, e.g. by (Han et al., 2015, Zagoruyko and Komodakis, 2017), has lead to its application to multi-modal matching within remote sensing, e.g. in (Mou et al., 2017, Merkle et al., 2017a, Hughes et al., 2018b). Despite deep networks being universal function approximators, the results of their application to the SAR and optical matching problem have been mixed and with varying degrees of robustness and generalizability. These effects can be attributed to three main challenges:

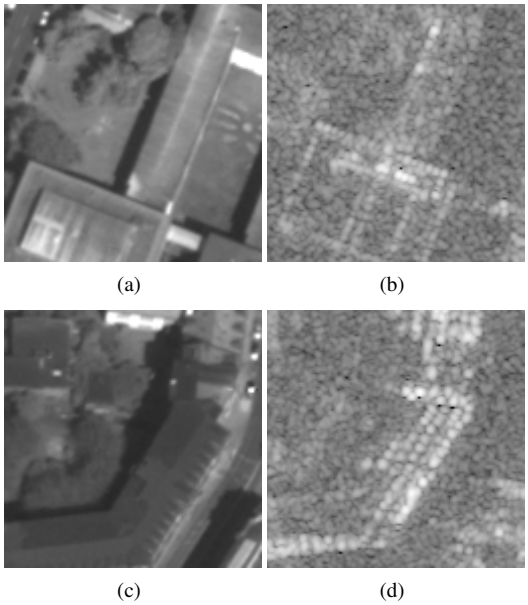


Figure 1: An example of corresponding SAR and optical patch pairs. Matching the image pairs in (**a**,**b**) and (**c**,**d**) proves to be a challenging task, even for domain experts.

firstly the intractability of creating large-scale annotated datasets due to SAR imagery being difficult, even for experts, to interpret; secondly the complex nature of SAR image formation which prevents the creation of realistic, synthetic datasets and finally the natural ineffectiveness of transfer learning techniques to extract meaningful feature representations from SAR, and lesser so, from optical space-borne imagery. These factors are all directly impacting on the feasibility of training the complex deep networks required to accurately determine correspondences between complex multi-modal data sources such as SAR and optical data.

To this end, we propose the use of semi-supervised learning to relax the requirements for large-scale labeled data in order to learn a well-generalizing SAR-optical image matching network. As semi-supervised learning has not yet been applied within this domain, the question still remains as to how much labeled data is required, and how well features learned in an unsupervised manner generalize to support supervised tasks. Additionally, we strive to understand the effects of data scarcity on the accuracy of learned SAR-optical descriptors, and the interplay between the unsupervised and supervised objectives. The main contributions of this paper can be summarized as follows: We formulate a semi-supervised approach to SAR-optical image matching and use this approach as a framework to assess the relative effect of data scarcity on the network's ability to learn meaningful descriptors for SAR-optical image matching.

2. RELATED WORK

2.1 Deep Learning for SAR-Optical Matching

Deep learning is becoming an increasingly important method in the toolbox of remote sensing practitioners, especially in the area of data fusion, and thus also SAR-optical matching (Zhu et al., 2017).

The first notable examples of this were provided in short succession by (Merkle et al., 2017b) and (Mou et al., 2017) who both proposed variants of a 2-stream architecture. (Merkle et al., 2017b) trained a siamese network to predict the relative shift between SAR and optical patches in order to improve the geolocalization accuracy of the optical data, while (Mou et al., 2017) trained a pseudo-siamese variant as a binary correspondence classifier. Taking inspiration from these seminal works, we extended the network proposed by (Mou et al., 2017) by enhancing the feature fusion stage and converting the output to a similarity score based on the soft-max probability (Hughes et al., 2018b).

Taking a different approach to the problem, (Merkle et al., 2018) proposed the use of a generative adversarial network (GAN) to generate SAR-like templates from optical image patches. These templates were then used as input to standard template matching approaches such as mutual information (MI) or normalized cross correlation (NCC).

These works all make use of supervised learning, which require large-scale labeled datasets – in this case, corresponding SAR-optical patch pairs. As such many of them lack robustness and generalizability, due to the intractability of creating large datasets of pixel-wisely matched VHR imagery of urban scenes.

In an attempt improve on this, we proposed a novel hard-negative mining strategy which does not increase the requirements for training data in previous work (Hughes et al., 2018a). To do this, we trained a conditional GAN to generate SAR patches which could be used directly, along with a corresponding optical image, for hard-negative mining. However, this approach is computationally expensive and does not completely resolve the problems caused by the scarcity of labeled data in SAR-optical matching problems.

2.2 Semi-supervised Learning

Semi-supervised learning constitutes a set of techniques for exploiting large-scale unlabeled datasets in order to support the learning in environments where labeled data is scarce (Chapelle et al., 2009). While many such methods exist, they all are centered around the same basic principles. Namely, to exploit unlabeled data in an unsupervised, or self-supervised manner to learn generalizable features, and to use small amounts of labeled data to steer learning towards a specific task.

