ISPRS Journal of Photogrammetry and Remote Sensing 169 (2020) 166–179

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

A deep learning framework for matching of SAR and optical imagery

Lloyd Haydn Hughes [a], Diego Marcos [b], Sylvain Lobry [b,c], Devis Tuia [b,d,*], Michael Schmitt [a,e,*]

[a] Signal Processing in Earth Observation, Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany
[b] Laboratory of Geo-Information Science and Remote Sensing, Wageningen University, The Netherlands
[c] Université de Paris, LIPADE EA 2517, Paris 75006, France
[d] Environmental Computational Science and Earth Observation Laboratory, EPFL, 1950 Sion, Switzerland
[e] Department of Geoinformatics, Munich University of Applied Sciences, Karlstr. 6, 80333 Munich, Germany

ARTICLE INFO

Keywords:
Multi-modal image matching
Image registration
Feature detection
Deep learning
Synthetic Aperture Radar (SAR)
Optical imagery

ABSTRACT

SAR and optical imagery provide highly complementary information about observed scenes. A combined use of these two modalities is thus desirable in many data fusion scenarios. However, any data fusion task requires measurements to be accurately aligned. While for both data sources images are usually provided in a georeferenced manner, the geo-localization of optical images is often inaccurate due to propagation of angular measurement errors. Many methods for the matching of homologous image regions exist for both SAR and optical imagery, however, these methods are unsuitable for SAR-optical image matching due to significant geometric and radiometric differences between the two modalities. In this paper, we present a three-step framework for sparse image matching of SAR and optical imagery, whereby each step is encoded by a deep neural network. We first predict regions in each image which are deemed most suitable for matching. A correspondence heatmap is then generated through a multi-scale, feature-space cross-correlation operator. Finally, outliers are removed by classifying the correspondence surface as a positive or negative match. Our experiments show that the proposed approach provides a substantial improvement over previous methods for SAR-optical image matching and can be used to register even large-scale scenes. This opens up the possibility of using both types of data jointly, for example for the improvement of the geo-localization of optical satellite imagery or multi-sensor stereogrammetry.

1. Introduction

Two of the most used modalities for space-borne remote sensing are Synthetic Aperture Radar (SAR) and optical imagery, since the information they provide about observed scenes is highly complementary. Thus SAR-optical data fusion has become a relevant area of research within the field of remote sensing (Schmitt et al., 2017).

As with any data fusion task, a fundamental first step is the alignment of the various data sources. In the case of image-based data fusion this alignment usually takes place through the process of image matching. More specifically this relates to the determination of corresponding points or regions across images which have different viewpoints, resolutions and may have been acquired by different sensors.

In classical computer vision, where problems are often restricted to a single modality or sensor, the task of image matching is largely considered to be solved to the degree of being usable. However, this is not true when dealing with highly heterogeneous datasets and multiple modalities such as in the case of SAR-optical image matching. Although remote sensing imagery often contains geographical coordinates for each pixel, we cannot rely on this geocoding to provide accurate correspondences as optical imagery often contains significant geolocalization errors (Merkle et al., 2017; Müller et al., 2012). Thus we need to rely on an image matching process which is subject to many complexities related to the large geometric and radiometric differences between the SAR and optical modalities (Schmitt et al., 2017; Hughes et al., 2019). For instance, the geometric distortions present in SAR imagery, such as layover, foreshortening and radar shadow, have no direct analogues in the optical domain. Optical imagery, on the other hand, suffers from illumination effects, related to clouds, object shadows, and the global scene illumination.

To tackle these challenges, researchers took inspiration from classical computer vision and developed a number of approaches for SAR-optical matching. Suri and Reinartz (2010) used mutual information to create a histogram-based method of registering SAR and optical imagery. Later a multitude of hand-crafted approaches were developed which were aimed at improving the performance of the scale-invariant feature transform (SIFT) detection and description algorithm (Lowe,

* Corresponding authors.
E-mail addresses: devis.tuia@wur.nl (D. Tuia), m.schmitt@tum.de (M. Schmitt).

https://doi.org/10.1016/j.isprsjprs.2020.09.012
Received 22 April 2020; Received in revised form 29 August 2020; Accepted 14 September 2020
Available online 23 September 2020

2004), by adapting the gradient operator and scale-space to be more suited to the properties of SAR imagery (Dellinger et al., 2015; Gong et al., 2014; Suri et al., 2010). These approaches were relatively successful in matching images in the SAR domain, however, they failed to match across modalities as the detected and described features were independent of those features detected in the optical domain Ma et al. (2017).

This is partially due to the vast radiometric differences between SAR and optical imagery. To address this, (Ye and Shen, 2016) proposed the histogram of oriented phase congruency (HOPC) descriptor whereby phase congruency was used as a proxy for gradient information. This ensures a commonality between features and descriptors in both modalities. Xiang et al. (2018) argued for the use of modality-specific gradient operators with a Harris scale-space to better handle the large radiometric differences while still allowing for repeatable features to be detected across modalities. Li et al. (2020) combined these previous approaches and the use of phase congruency to create Radiation-variation Insensitive Feature Transform (RIFT), which was shown to be less sensitive to rotational and radiometric differences across modalities while still providing repeatable features.

While feature-based methods are able to find correspondences between SAR and optical modalities, their success is limited to imagery which obeys specific geometric and radiometric constraints. As feature-based methods rely on a small region of support, they are not able to deal with large differences in the geometric structure of features. Furthermore, the radiometric differences between the modalities lead to different descriptions of features and thus further reduce matching performance. Therefore, feature-based methods are better suited to matching flat, semi-urban and rural environments where the vast differences in local feature appearance are less severe and the radiometric properties are more strongly correlated (Li et al., 2020; Xiang et al., 2018; Ye and Shen, 2016).

At a higher level, the constraints on geometric and radiometric differences are a consequence of the hand-crafted nature of the feature detectors and descriptors and thus also exist in single domain matching problems. For instance, in classical computer vision, many handcrafted approaches break down under large baselines or strong radiometric differences. For this reason, and with the advent of modern deep learning techniques, there has been a strong movement towards deep matching to solve the image matching problem directly from data (Kuppala et al., 2020).

Fischer et al. (2014) demonstrated that features extracted from the last layer of a Convolutional Neural Network (CNN), pretrained on ImageNet, can outperform the SIFT descriptor in image matching tasks. This lead to the development of several CNN-based descriptors, which learned similarity metrics directly from corresponding image patch pairs. Simo-Serra et al. (2015) proposed the use of a siamese network trained with pairs of corresponding and non-corresponding patches, and a Euclidean distance metric to learn a 128-dimensional descriptor for image matching. A similar approach was proposed in Zagoruyko and Komodakis (2015), however, an additional network was added to focus the matching around the center of the image patch pair. Building on these approaches Han et al. (2015) proposed MatchNet, which made use of a triplet loss and hard negative mining to better discriminate between corresponding and non-corresponding patch pairs. In Balntas et al. (2016a,?), a triplet-based approach was proposed, which used a simple shallow network and thus lead to a drastic improvement in computational and training performance without sacrificing accuracy. Taking a different approach, Yi et al. (2016) proposed a learned variant of SIFT, in which each component of the SIFT matching pipeline was implemented as an independent CNN trained using SIFT as the ground truth. More recently, Dusmanu et al. (2019) proposed an alternative formulation which jointly solves the problem of feature detection and description using a single network to output a pixel-wise map of feature likelihood and a joint descriptor. Similarly, Revaud et al. (2019) propose a single network which jointly learns feature detection and description, as well as a measure of the local feature discriminative power, which is used to identify repeatable features for matching.

Driven by these successes, remote sensing practitioners turned to deep learning to address the various shortfalls of handcrafted approaches for matching SAR and optical imagery (Hughes et al., 2019). To this end a number of approaches have been developed which specifically account for the multi-modal and inherently heterogeneous nature of the imagery. The first notable examples of deep SAR-optical matching made use of (pseudo-)siamese networks: Merkle et al. (2017) proposed a siamese network to directly predict the relative shift between a larger SAR search patch and a smaller optical template patch. Similarly, Mou et al. (2017) framed the matching as a binary classification problem and trained a pseudo-siamese network to predict the correspondence of the center pixel between SAR and optical patches. Taking inspiration from these initial works we extended the pseudo-siamese network proposed in Mou et al. (2017) to include a more robust fusion network and modified the binary classification problem to output a similarity index based on a soft-max activation (Hughes et al., 2018). Citak and Bilgin (2019) proposed the use of SAR and optical visual saliency maps as an attention mechanism in the feature extraction arms of a siamese matching network. Wang et al. (2018) use a self-learned deep neural network to directly learn the mapping between a source and reference image with the goal of applying this mapping remote sensing image registration. Bürgmann et al. (2019) proposed modifications to HardNet (Mishchuk et al., 2017) and applied it to matching SAR Ground Control Points (GCPs) in optical imagery. Hoffmann et al. (2019) trained a Fully Convolutional Network (FCN) to learn a similarity metric which was invariant to small affine transformations between SAR and optical patch pairs. Ma et al. (2019) proposed a two-step, coarse-to-fine registration method based on features extracted from fine-tuned VGG16 model (Simonyan and Zisserman, 2015).

Although we have seen significant progress in the matching of SAR and optical imagery, these approaches rely on the selection of good feature points for the extraction of matchable candidate search and template patches. Given the large differences between SAR and optical imagery it is often the case that salient features are not visible in both domains. Thus the selection of candidate patches in previous works has largely relied on features extracted from a single modality (Bürgmann et al., 2019; Merkle et al., 2017; Hughes and Schmitt, 2019) or assumed correspondence based on geo-localization (Citak and Bilgin, 2019; Hoffmann et al., 2019; Ma et al., 2019). For instance in Merkle et al. (2017) the locations of road intersections extracted from OpenStreetMap (OSM) data were used as features for extracting candidate regions for matching. While this showed reasonable results, OSM data is known to have varying accuracy and is not globally consistent (Vargas-Muñoz et al., 2019). Furthermore, the approach also required significant preprocessing and manual intervention. Bürgmann et al. (2019) made use of GCPs derived from a geodetic stereo SAR approach as features for the extraction template patches from the SAR image. The generation of these GCPs is computationally complex and requires multiple SAR acquisitions of the same scene with specific acquisition geometry. Furthermore, these GCPs are not generic features and often do not exist in rural areas.

Even in the best case scenario, where the proposed candidate patches meet all the requirements for increasing the likelihood of matching, outliers and incorrect matches will still exist. This is both due to the ambiguity and the complexity of the task of matching under extreme heterogeneity. The task of identifying and removing outliers in classical computer vision usually falls on statistical approaches such as the RANdom Sampling and Consensus (RANSAC) algorithm (Fischler and Bolles, 1981). These approaches, however, have not seen use in SAR-optical matching due to the complexity of modeling the feature transfer between domains in the presence of large geometric differences. Therefore, the removal of outliers in SAR-optical matching approaches has largely relied on filtering matches based purely on the

similarity score. Thus many of the previously mentioned approaches suffer from high false positive rates, which degrade the performance of downstream tasks.

In this paper we propose a fully-automated, multi-scale SAR-optical matching framework to address some of the shortfalls and constraints of previous approaches. This framework is comprised of three neural networks used in sequence: first is a *goodness network*, made of domain-specific sub-networks. This first network highlights regions with a high likelihood of containing salient features which are matchable across modalities. Second is a *multi-scale matching network*, architected around a feature space correlation function, which produces correspondence heatmaps for the matching of candidate patches. Finally, an *outlier reduction network* is used to directly estimate the quality of the matching result and allow for the removal of incorrect matching results. We evaluate the effectiveness of the individual sub-components, as well as the complete SAR-optical matching pipeline on a large and diverse dataset of high resolution SAR and optical imagery.

2. Multi-modal feature proposal and matching framework

In this section we detail the architecture and design of the three components which make up the proposed end-to-end SAR-optical matching framework. An overview of the framework and definition of these main components is depicted in Fig. 1.

2.1. Goodness network

The first stage of our framework aims at extracting the candidate patches which are used, by the correspondence network, for matching SAR and optical imagery. To extract these patches, we assess the *goodness* of regions for matching, *i.e.* the suitability of a region for matching.

This assessment is made using two identical yet independent (i.e. the architecture is the same but weights are not shared) domain-specific CNNs, each one producing a map indicating the likelihood of a region being matchable. With each network being trained on a single modality, but supervised by the matching loss generated by the correspondence network (see Section 3.2 for details), we expect the domain specific CNNs to learn which features are likely to be discernible in the other modality. These two maps are then merged into a *cross-modality scene goodness map*.

To cope with the geo-coding errors which exist in optical remote sensing imagery and the large differences in geometry between SAR and optical imagery, we generate the *goodness maps* at a reduced resolution. This allows for the coarse alignment of the SAR and optical goodness maps, and thus the extraction of jointly good regions, i.e. regions which have a high goodness in both domains. While this alone solves the correspondence problem, it only does so at a significantly reduced resolution and thus the identified regions are used to extract candidate patches for higher resolution matching with the correspondence network presented in the next section. Furthermore, in identifying regions for matching in this manner we reduce the overall number of candidate points. However, many downstream applications of the determined correspondences only require a few, well distributed and accurately matched feature points.

The domain-specific networks are based on the VGG11 architecture (Simonyan and Zisserman, 2015). This base was chosen due to its simplicity, relatively low number of parameters and proven performance in a variety of tasks (Ma et al., 2019; Iglovikov and Shvets, 2018; Hughes and Schmitt, 2019). The backbone architecture consists of four blocks of two 3×3 convolutional layers, with each convolutional layer being a sequence of convolution, activation by a rectified linear unit (ReLU) and batch normalization (BN). The first three convolutional blocks are downsampled by a factor of 2 using max-pooling. The head of the network consists of two convolutional layers with a stride of 2 (thus downsampling the spatial dimension of the tensor by a factor of 2), followed by fully connected layers implemented using a 1×1 convolutional block. Thus creating a network with an effective downsampling factor of 32, which is slightly larger than the maximum expected offset, between high resolution SAR and optical imagery, as reported by Merkle et al. (2017). Finally, an average pooling layer, with a kernel size $\mathcal{N}_p$ and stride $\mathcal{N}_k$ ensures a receptive field that accounts the maximum expected offset between the domains, as well as the size of the desired template patch. Thus the goodness network can identify regions of high goodness, with a size of $32\mathcal{N}_p \times 32\mathcal{N}_p$ pixels, where a maximum offset of $32\mathcal{N}_k$ pixels exists between the SAR and optical modalities. An overview of the modality specific network architecture is depicted in Fig. 2.

The SAR and optical domain-specific goodness networks are trained in an independent manner, using co-registered SAR-optical patch pairs, $\mathbf{I}_s, \mathbf{I}_o$, and a shared binary label, y_m, derived from the matching result of the image pair, as described in Section 3.2. During training, the domain-specific goodness networks are only trained on the patches from their respective modalities, while the binary label provides details as to how *good* the input patch was for matching when matched against the corresponding image of the other modality. A Binary Cross Entropy (BCE) loss function is used to supervise the learning process:

$$\mathcal{L}_g = -\frac{1}{N}\sum_{i}^{N} y_{mi}\log(\tilde{y}_{mi}) + (1 - y_{mi})\log(1 - \tilde{y}_{mi}), \tag{1}$$

where y_{mi} is a binary label indicating if the ith patch pair can be matched, $\tilde{y}_{mi}$ is the Sigmoid activated output of the domain-specific goodness network, and N is the total number of samples.