(Zhang et al., 2016) and (Rasmus et al., 2015) proposed combining supervised classification with an unsupervised autoencoder-based reconstruction loss for image recognition. (Lai et al., 2017) trained a deep network using an adversarial loss to predict the flow field between a pair of images. This method used sparse depth information from LiDAR for supervision, while using an image consistency loss for unsupervised training. (Mukherjee et al., 2017) proposed the use of deep matching autoencoders to learn a common latent space between multi-modal data. This was achieved using a statistical dependency measure to pair unlabeled data during training and supervised with corresponding training pairs. Using a multi-phase training approach (Bui et al., 2018) pretrained a classifier for each domain in a supervised manner and then used a second training phase to learn a transformation between the learned embeddings for cross-domain image retrieval.

Autoencoders and reconstruction losses form a fundamental part of many semi-supervised learning approaches. However, they are still most often used as an auxiliary loss in supervised learning for matching multi-modal data (Ngiam et al., 2011, Liu et al., 2018). This is largely due to increased complexity of semi-supervised learning and the fact that these techniques lend themselves best to well conditioned problems (Cholaquidis et al., 2018). While the image matching problem is known to be ill-conditioned, autoencoders have still shown success in the domain of supervised learning for multi-modal matching. Thus in this paper, we will propose extensions to supervised autoencoder based matching techniques to allow for semi-supervised learning in within this domain.

3. SEMI-SUPERVISED SAR OPTICAL MATCHING

In this section, we describe our proposed SAR-optical matching network, including the use of autoencoders for semi-supervised learning of descriptors from labeled and unlabeled data, and the use of an adversarial loss for aligning these descriptor latent spaces. Further, we describe the training procedure and how matching can be achieved using the final trained network. An overview of the proposed architecture can be seen in Figure 2.

3.1 Network Architecture

In a similar vein to the matching networks proposed by (Liu et al., 2018) and (Mukherjee et al., 2017), we propose a dual autoencoder network in order to learn SAR and optical descriptors which can later be matched in a computationally efficient manner. In doing so we are able to exploit the self-supervised nature

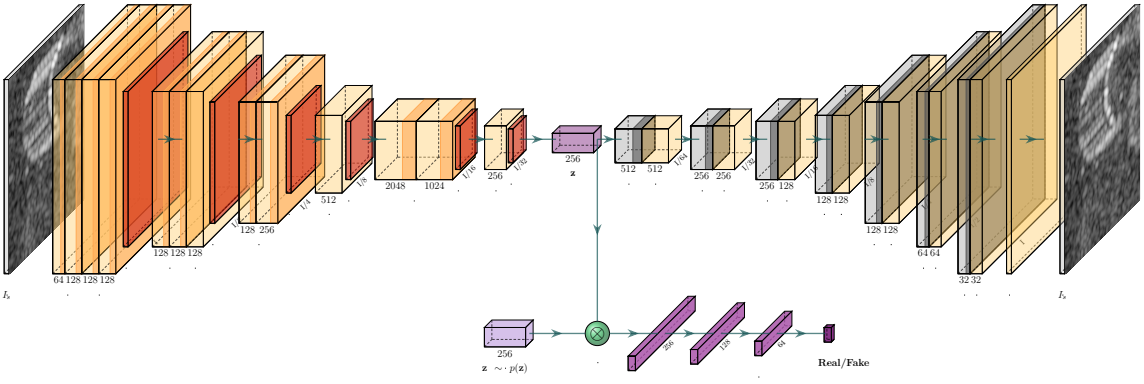


Figure 2: A single branch of the proposed network architecture. The autoencoder learns a meaningful latent code space $\mathbf{z}$ by learning to reconstruct the input image, while the discriminator network conditions the distribution of the latent codes using adversarial training and an arbitrary prior distribution. The optical branch is an exact mirror of the SAR branch and the discriminator network is shared between the branches.

of autoencoders to learn useful features from unpaired SAR and optical imagery. Furthermore, we use the latent code generated in the bottleneck as a natural descriptor and jointly train each domain specific autoencoder to align these latent codes. This alignment is achieved through the incorporation of a supervised loss function which is optimized using a small dataset of corresponding SAR-optical patch pairs.

Autoencoders typically consist of two networks, namely, an encoder and a decoder. Our proposed encoder network is based on the VGG11 (Simonyan and Zisserman, 2015) architecture. This architecture was chosen as a base due to its relative simplicity and low number of parameters. Furthermore, it has been used as a base to achieved state-of-the-art results in a variety of tasks (Iglovikov and Shvets, 2018), and is thus considered to be a good starting point for the exploration of semi-supervised learning for SAR-optical matching. The decoder network is based on a combination of convolutional and transposed convolution layers which are used to upsample the latent code in order to reconstruct the original image. The autoencoders for each modality (i.e. SAR and optical) have identical architectures and do not share any layers or weights. This allows for the learning of modality-specific features. As shown in Figure 2, the encoder network consists of blocks of 3×3 convolutions, batch normalization and activation with a Leaky ReLU function with a negative slope of 0.2. Similarly, the decoder network is made up of blocks of 3×3 transposed convolutions with a stride of 2 and ReLU activation, followed by a 3×3 convolutional layer and a ReLU activation. The depth of the feature maps are detailed in Figure 2.

For a given a SAR-optical image pair I_s, I_o we train the encoders $\text{Enc}_s, \text{Enc}_o$ to generate a descriptive latent code, $\mathbf{z}_s$ or $\mathbf{z}_o$ respectively, such that the decoder networks, $\text{Dec}_s, \text{Dec}_o$, can create an approximate reconstruction of the original inputs from the latent code. For a non-corresponding SAR-optical patch pair we seek to minimize the reconstruction loss such that,

$$\mathcal{L}_{recon} = \|I_s - \tilde{I}_s\|_2 + \|I_o - \tilde{I}_o\|_2, \quad (1)$$

where $\tilde{I}_s$ and $\tilde{I}_o$ are the reconstructed images generated by

$$\mathbf{z} \sim \text{Enc}(I), \quad (2)$$

$$\tilde{I} \sim \text{Dec}(\mathbf{z}) \quad (3)$$

using the appropriate, domain specific encoder and decoder networks.

For a pair of images labeled as either corresponding or non-corresponding, we augment the reconstruction loss, $\mathcal{L}_{recon}$, with a contrastive matching loss,

$$\mathcal{L}_{match} = y(\|\mathbf{z}_o - \mathbf{z}_s\|_2^2) + (1-y)\{\max\left(0, m - \|\mathbf{z}_o - \mathbf{z}_s\|_2^2\right)\}, \quad (4)$$

where y is the target label (zero for non-corresponding and one for corresponding), and m is the margin. The contrastive loss encourages the network in learning a latent space where corresponding pairs are near to each other, while non-corresponding pairs have a squared norm distance of at least margin m (Chopra et al., 2005). To ease the tuning of the margin hyperparameter, we took the L_2 norm of the each of the descriptor vectors $\mathbf{z}_o$ and $\mathbf{z}_s$ prior to the calculation of the contrastive loss. This ensures that both descriptors are on the hypersphere before matching and allows the use of normalized measures such as the cosine distance for matching the descriptors. This is significantly more efficient than descriptor-specific matching networks as the descriptors can be precomputed for each image patch.

In the end, the semi-supervised matching network is trained by minimizing the respective reconstruction losses $\mathcal{L}_{recon}$ for all SAR and optical data (paired and unpaired), while additionally minimizing the matching loss $\mathcal{L}_{match}$ for labeled, i.e. paired, data:

$$\mathcal{L}_{semisuper} = \sum_{i \in D_a} \left[\mathcal{L}_{recon}\left(I_s^i, \tilde{I}_s^i\right) + \mathcal{L}_{recon}\left(I_o^i, \tilde{I}_o^i\right)\right] + \sum_{j \in D_l} \mathcal{L}_{match}\left(\text{Enc}(I_s^j), \text{Enc}(I_o^j)\right), \quad (5)$$

where D_a and D_l represent the datasets of all, and labeled (corresponding and non-corresponding) SAR-optical patch pairs, respectively. Optimizing both the modality-specific reconstruction loss as well as the joint matching loss enables the network to learn to extract important features and generate descriptive latent codes from unlabeled data, while learning to align these latent spaces using a smaller labeled dataset.

While autoencoders are capable of learning complex data manifolds, these manifolds are often poorly conditioned with weak

supports. Thus they often do not extend well to unseen data, such as imagery with a slightly different data distribution or from a different spatial region. This is due to the fact that the manifold is only smooth near to existing samples, i.e. the training samples. To reduce these effects, and simplify the alignment between the modality specific latent distributions we propose to impose a continuous prior distribution $p(\mathbf{z})$ on the respective latent codes. This is realized through the reformulation of our modality specific autoencoders as adversarial autoencoders with a joint adversary, and is described in the following.

3.2 Adversarial Training

An adversarial autoencoder is an autoencoder which is regularized by matching the generated posterior $q(\mathbf{z})$ to an arbitrary prior $p(\mathbf{z})$. This is achieved through a min-max game in which the generator network, the encoder (Enc) of the autoencoder, learns to maximize the error of a discriminator network (Dis), while the discriminator learns to minimize the classification error of samples coming from the prior and the posterior (Makhzani et al., 2016). This objective function can be expressed as:

$$\min_{Enc} \max_{Dis} \underset{\mathbf{z}\sim p(\mathbf{z})}{E}[\log(\text{Dis}(\mathbf{z}))] + \underset{I\sim D_a}{E}[\log(1-\text{Dis}(\text{Enc}(I)))]. \tag{6}$$

In order to prevent the discriminator being able to learn the prior and posterior distributions too easily, the discriminator network is kept relatively shallow and simplistic. In our case, the discriminator is comprised of three fully connected layers of decreasing size, each of which is followed by a Leaky ReLU activation with a negative gradient of 0.2. The last layer of the discriminator uses a sigmoid activation to classify the input vector as either coming from the prior or posterior distribution. This network structure can be seen in Figure 2.

As we wish for the SAR and optical latent spaces to be aligned, such that corresponding pairs appear nearby in the code space, we impose the same prior on both latent distributions and solve the min-max problem over both encoders and the shared discriminator. This is done by alternating between updating the discriminator network and the generator (encoder) network using samples from the full dataset of labeled and unlabeled SAR-optical pairs.