In order to identify common regions of high goodness, it is imperative to fuse the outputs of the domain-specific goodness networks. This is achieved by combining the trained domain-specific goodness networks with a simple fusion stage, in order to form the final goodness network architecture (Fig. 2). The fusion stage is responsible for combining the two domain-specific goodness maps to create the cross-modality scene goodness map $\mathbf{G}$, and boils down to a pixel-wise merge operator. The merge-operator can be any pixel-wise operator, however, in this paper we investigate the minimum and maximum operators which represent continuous domain proxies for the intersection and union of the domain-specific goodness maps. This fusion is followed by a non-local-maximum suppression (NMS) operation, as proposed in Dusmanu et al. (2019):

$$\hat{G}_{ij} = \frac{\exp(G_{ij})}{\sum_{kl \in \mathcal{N}_{ij}} \exp(G_{kl})}, \tag{2}$$

where G_{ij} is the value of $\mathbf{G}$ at pixel (i, j), and $\mathcal{N}_{ij}$ is the pixel neighbourhood centered on (i,j). The NMS operation is performed using non-overlapping 3×3 pixel neighborhoods. This neighborhood size was selected based on the fact that a 3×3 pixel neighborhood is the smallest neighborhood which allows for the suppression of redundant points (points which lead to template with a high degree of overlap), while preserving the overall number of identified points of high goodness. Examples of the domain-specific, and cross-modal scene goodness maps are depicted in Fig. 3.

Points of high goodness are identified by selecting all pixels which exceed a threshold of 0.5 in the cross-modality scene goodness map. The pixel co-ordinates of the points of high goodness are then transformed into pixel co-ordinates in the original image space by undoing the pooling and stride operations, such that,

$$\mathbf{c}_{i'j'} = 2^{3+\lfloor \frac{\mathcal{N}_k}{2} \rfloor}\mathcal{N}_p\mathbf{p}_{ij}, \tag{3}$$

where $\mathbf{c}_{i'j'}$ is the location of the point in the original image, $\mathbf{p}_{ij}$ is the point location in the joint scene goodness map and $(\mathcal{N}_p, \mathcal{N}_k)$ are the pooling and stride parameters of the goodness network. Finally, these transformed point locations are used as center points for extracting candidate search and template patches from the SAR and optical images, respectively.

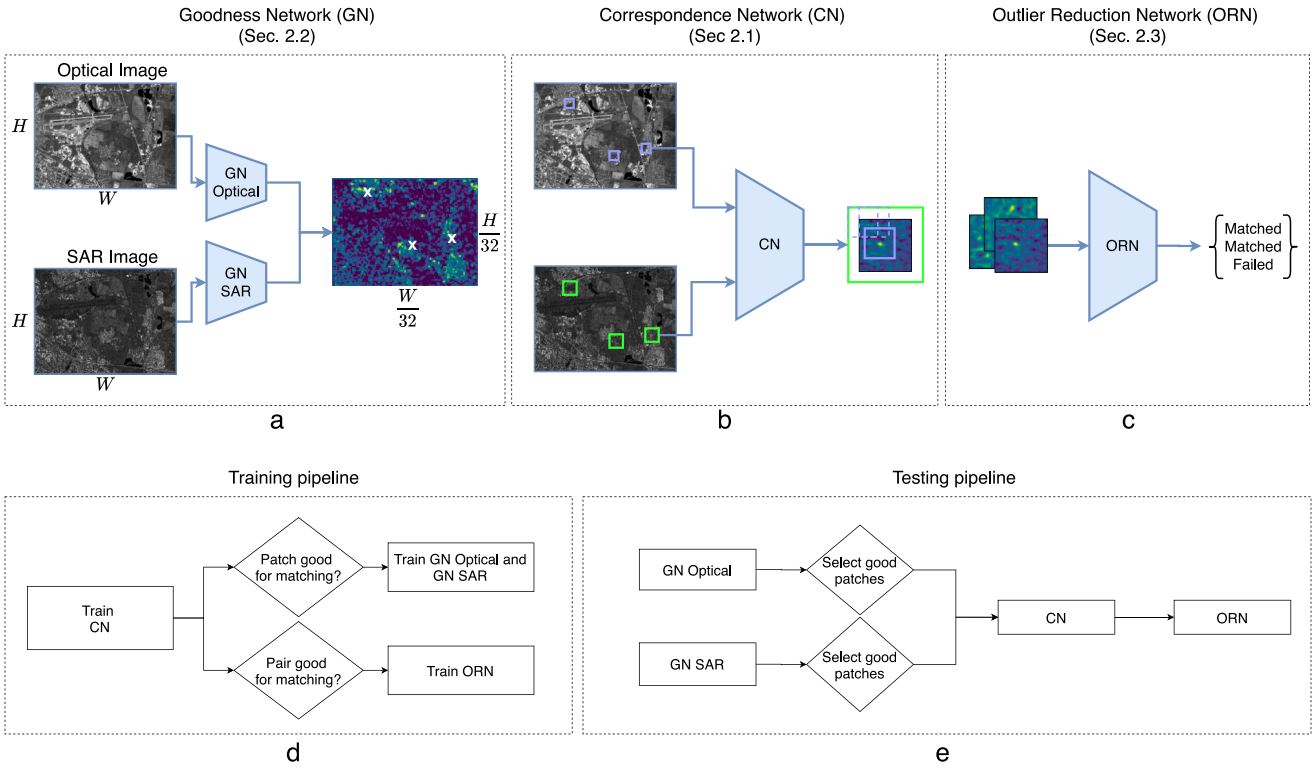

Fig. 1. The proposed SAR-optical matching framework. At test time (e), the SAR and optical images of the scene are first processed by the *goodness network* (a) to create a scene-wise map of suitability of regions for matching, from which points of high goodness are identified (white x's). Candidate search (green boxes) and template patches (blue boxes) are then extracted from these local maxima and the *correspondence network* (b) is used to determine the point of correspondence via feature-space cross correlation, thus producing a correspondence heatmap. The quality of the match is then assessed by the *outlier reduction network* (c) to filter out incorrect or ambiguous correspondences. During training (d), first the correspondence network is trained. This allows us to know which training patches resulted in good matches, providing us with the ground truth required to train the two goodness networks and the outlier reduction network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