Due to instabilities which can arise during the optimization of the min-max game (Equation 6), we replace the traditional generative adversarial loss with a Wasserstein-distance-based loss (Gulrajani et al., 2017). The Wasserstein loss strives to optimize the min-max game in terms of distributions rather than directly as a classification problem, and is thus more robust against gradient explosion and problems of mode collapse. Thus our final semi-supervised matching network is trained by minimizing the discriminator and autoencoder objective functions,

$$\mathcal{L}_{dis} = \sum_{i\in D_a} \left(\text{Dis}(\text{Enc}(I_s^i)) + \text{Dis}(\text{Enc}(I_o^i)) - 2\left(\text{Dis}(\mathbf{z}_p^i)\right)\right), \tag{7}$$

$$\mathcal{L}_{ae} = \mathcal{L}_{hnet} - \sum_{i\in D_a} \left(\text{Dis}(\text{Enc}(I_s^i)) + \text{Dis}(\text{Enc}(I_o^i))\right), \tag{8}$$

where $\mathbf{z_p^i}$ is a sample from an arbitrary prior distribution $p(\mathbf{z})$. In our case, we define $p(\mathbf{z})$ as a normal distribution such that $p(\mathbf{z}) \sim \mathcal{N}(0, 5)$.

3.3 Implementation Details

We implement the proposed approach using the PyTorch deep learning framework (Paszke et al., 2017). The optimization of the autoencoders is performed using the Adam solver with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ and a weight decay of 10^{-4}. The discriminator network is optimized using stochastic gradient descent (SGD) with a momentum of 0.9, weight decay equal to $3 \cdot 10^{-4}$ and a learning rate of $4 \cdot 10^{-3}$.

The learning rate for the Adam optimizer was determined using the search method proposed by (Smith and Topin, 2017), whereby the learning rate is rapidly increased from a small value, 10^{-7}, over consecutive batches while the loss is recorded. The learning rate is then selected to be in the region where the loss decreased in a smooth and constant manner (region of highest gradient). Using this approach we found the optimum learning rate for the Adam optimizer to be in the range of $5 \cdot 10^{-5}$ and $5 \cdot 10^{-4}$. This learning rate range was then used to initialize a one-cycle policy learning rate scheduler to dynamically vary the learning rate during training (Smith and Topin, 2017). The full network was then trained in an end-to-end manner for 100 epochs with a batch size of 32.

To improve the stability of the adversarial training the discriminator was trained using an update schedule with five times the frequency of that of the generator. Furthermore, the discriminator weights were clipped to the range of $[-0.1, 0.1]$ in order to preserve the *1-Lipschitz* constraints required for the Wasserstein loss (Petzka et al., 2017, Gulrajani et al., 2017).

Data augmentation was used to improve generalization and prevent overfitting due to the relatively small supervised dataset which we used. The data augmentation scheme included 1) horizontal and vertical flipping with a probability of 0.5 for each corresponding image pair, 2) the addition of Gaussian white noise with a standard deviation of $\sigma = 0.02$ to the optical image, and 3) scaling of image intensities by a randomly selected factor of $[0.95, 1.05]$, with a probability of 0.2. In order to preserve the accuracy of the labeled dataset, the same flipping and scaling transformations were applied to each image in the image pair. For the unlabelled dataset, these transformations are applied independently to each image.

4. EXPERIMENTS

4.1 Experimental Setup

As large-scale SAR-optical correspondence datasets are difficult to produce for very high resolution imagery, especially in urban areas, we make use of the UrbanAtlas dataset and reduce the region of interest for matching to areas which are mainly comprised of rural and semi-urban areas. In doing so we can limit the geometric differences between the SAR and optical imagery, and thus can derive corresponding points using the geo-localization information. While this approach may contain inaccuracies, these are assumed to be small at the spatial resolution of the dataset.

The UrbanAtlas dataset is comprised of high resolution ($2.5m$ GSD) TerraSAR-X and PRISM imagery of 23 cities across Europe. In order to increase the probability of salient features being present in both images we applied a Harris corner detector to the optical domain and applied a non-maximal suppression filter with a spatial constraint to ensure a minimum distance of 128

Figure 3: The distribution of the cities which were used for training (yellow), testing (black) and validation (white).

pixels between feature points. These feature points were then used as the center point when cutting SAR and optical patches, of 256×256 pixels, from the scenes. For training we extracted 50,000 patch pairs from 12 cities; 10,000 patch pairs from 3 cities for validation and 10,000 patch pairs from 8 cities for testing. The distribution of the cities into training, testing and validation sets is depicted in Figure 3.

In order to optimize the supervised loss we require both positive and negative training pairs. In order to achieve this we utilized a center crop of 128×128 pixels as the positive training pair, and an off center random crop of 128×128 to form a non-corresponding negative pair. The motivation for cropping both the positive and negative pair from the same patch was that nearby regions are likely to be more similar, giving the negative pair a similar distribution to the positive pair. This is expected to provide harder negative examples than purely random patch selection.

During pre-processing, all image patches were scaled to the range $[0, 1]$ and then standardized to zero mean and one standard deviation using the normal distributions as calculated from the SAR and optical images of the training set, i.e. $\mathcal{N}_{SAR}(0.5, 0.2)$ and $\mathcal{N}_{Opt}(0.45, 0.15)$. All other hyper-parameters were kept fixed for each scenario, such that the only variable was the degree of supervision.