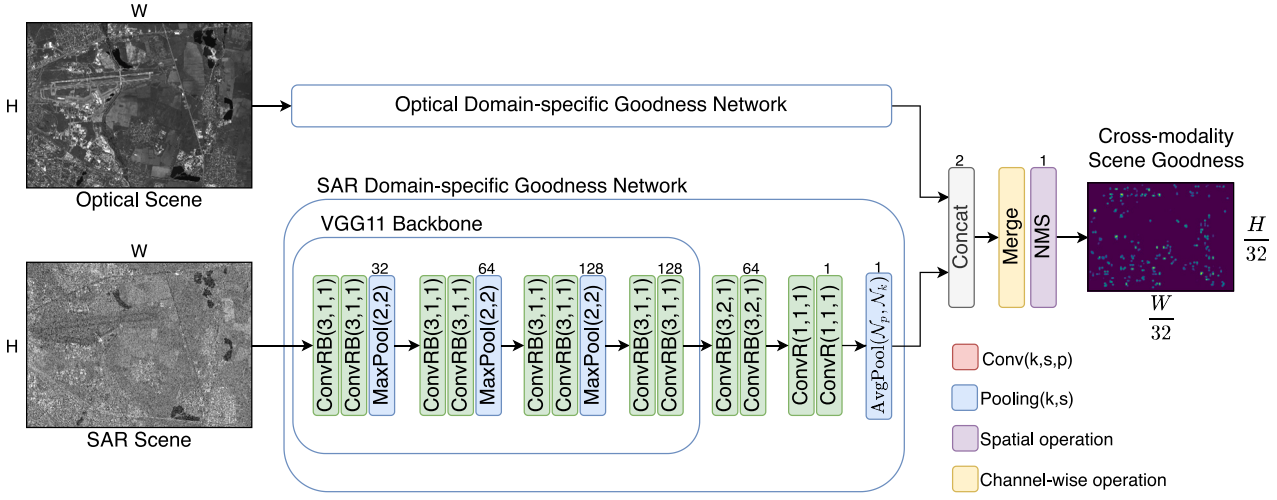

Fig. 2. Overview of the goodness network architecture showing the layer details for the SAR domain-specific branch. Conv(k, s, p) and (Max/Avg)Pool(k, s), represent a convolutional layer, and pooling layer, with a kernel of size k, stride of s, and padding of p, respectively. Convolution followed by ReLU is represented as ConvR(k, s, p), and the addition of batch normalization as ConvRB(k, s, p). The number of output channels at each layer is described by a number above the block and the optical stream has an identical structure to the SAR stream, although they do not share weights.

2.2. Correspondence network

The goodness network informs about regions of the two images that seem to be interesting to find matching keypoints, but does that only at a coarse resolution. The next step is to find a fine grained matching keypoint between the two. To do so, a second network, namely the correspondence network, slides a small subpatch of the optical image (*template patch*, $\mathbf{I}_t$, of size $\mathcal{N}_t \times \mathcal{N}_t$) over the wider SAR image (*search patch*, $\mathbf{I}_s$, of size $\mathcal{N}_s \times \mathcal{N}_s$) in search of a match. In other words, the correspondence network aims to determine the most likely point of correspondence for the center pixel of the template patch within the search region. SAR imagery was chosen for the search patches for two reasons: firstly, many regions exhibit low texture, which leads to a uniform response when comparing the template patch over these regions.

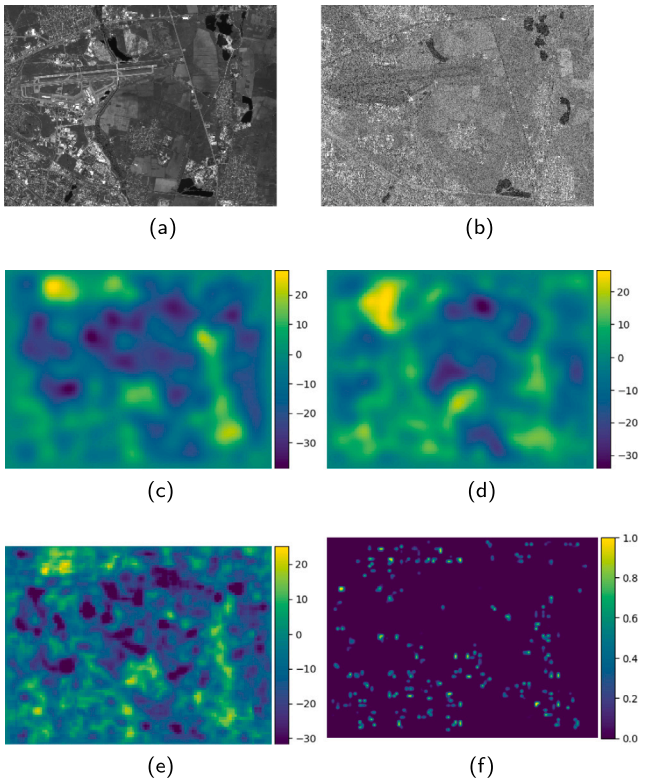

Fig. 3. Example of scene goodness maps produced by the domain specific networks, and the final, fused goodness map. (a) and (b) are the optical and SAR images of the scene. (c) and (d) are the respective domain-specific goodness maps, after average pooling with $\mathcal{N}_p = 4$ and $\mathcal{N}_k = 1$, (e) is the minimal response cross-modality goodness map $\mathbf{G}$ and (f) the final cross-modality scene goodness map $\hat{\mathbf{G}}$, where points of high goodness are clearly visible. It should be noted that images (c–f) have the same spatial extent as the input images (a–b) but have a resolution which is 32 times lower.

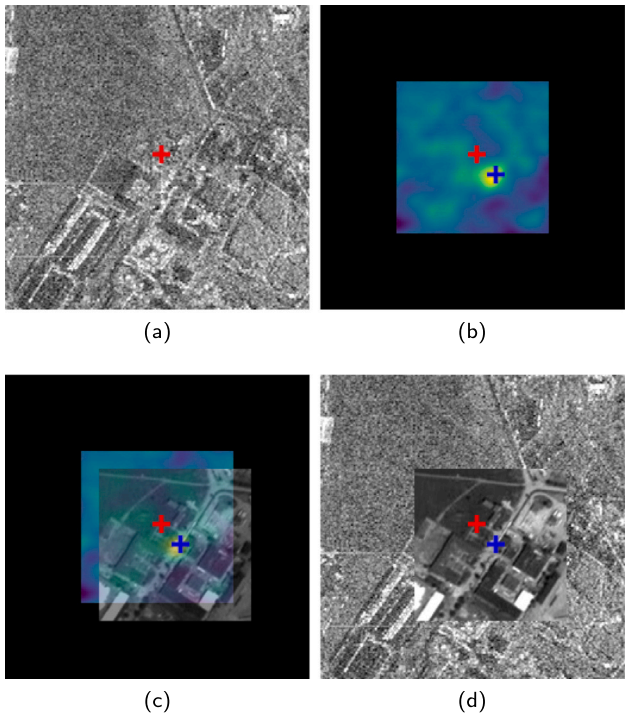

Fig. 4. Example of the process by which the correspondence heatmap can be used to determine the corresponding point for the center pixel of the optical template patch. (a) The search window with its center pixel marked by a red plus, (b) the resultant heatmap from the correspondence network with its center pixel aligned to that of the search window, and the peak point of correspondence depicted by a blue plus. (c) The center of the optical template patch is aligned to the peak point of correspondence, (d) the final alignment of the optical template patch, with the located point of correspondence marked by the blue plus. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

And secondly, SAR imagery is significantly better geo-referenced that optical imagery, which means that the identified correspondences are directly related to the correct geo-localization of the optical template. An example of the matching process using a candidate patch pair, and the output correspondence map is depicted in Fig. 4.

Existing approaches to SAR-optical matching largely rely on features extracted from the final layers of deep CNNs. While these features contain global semantic information they are low resolution and invariant to disturbances such as translation. Thus it can be argued that they lack the fine detailed features, required to accurately determine correspondence between images. For this reason we architected our correspondence network around the concept of convolutional hypercolumns (Hariharan et al., 2015) which are constructed by stacking feature maps extracted from multiple levels of a shallow CNN. These hypercolumns can thus be interpreted as a multi-scale feature pyramid which contains both feature primitives (i.e. lines, blobs, corners) as well as high-level contextual information.

The correspondence network consists of two identical, yet independent, four-layer CNNs, one for each modality, from which feature maps are extracted to form the modality specific hypercolumns. The number of channels in each hypercolumn is then reduced by a modality specific feature reduction network, before being matched using a feature space correlation operator.

The hypercolumn is constructed by extracting feature maps at each of the four layers of the feature extraction network. These feature maps are then upsampled, using bi-linear interpolation and stacked into a hypercolumn. The depth is then reduced to the desired number of features, $\mathcal{N}_d$, using a series of 1×1 convolutional layers. To improve response of salient features in each modality, a spatial attention map, as proposed in Woo et al. (2018), is created and applied to each hypercolumn. The reduced hypercolumn is then normalized along the channel dimension using $L2$ normalization.

The search and template hypercolumns are then matched in feature space using a correlation operation. As the search region is extracted from a larger image, and the area in which to search for a correspondence is known by design, the correlation operation is applied using *valid* padding. This refers to the fact that the correlation is only computed over the inner region of the search patch, to avoid the need for additional padding to deal with edge cases. Finally, the result is upsampled and padded to match the extent of the search window. The output of which is a heatmap containing the matching scores for each offset of the template window within the search window. The full architecture of the correspondence network, as well as the input and output datum, is depicted in Fig. 5.

We can train the network using a 2D Kronecker delta function as the ground truth, whereby the position of the unit impulse is parameterized as the true point of correspondence of the template patch within the search patch. The network is then trained via backpropagation using a modified mean-squared error (MSE) loss,

$$\mathcal{L}_{mse} = \frac{1}{\mathcal{N}_1 + \mathcal{N}_0} \sum_i \mathbf{w}_i \left(\mathbf{y}_i - f_{ss}(\tilde{\mathbf{y}}_i)\right)^2, \tag{4}$$

$$\mathbf{w}_i = \mathbf{y}_i \frac{\mathcal{N}_0}{\mathcal{N}_1} + (1 - \mathbf{y}_i), \tag{5}$$

where, $\mathbf{y}_i$ and $\tilde{\mathbf{y}}_i$ represent the target labels and the predicted heatmap of the ith sample. The function f_{ss} is a spatial softmax operation which is applied to the predicted heatmap in order to convert the matching scores into a probability distribution with the peak at the point of correspondence. The softmax activation relates all points in the heatmap

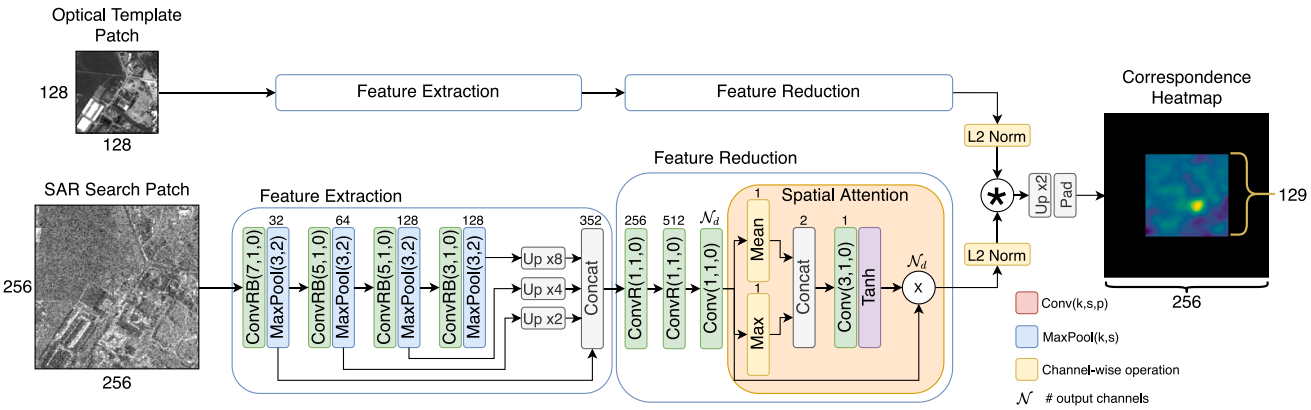

Fig. 5. The correspondence network architecture showing the layer details for the SAR branch with Conv(k, s, p) and MaxPool(k, s), representing a convolutional layer, and pooling layer, with a kernel of size k, stride of s, and padding of p, respectively. Convolution followed by ReLU is represented as ConvR(k, s, p), and the addition of batch normalization as ConvRB(k, s, p). The optical feature extraction and reduction networks follow an identical structure to those of the SAR branch, but the branches do not share any weights.

and thus to obtain a strong peak it encourages the suppression of the matching score in other regions. As the ground truth map contains only a single non-zero value we make use of a weighting vector, $\mathbf{w}_i$, to ensure the loss at the peak is given the same importance as the loss created by all non-corresponding points in the heatmap. This further exaggerating the requirement for a strong peak in the heatmap. Thus $\mathcal{N}_1$ and $\mathcal{N}_0$ represent the count of the number of zero- and non-zero pixels in $\mathbf{y}_i$.

Due to the spatial softmax operation f_{ss}, which normalizes $\sum_{x,y} \tilde{\mathbf{y}} = 1$, and the loss function which prioritizes peakiness, the network tends to overfit the training dataset. It achieves this by exploiting the peak-to-peak range of the pre-activated heatmaps, $\hat{\mathbf{y}}_i$. To reduce overfitting, encourage sparsity, and limit the dynamic range of $\hat{\mathbf{y}}_i$, we augment our $\mathcal{L}_{mse}$ loss with an L_1 regularization term. Thus the overall loss function can be expressed as

$$\mathcal{L}_{cor} = \mathcal{L}_{mse} + \lambda \sum_i |\hat{\mathbf{y}}_i|, \tag{6}$$

where λ is a hyperparameter to adjust the strength of the regularization.

2.3. Outlier reduction network

Due to the nature of the spatial softmax operation, which is applied to the correspondence heatmaps, $\tilde{\mathbf{y}}$ will likely always contain a small cluster of pixels which exhibits a strong response. However, the magnitude and location of these pixels is insufficient to discern the quality of the matching result. Thus we hypothesize that a better approach in determining the matching quality is to analyze the topology of the pre-softmax heatmap, $\hat{\mathbf{y}}$. We base this hypothesis on the observation that good matches tend to exhibit a single narrow peak, while incorrect matches are often multi-modal, or have a wide spread. Examples of various correspondence heatmaps are presented in Fig. 6.

To this end we train an Outlier Reduction Network (ORN) on $\hat{\mathbf{y}}$ to classify good and bad matches. The overall goal of the ORN is to reduce the overall number of inaccurate matches of the correspondence network, as a low false positive rate is more important than a high number of matches for many downstream applications of SAR-optical matching such as, for example, image coregistration or stereogrammetry (Müller et al., 2012; Merkle et al., 2017; Qiu et al., 2018; Bagheri et al., 2018).

The ORN is based on the same architecture as the correspondence feature extraction network, with some minor modifications. As the heatmaps produced by the correspondence network are not normalized and have a variable dynamic range, they cannot be assumed to have been drawn from the same distribution. Thus we adapt the input layer to use instance normalization (IN), instead of BN, as it operates on each sample independently. We formulate the problem of determining outliers as binary classification, and thus we need to adapt the head of the network to be suitable for this task. This modification includes the addition of an adaptive average pooling layer (AdaptAvgPool), which pools the entire spatial extent to output a single value. A Binary Cross Entropy (BCE) loss function is used to supervise the learning process:

$$\mathcal{L}_g = -\frac{1}{N} \sum_i^N y_{mi} \log(\tilde{y}_{oi}) + (1 - y_{mi}) \log(1 - \tilde{y}_{oi}), \tag{7}$$

where y_{mi} is a binary label indicating if the ith patch pair can be matched, $\tilde{y}_{oi}$ is the Sigmoid activated output of the ORN, and N is the total number of samples.

Training is then supervised using ground truth labels which are derived based on the accuracy of the matching result as reported by the correspondence network, this process is described in detail in Section 3.3. The problem can be summarized as: given a correspondence heatmap $\hat{\mathbf{y}}$, is it more likely to represent a successful or unsuccessful match. The full architecture is described in Fig. 7.