For prediction at test time, we make use of a sliding window search procedure with a fixed optical template patch and a 256×256 SAR image search region. Matching is performed by calculating a descriptor for the central optical patch, and comparing this to the descriptors generated from a 128×128 sliding window over the SAR image. Thus, we obtain a $256 \times 128 \times 128$ descriptor volume for the SAR search region. The final correspondence map is then computed by calculating the cosine similarity between the descriptor volume and the descriptor of the optical template patch.

4.2 Matching under Data Scarcity

In order to assess our proposed network's ability to learn robust and discriminative features under conditions of data scarcity, we train the network with varying degrees of supervision. This further allows us to assess the effects of data scarcity on training the network, as well as the dynamics between the supervised and unsupervised loss functions.

We split the training dataset into supervised and unsupervised subsets with ratios of 100%, 75%, 50%, 25% and 5% supervised data to unsupervised data. The supervised subset is then oversampled to ensure that the distribution remains balanced. The network is then trained using alternating batches of unsupervised and supervised data.

The results of matching for these various scenarios are depicted as histograms/density functions of the pixel distance between the detected matching point and the ground truth location, as seen in Figure 4.

From Figure 4 it is clear that there is a non-linear relationship between the level of supervision and the number of well matched pairs. This relationship is particularly evident when observing the 1-percentile for each of the scenarios. The overall shape of the distribution should be noted too as it provides important insights into the network's matching abilities.

Due to the complexities of matching SAR and optical imagery it is expected that matching efforts will only yield a few correspondences. Thus it is often easier to obtain an intuition for the performance of a matching algorithm through a qualitative investigation of the correspondence maps for successful and unsuccessful matches. To this end Figure 5 depicts a few such examples for test scenes of varying building density and difficulty.

In an ideal matching scenario we would expect the correspondence maps, as shown in Figure 5, to have a single point of correspondence (red pixel) at the center, with the values at other offsets being relatively low in comparison (blue). However, in reality it is much more common to see a Gaussian like spread around the point of correspondence, with the peak value indicating the correct shift for maximal correspondence. From Figure 5 we can clearly see these point spread functions which depict the point of correspondence.

5. DISCUSSION

5.1 Semi-Supervised Matching

The examples in Figure 5 were selected as a fair depiction of the range of results which were obtained. From these examples, and in a qualitative manner, it is clear that the network is able to achieve SAR-optical matching, specifically in rural and semi-urban areas, across many levels of supervision.

On the other hand, the number of accurately matched points remains low, as evident from Figure 4. However, a large majority of data fusion tasks (such as stereogrammetry or image registration) require only a few reliable matches, i.e. they rely on a low false positive rate instead of only a high true positive rate. In conjunction, a high number of false negatives does not negatively impact follow-on applications.

The low number of detected correspondences is related to the vast differences in geometry between SAR and optical imagery which leads to salient points in the optical domain not always being visible in the SAR domain. Thus, the matching of these specific points becomes intractable even in the case of a fully supervised approach – which by nature of having more examples to learn from – should perform better than a semi-supervised approach. This outcome is also depicted in results corresponding to the SAR scene in Figure 5c, whereby the sharp edges and corners of the

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-2-W7-71-2019 | © Authors 2019. CC BY 4.0 License.

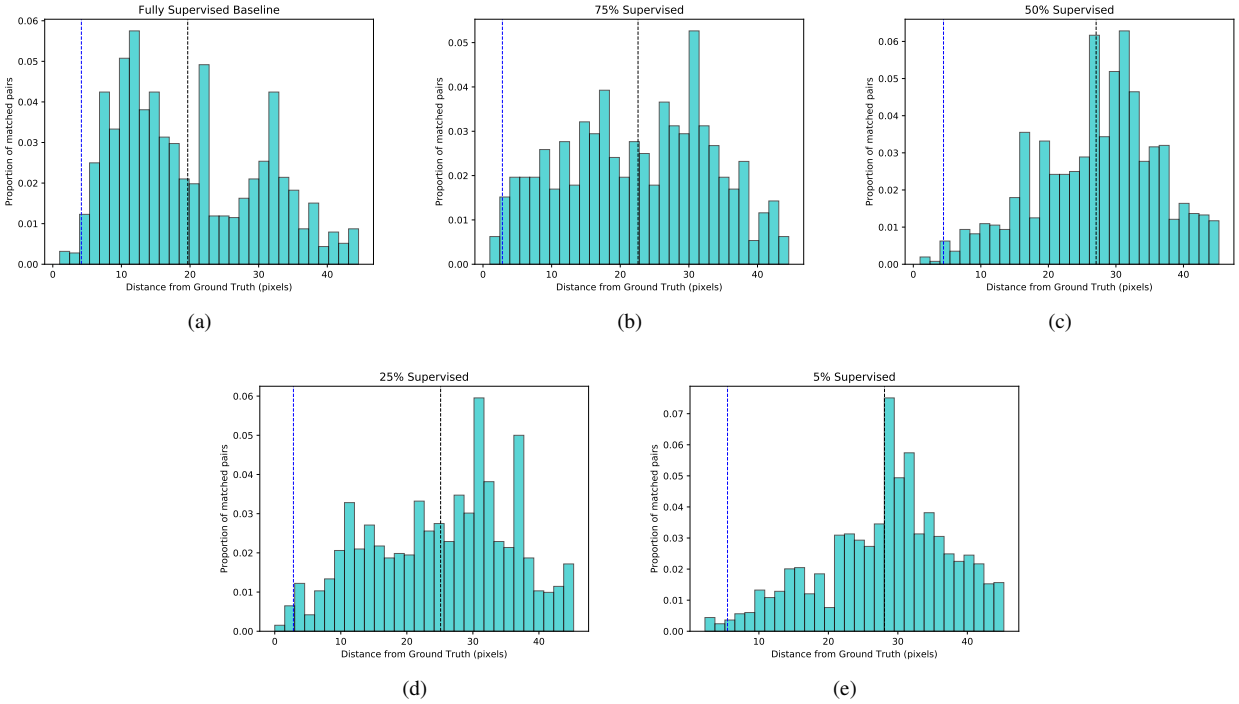


Figure 4: Histograms reflecting the precision of the determined matched point when compared to the ground truth location for varying degrees of supervision. The dashed black line represents the mean matching distance while the dashed blue line represents the 1-percentile matching distance.

building in the optical domain is not clearly visible in the SAR domain.

The relative consistency of these correspondence maps, across multiple levels of data scarcity, support the hypothesis that using a shared adversary and supervised objective function, we are able to align these latent spaces in a meaningful way for cross domain matching; even with very little data.

Furthermore, we note from Figure 5 that the spread of the correspondence peak appears to grow as we decrease the amount of supervision. This is providing insight into the increased uncertainty in the matching process as the latent distributions are only aligned at a small number of locations. Furthermore, and perhaps more importantly, we note that in the case of failed correspondences the correspondence map no longer represents a Gaussian like distribution and instead becomes multi-modal or somewhat random – as depicted in the results corresponding to Figure 5c. This observation could perhaps be exploited in future work to filter out failed correspondences, or to design more sophisticated correspondence point selection schemes; as selecting the point of correspondence based on a single value rather than based on the distribution of values is susceptible to noise.