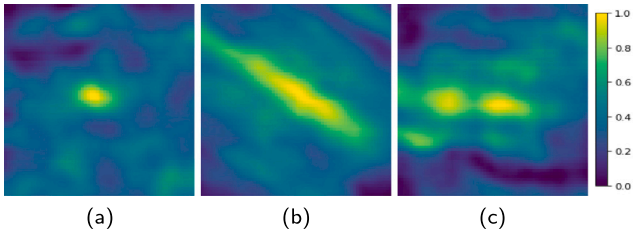

Fig. 6. Examples of common patterns seen in the correspondence heatmaps. For brevity only the *valid* region of the heatmap is depicted. (a) High likelihood of an accurate match as the heatmap contains only a single, strong response with a low spread. (b) A matching ambiguity exists along a single axis, which leads to a lower likelihood of the correct point of correspondence being identified. (c) A strongly multimodal response, with a wide spread which leads to multiple ambiguities in matching and thus a lower confidence.

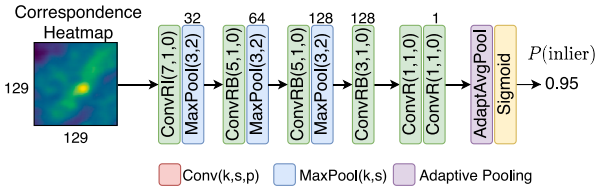

Fig. 7. An overview of the outlier reduction network (ORN) architecture. The network takes a correspondence heatmap as input and outputs the probability of the heatmap representing a successful correspondence.

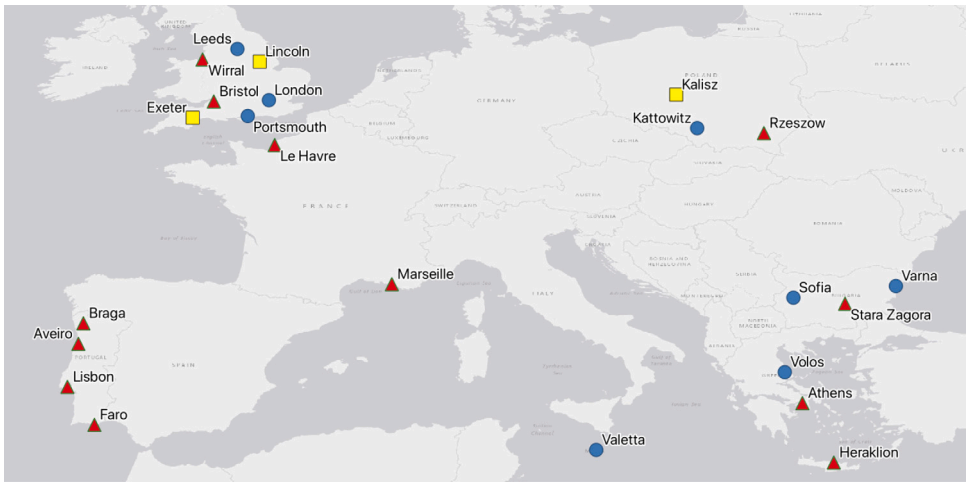

Fig. 8. The distribution of cities in the Urban Atlas dataset. The cities used for training, validation and testing are depicted as red triangles, yellow squares and blue circles respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Datasets and workflow

While the logical structure of the framework follows from the goodness network via the correspondence network to the outlier reduction network, as depicted in Fig. 1, this is not the case for training. As the training of goodness and outlier reduction networks rely on a trained correspondence network, we start by describing the dataset for the correspondence network followed by the description of the datasets, derived from the correspondence network outputs, which are used for training the goodness and outlier reduction networks. We further provide insights into the assumptions which were made and outline the way in which training, validation and testing samples were selected.

3.1. SAR and optical correspondence

To train the correspondence network we require a large dataset of salient candidate search and template patches with known points of correspondence. Due to the complexity of creating such a dataset, and the intractability of manually annotating correspondence across heterogeneous domains, we rely on simplifying assumptions (such as the correspondence of points at ground level in co-registered imagery) and the Urban Atlas dataset (Schneider et al., 2010) to generate our training and validation data.

The Urban Atlas dataset consists of manually co-registered, high resolution SAR and optical imagery acquired over 23 European cities. The imagery covers types of scenes including rural, industrial, suburban and urban environments, and has a total coverage area of approximately 20,000 km^2. The optical imagery was acquired using the Panchromatic Remote-sensing Instrument for Stereo Mapping (PRISM) sensor with a spatial resolution of 2.5 m, while the SAR imagery is based off of Enhanced Ellipsoid Corrected (EEC) TerraSAR-X data products and thus have square pixels with a pixel spacing of 1.25 m. Within the framework of the Urban Atlas project (Schneider et al., 2010), these SAR and optical images were accurately co-registered using hundreds of manually selected tie points to align each image pair. Based on this manual alignment it is reported that the residual co-registration error is within the range of 3 to 5 m (Merkle et al., 2017).

To reduce the complexity of the matching problem, as well as to allow larger batch sizes during training, we downsample the TerraSAR-X imagery using bi-linear interpolation, to a 2.5 m pixel spacing. The 23 cities are then divided into three groups for training, validation and testing. The process of dividing the cities between the three datasets is formulated as a discrete optimization problem over the total area of the imagery for each scene. Within this formulation the objective is to assign the cities to the three datasets such that the final dataset exhibits an approximately 80/20 split between the number of training, and testing samples. The final assignment and the spatial distribution of the cities can be seen in Fig. 8.

We then apply a Harris corner detector to the optical images to select points which are salient in at least one modality. Using these points, and the knowledge that the SAR and optical data in the Urban Atlas data set has been accurately co-registered (Schneider et al., 2010), we select the corresponding points from the SAR imagery using the geo-reference information for each pixel. We then use OpenStreetMap data and non-maximal suppression to reduce the overall point set to contain points which are more likely to be at ground level, such as near roads, and away from buildings and forested areas. This step is performed as in the case of co-registered data, the assumption of correspondence, at the same geo-location, only holds for points with no height above the ground.

We then cut 256 × 256 pixel patches from the SAR and optical imagery, centered around the identified points of correspondence. Then during training, we randomly crop a 128 × 128 template patch from the optical patch, with a maximum offset of 32 pixels around the center (accounting for the maximum shift (Merkle et al., 2017)). In doing so we ensure that the correspondence network learns to match the template image to the search window under realistic conditions, while allowing for the generation of ground truth data for the supervision and evaluation of the training process. An example of a candidate patch pair and the corresponding ground truth label is depicted in Fig. 9.

The 128 × 128 pixel extent of the optical template patch was chosen such that it captures sufficient spatial context to enable matching under the assumed worst case scenario, while remaining small enough to allow for better selectivity and finer grained matching. The extent of the SAR search patch was then selected such that is allowed for a maximum matching offset of up to 32 pixels (Merkle et al., 2017), while ensuring that even under extreme cases there is sufficient spatial context for matching.

We then standardize the dynamic range of the SAR imagery, and convert the speckle into an approximate additive Gaussian noise model. This is done by converting the pixel values to Decibels (dB), then clipping their range to the 3σ range of the training images, approximately $[10, 30]$dB, and finally scaling the resultant imagery such that $I_{SAR} \in [0, 1]$. For the optical imagery we simply normalize the values to the range of $I_{opt} \in [0, 1]$ by dividing through by 255.