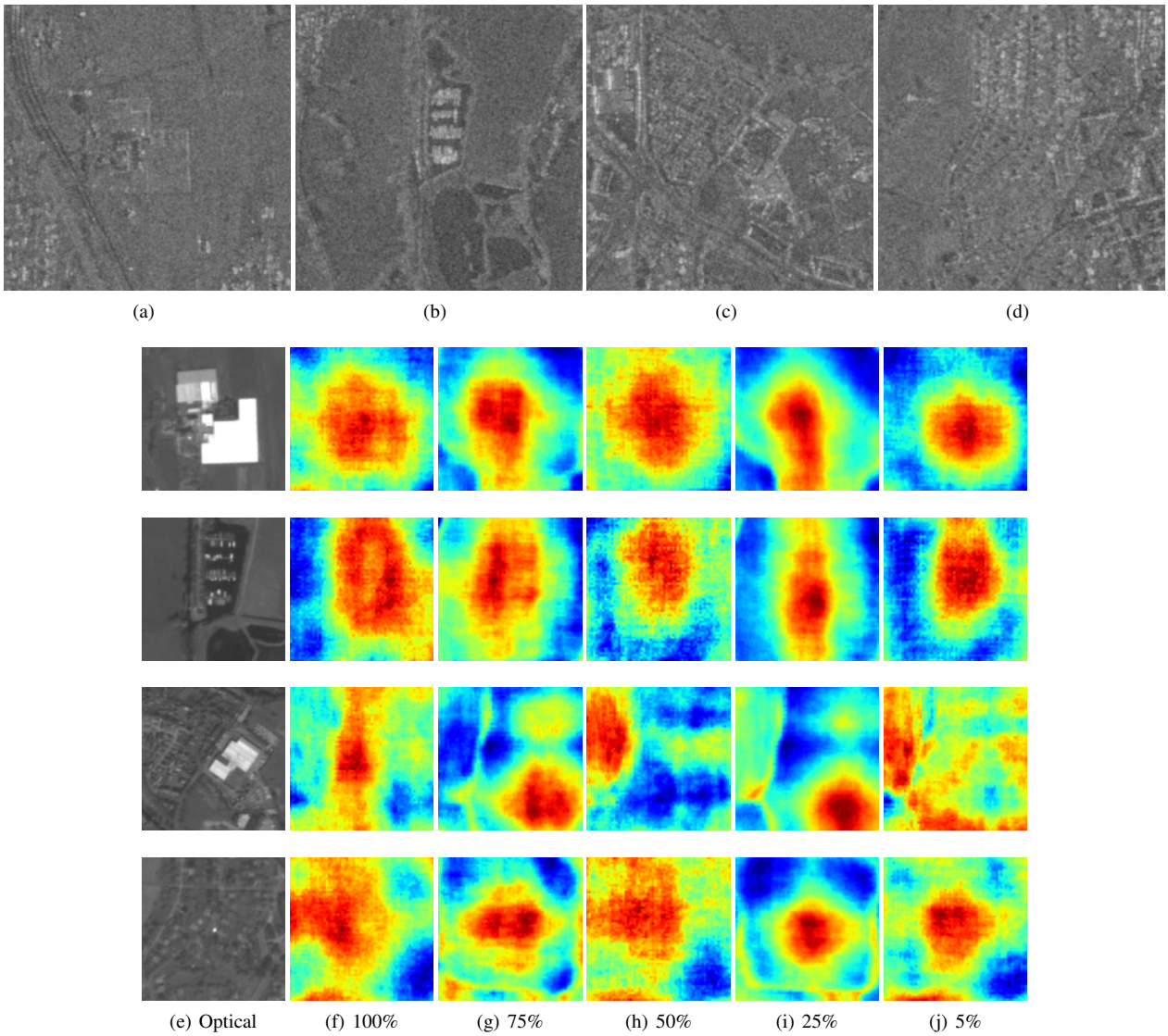
5.2 Effects of Data Scarcity

From the examples depicted in Figure 5 the impression arises that the proposed network performs best in semi-urban scenes (cf. Figure 5b), while it also shows reasonable performance in rural scenes (cf. Figure 5a). In urban scenes (Figure 5c-d), however, the matching accuracy varies significantly at different levels of supervision with the corresponding point shifting to a variety of locations. The reason for the better performance in semi-urban environments is likely due to the well distributed nature of objects in the scene, which allows the network to observe enough diversity in a patch that the descriptor can accurately capture the inherent details. In rural scenes, more often than not, there are fewer visual features and the scene has a relatively high self-similarity index, and thus the descriptors at multiple locations are similar. In urban scenes, the dense spacing of buildings, and thus the increased layover effects coupled with the $2.5m$ resolution obfuscate features and degrade the lower level structure of the scene, thus creating regions which have similar visual appearance, and in turn similar descriptors and multiple peaks in the correspondence map.

From Figure 4 the effects of data scarcity are visible in the overall distribution of the matching errors. As the amount of supervision is decreased the histogram becomes more skewed towards the right, and the number of successful matches for lower threshold values decreases significantly. This can be clearly observed when comparing the histograms of the fully supervised baseline (cf. Figure 4a) network to that of the scenario where only 5% supervision (cf. Figure 4d) was employed, where the former has a tighter distribution with a lower mean matching error, while the latter has a long tail and a very right-skewed distribution.

From further evaluation of Figure 4 it is clear that there isn't a linear trend between the number of accurately matched pairs and the amount of supervision used during training. This is evident in the accumulation of the number of matches which fall in the 1-percentile. Through this observation it is clear that 75% supervision and 25% supervision both have a higher number of low-error matches than the baseline approach.

At first glance this outcome can seem counter intuitive, however, an analysis of the literature (Dai et al., 2017) leads to the hypothesis that the unsupervised reconstruction loss and supervised matching loss are orthogonal to some degree. Thus, by optimizing for both losses in the baseline method the network ends up in a local minimum which is not necessarily best suited to either task. The reduction in the amount of supervision in the net-



(a) (b) (c) (d)

(e) Optical (f) 100% (g) 75% (h) 50% (i) 25% (j) 5%

Figure 5: Correspondence maps produced under varying conditions of data scarcity, on example scenes of differing density. (a-d) exemplary SAR test scenes, corresponding rows depicting (e) optical image patch, and (f - j) correspondence maps when trained with supervision percentage of 100%, 75%, 50%, 25% and 5% respectively.

work can be likened to applying some weighting function to the loss functions, and thus prioritizing the one objective over the other. In doing so the network is able to find a better optimum for the latent space generation task (reconstruction and adversarial losses) and the alignment of these spaces becomes an auxiliary task. While we would prefer to improve matching over reconstruction, it appears from the results that the prioritization of the adversarial task (by decreasing the supervision level) does in turn improve the matching task in some situations. This, however, would need to be subject to further investigation to fully understand the dynamics at play.

6. SUMMARY AND OUTLOOK

In this work, we proposed a semi-supervised approach to learn modality-specific features which are matchable via a simple distance-based metric, in our case cosine similarity. The approach consists of modality-specific autoencoders, which learn feature representations from unlabeled data, and are trained in an adversarial manner to enforce smoothness on the latent space. These learned representations (descriptors) are then aligned, using a supervised matching loss such that matching can be performed.

We further evaluated the effects of data scarcity on learning meaningful feature descriptors for SAR-optical matching by training our proposed network at varying levels of supervision and analysing the matching results in the form of correspondence maps, as well as the precision achieved for matching on our test set.

Overall we showed that even under very low data conditions, i.e. only 5% of supervision, we were able to obtain accurate correspondences in rural and semi-urban areas. While the overall number of accurate (1-percentile) correspondences was low, the strong structure of their correspondence maps leads us to believe that they could be filtered out during a post-processing step. This will be subject to further investigation in future work.

Furthermore, we found that the unsupervised and supervised objective functions are not fully complementary. That leads to a stunted baseline approach due to the strong trade-offs in the feature space required for each task. However, it was found that

decreasing the amount of supervision can be sufficient to enable the network to learn a better latent distribution, and thus achieve higher accuracy in matching. This paper provides an initial contribution to the use of semi-supervised learning to exploit unlabelled training data in order to support SAR-optical matching, where training data is usually scarce and difficult to obtain. In future work we will investigate post-processing methods for extracting high accuracy correspondences based on the structure of their correspondence maps. We will further investigate the hypothesis that lowering supervision signals is equivalent to applying a weighting between the loss functions, and then will investigate ways of automatically learning an inverse weighting to reprioritize the matching/alignment task objective over the unsupervised objectives.