While the test scenes are processed in the same manner as the training and validation scenes, the candidate patches are only useful for the evaluation of the correspondence network. Thus to evaluate the entire pipeline in an end-to-end manner we also create larger test scenes which can be used for evaluation. The train, test and validation

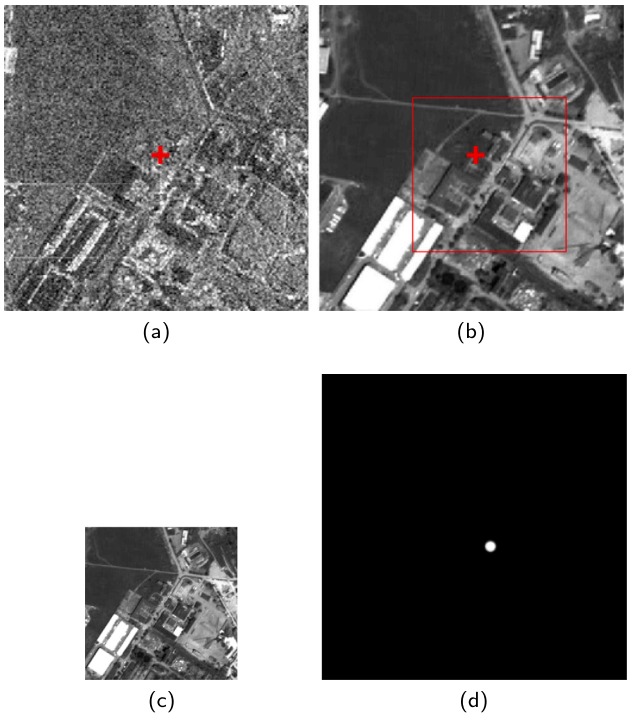

Fig. 9. A single training sample from our correspondence dataset. (a) The SAR search patch cropped around the location of the optical Harris corner (represented by the red cross), (b) the optical patch from which we extract the template search patch with random offset during training (depicted by the red box), (c) The extracted template patch, and (d) the derived ground truth label representing the true point of correspondence. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

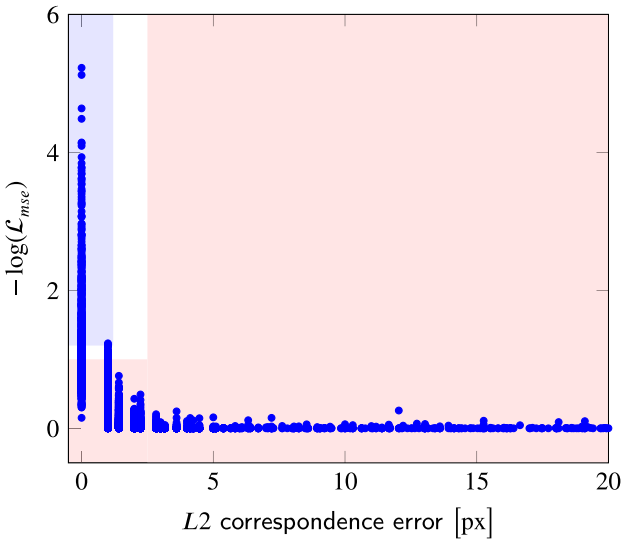

Fig. 10. A plot of the negative log matching loss versus the $L2$ pixel error for the training dataset. The region from which positive samples are drawn is highlighted in blue, and the negative samples are drawn from the area in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

patches are extracted from spatially distinct regions with a maximum patch overlap of 50%, while the 8 larger test scenes are created from each testing city and are thus spatially diverse and contain no overlap. The final dataset consists of 40,314 training candidate patch pairs, 4,205 validations pairs and 6,353 testing pairs, as well as, 8 larger test scenes.

3.2. Goodness

As no goodness dataset exists, and the creation of such a dataset is non-obvious for manual annotation, we rely on the trained correspondence network to identify patches which can act as positive and negative samples for training and evaluating the goodness network.

To do this we make use of the SAR and optical patches from the correspondence datasets, as well as the recorded matching loss, $\mathcal{L}_{mse}$, and an L_2 correspondence point error for each sample $\mathcal{L}_e$. We then create binary goodness labels for each sample by thresholding $-\log(\mathcal{L}_{mse})$ and the L_2 error. The negative log loss is used to invert the loss and reduce the dynamic range, which makes the task of selecting thresholds easier. We label the patch pairs such that,

$$y_{mi} = \begin{cases} 1 \text{ if } -\log(\mathcal{L}_{mse}) \geq 1.2 \text{ and } \mathcal{L}_e \leq 1 \\ 0 \text{ if } -\log(\mathcal{L}_{mse}) \leq 1 \text{ or } \mathcal{L}_e \geq 2.5 \end{cases} \tag{8}$$

where y_{mi} is the label for the ith patch pair, and label $y_{mi} = 1$ represents a patch pair which is good for matching, while 0 represents patch pairs which lead to inaccurate or unsuccessful matching. The values for the thresholds were chosen based on the training dataset such that we avoid possibly ambiguous samples, this process is depicted in Fig. 10. The negative log loss allows for easier selection of patches which produce correspondence heatmaps with desirable properties (low matching loss), such as a single peak with a narrow spread and small values everywhere else, while the $L2$ threshold ensures that these heatmaps actually correspond to positive matches.

As the correspondence dataset was created with no guarantees of mutually visible features, there is a large imbalance in the final goodness dataset with many more negative examples being present. To correct this we reduce the number of negative samples, by random selection, to be equal to the number of positive samples.

The final step in creating the goodness dataset is to crop the SAR search patches to the same extent as the corresponding optical template patch. This is done as the goodness score is derived only from the maximum point of correspondence, thus regions beyond the extent of the template patch do not contribute to whether the patch was good for matching or not.

3.3. Outlier reduction

To train the outlier removal network we make use of the *valid* region of the heatmaps generated from the correspondence network. These heatmaps are used as inputs to the outlier reduction network and the binary training labels indicate whether they were the result of a successful or unsuccessful matching result.

The generation of the heatmap labels follows the same approach to labeling as the previously described goodness dataset. However, we only apply the L_2 threshold as the label relies solely on whether the patch was accurately matched. Some labeled examples from the training dataset are shown in Fig. 11.

4. Implementation details

Due to the data requirements discussed in Section 3, we first train the correspondence network and then use the results of this training to generate the data needed to train the goodness and outlier reduction networks.

The average pooling parameters of the goodness network were set as $\mathcal{N}_p = 4$ and $\mathcal{N}_k = 1$. This corresponds to creating a receptive field of 128 × 128 pixels, which is large enough to account for co-registration errors of up to 160 m between domains, while exhibiting a 75% overlap between the evaluated regions. Furthermore, the hypercolumn depth

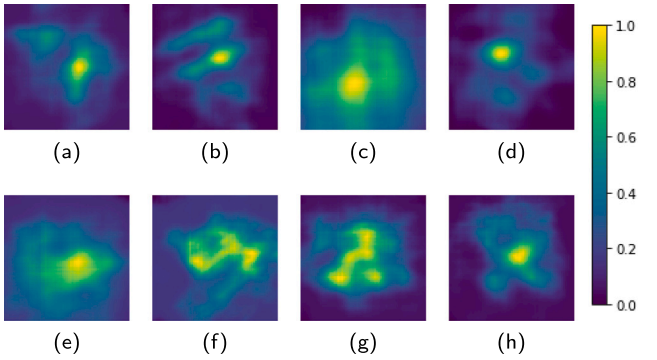

Fig. 11. Examples of the positive (a–d) and negative (e–h) correspondence heatmaps used to train the outlier reduction network. Only the *valid* region of the heatmap is used as the padded area contains no additional information.

Table 1
The hyperparameters used for training each of the sub-networks, where lr is the learning rate, β_1 and β_2 control the momentum.