ACKNOWLEDGEMENTS

This work is supported by the German Research Foundation (DFG) as grant SCHM 3322/1-1

REFERENCES

Bui, T., Ribeiro, L., Ponti, M., Collomosse, J., 2018. Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression. *Computers & Graphics*, 71, 77–87.

Chapelle, O., Scholkopf, B., Zien, A., 2009. Semi-supervised learning. *IEEE Trans. Neural Netw.*, 20(3), 542–542.

Cholaquidis, A., Fraimand, R. and Sued, M., 2018. Semi-supervised learning: When and why it works. *arXiv preprint arXiv:1805.09180*.

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. In: *Proc. CVPR*, 539–546.

Dai, Z., Yang, Z., Yang, F., Cohen, W.W., Salakhutdinov, R. R., 2017. Good semi-supervised learning that requires a bad GAN. In: *Proc. NeurIPS*, 6510–6520.

Dellinger, F., Delon, J., Gousseau, Y., Michel, J., Tupin, F., 2015. SAR-SIFT: a SIFT-like algorithm for SAR images. *IEEE Trans. Geosci. Remote Sens.*, 53(1), 453–466.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. C., 2017. Improved training of Wasserstein GANs. In: *Proc. NeurIPS*, Long Beach, USA, 5769–5779.

Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C., 2015. MatchNet: Unifying feature and metric learning for patch-based matching. In: *Proc. CVPR*, 3279–3286.

Hughes, L.H., Schmitt, M., Zhu, X.X., 2018a. Mining hard negative samples for SAR-optical image matching using generative adversarial networks. *Remote Sensing*.

Hughes, L.H., Schmitt, M., Mou, L., Wang, Y., Zhu, X.X., 2018b. Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN. *IEEE Geosci. Remote Sens. Lett.*, 15(5), 784–788.

Iglovikov, V., Shvets, A., 2018. Ternausnet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation. *CoRR*.

Lai, W.-S., Huang, J.-B., Yang, M.-H., 2017. Semi-supervised learning for optical flow with generative adversarial networks. In: *Proc. NeurIPS*, 354–364.

Liu, W., Shen, X., Wang, C., Zhang, Z., Wen, C., Li, J., 2018. H-net: Neural network for cross-domain image patch matching. In: *Proc. IJCAI*, 856–863.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., 2016. Adversarial autoencoders. In: *Proc. ICLR*.

Merkle, N., Auer, S., Müller, R., Reinartz, P., 2018. Exploring the potential of conditional adversarial networks for optical and SAR image matching. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 11(6), 1811–1820.

Merkle, N., Fischer, P., Auer, S., Müller, R., 2017a. On the possibility of conditional adversarial networks for multi-sensor image matching. In: *Proc. IGARSS*, Fort Worth, USA, 2633–2636.

Merkle, N., Luo, W., Auer, S., Müller, R., Urtasun, R., 2017b. Exploiting deep matching and SAR data for the geo-localization accuracy improvement of optical satellite images. *Remote Sensing*, 9(6), 586.

Mou, L., Schmitt, M., Wang, Y., Zhu, X., 2017. A CNN for the identification of corresponding patches in SAR and optical imagery of urban scenes. In: *Proc. JURSE*, Dubai, U.A.E.

Mukherjee, T., Yamada, M., Hospedales, T.M., 2017. Deep matching autoencoders. *CoRR*.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y., 2011. Multimodal deep learning. In: *Proc. ICML*, 689–696.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in PyTorch. In: *Proc. NeurIPS*.

Petzka, H., Fischer, A., Lukovnicov, D., 2017. On the regularization of Wasserstein GANs. *arXiv preprint arXiv:1709.08894*.

Rasmus, A., Valpola, H., Honkala, M., Berglund, M., Raiko, T., 2015. Semi-supervised learning with ladder network. *CoRR*.

Schmitt, M., Zhu, X.X., 2016. Data fusion and remote sensing – an ever-growing relationship. *IEEE Geosci. Remote Sens. Mag.*, 4, 6–23.

Schmitt, M., Tupin, F., Zhu, X., 2017. Fusion of SAR and optical remote sensing data - challenges and recent trends. In: *Proc. IGARSS*, Fort Worth, TX, USA, 5458–5461.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: *Proc. ICLR*.

Smith, L.N., Topin, N., 2017. Super-convergence: Very fast training of residual networks using large learning rates. *CoRR*.

Ye, Y., Shen, L., 2016. HOPC: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching. *ISPRS Annals*, 3, 9.

Ye, Y., Shan, J., Bruzzone, L., Shen, L., 2017. Robust registration of multimodal remote sensing images based on structural similarity. *IEEE Trans. Geosci. Remote Sens.*, 55(5), 2941–2958.

Zagoruyko, S., Komodakis, N., 2017. Deep compare: A study on using convolutional neural networks to compare image patches. *CVIU* 164, 38–55.

Zhang, Y., Lee, K., Lee, H., 2016. Augmenting supervised neural networks with unsupervised objectives for large-scale image classification. In: *Proc. ICML*, 612–621.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.*, 5(4), 8–36.