Network	lr	β_1	β_2	Weight decay
SAR goodness	5×10^{-5}	0.9	0.999	0
OPT goodness	9×10^{-4}	0.9	0.999	0
Correspondence	1×10^{-4}	0.9	0.999	1×10^{-6}
Outlier reduction	1×10^{-4}	0.9	0.999	1×10^{-6}

Table 2
The probabilities used in the data augmentation pipeline while training each sub-network.

Network	HF	VF	IS	CS	CD
SAR goodness	0.5	0.5	0	0.7	0.8
OPT goodness	0.5	0.5	0	0.7	0.8
Correspondence	0.5	0.5	0	0	0
Outlier reduction	0.5	0.5	0.1	0	0.8

$\mathcal{N}_d$ of the correspondence network was set to 256. Similarly, the regularization weight λ, specified in the loss function for the correspondence network was set to 1×10^{-5}.

We make use of the PyTorch deep learning framework (Paszke et al., 2019) to implement all aspects of our proposed pipeline. The various sub-networks were randomly initialized using the method proposed by He et al. (2015), and are trained using the Adam solver (Kingma and Ba, 2014). The hyperparameters used for the solver are specified in Table 1. For each of the sub-networks the optimal learning rate was determined using the search method proposed by Smith (2017).

We make use of a fixed batch size of 16 samples, which constitutes the maximum batch size that could be used to train the correspondence network on a Nvidia GTX1080Ti GPU. This batch size further allowed for both goodness networks and the outlier reduction network to be trained simultaneously on the same GPU. The correspondence network was trained for 50 epochs, while the remaining sub-networks were trained for 1000 epochs due to the relatively small dataset size in comparison to the correspondence dataset size.

Data augmentation was used to improve generalization and reduce the risks of overfitting. This step was found to be of increased importance when training the goodness and outlier reduction networks due to the reduced dataset size. The data augmentation pipeline consisted of horizontal (HF) and vertical flipping (VF), image scaling (IS) by a factor of ± 0.1, intensity scaling (CS) by a random value between $(0.7, 1.3)$, and coarse image dropout (CD) of between $(1\%, 5\%)$ of the image pixels, taken on a version of the image which is downsampled to between $(80\%, 98\%)$ of the original size. Each of these augmentations is applied during training with a certain probability, as described in Table 2.

Table 3
Correspondence network configurations used in the ablation study. The use of a specific layer or inclusion of regularization is indicated by a yes (Y) or no (N).

Network	Attention	Spatial Softmax	L1 Reg.
CorrBase	N	N	N
CorrA	Y	N	N
CorrAS	Y	Y	N
CorrASL	Y	Y	Y

Table 4
Influence of attention (A), spatial-softmax activation (S) and $L1$ regularization (L) on the matching performance (evaluated on the validation dataset) of the correspondence network.

Network	Matching accuracy		Matching precision
	$\leq$ 1px [%]	avg. $L2$ [px]	mAP [px]
CorrBase	28.44	2.34	1.27
CorrA	28.13	2.36	**1.25**
CorrAS	44.42	3.0	1.99
CorrASL	**54.46**	**2.32**	1.53

To aid future development and in the interest of openness in science, a full implementation of the framework has been released.[1]

5. Experiments and results

In this section we first motivate our architectural choices by performing ablation studies. We further evaluate the performance of the individual sub-networks in comparison to existing methods, as well as their effects on the accuracy of the final set of correspondences. Finally, we evaluate the overall performance of the matching framework over a larger test scene.

5.1. Ablation study

To aid the design of the correspondence network described in Section 2.2 we performed an ablation study to compare the performance of the network as various architectural and regularization elements were added. We tested four variants of the correspondence network which are detailed in Table 3.

The networks were trained as previously described, and the random elements in the training process were made deterministic such that all networks were trained on the same data and augmentations. Finally, we evaluated the performance of the various networks using the validation dataset to prevent biasing our architecture selection to the test data.

We evaluate performance in terms of matching accuracy and precision. Whereby, matching accuracy is defined by the percentage of matches which have an $L2$ distance to the ground truth point of correspondence of at most one pixel, as well as the mean $L2$ error. Matching precision is defined as the mean average precision (mAP), where the standard deviation is used as a measure of precision. The results of the ablation study are described in Table 4.

From Table 4, it can be see that the addition of the spatial softmax operator leads to a significant improvement in terms of matching accuracy, however, this comes with a reduction in precision. The addition of the $L1$ regularization term further improves the matching accuracy while simultaneously only having slightly reduced precision over the baseline network with attention. Thus the *CorrASL* network was selected as the preferred architecture for our SAR-optical matching framework, and all further experiments are conducted with reference to this result.

[1] https://github.com/system123/SOMatch.

Table 5
A comparison of the matching accuracy and precision (evaluated on the testing dataset) of NCC (Burger and Burge, 2009), PSiam (Hughes et al., 2018) and our proposed correspondence network.

Network	Matching Accuracy		Matching Precision
	≤ 1px [%]	avg. $L2$ [px]	mAP [px]
NCC	8.2	7.85	6.81
PSiam	18.4	5.22	5.93
CorrASL	**46.9**	**2.1**	**2.62**

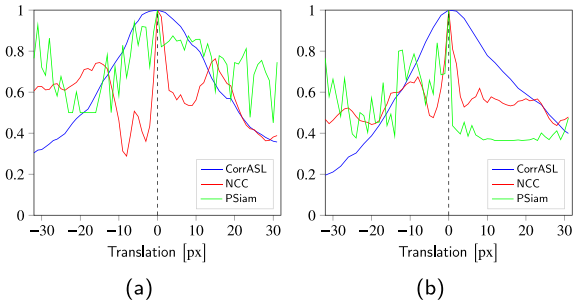

Fig. 12. The median correspondence heatmap peak shape along the (a) x-axis and (b) y-axis for each of the evaluated approaches.

5.2. Matching results

As the correspondence network plays a vital role in training the goodness and outlier reduction networks, it is imperative that we evaluate its performance relative to existing methods. To do so we make use of two relevant and available methods: Normalized Cross Correlation (NCC) (Burger and Burge, 2009), as well as the pseudo-Siamese matching approach (PSiam) presented in Hughes et al. (2018).

To ensure a fair comparison we retrained the pseudo-siamese approach on the same dataset, and under the same data augmentations and pre-processing as our correspondence network. As the pseudo-siamese network requires corresponding and non-corresponding SAR-optical patch pairs, for training, we applied random offsets for the creation of the non-corresponding pairs. Furthermore, both the SAR and optical pairs were cropped to an extent of 128 × 128 pixels. During the evaluation phase we apply the pseudo-siamese network over the full extent of the SAR search patch, using a sliding window approach, to generate a correspondence heatmap.

Table 5 shows the matching accuracy and precision for the baseline methods compared to the proposed method when assessed on our ground-level Harris corner derived test dataset.

From Table 5 it is clear that our proposed matching architecture provides a significant improvement in matching accuracy as well as precision over the selected baseline methods. The discrepancy between the test precision in and the validation precision reported in Table 4 is most likely due to a wider diversity of scenes being used for testing.

In Fig. 12 we evaluate the peakiness and smoothness of correspondence heatmaps generated by the various methods. Both of these are desirable properties as they lead to better selectivity and interpretability, while reducing ambiguity in the resultant heatmaps. To perform this evaluation we compare the shape of the heatmaps at locations surrounding the point of correspondence. We normalize the heatmaps of the successful matches, for each method, such that their dynamic range is comparable, and their peaks are aligned. We then generate the median heatmaps and analyze the row and column cross-sections, relative to the global maximum peak.

From Fig. 12, it is evident that both NCC and PSiam approaches suffer from a high number of local maxima which leads to a lower dynamic range in the heatmaps, and a less interpretable result. Our proposed solution on the other hand has a tendency to produce smooth heatmaps with a single global maximum for accurately matched results.

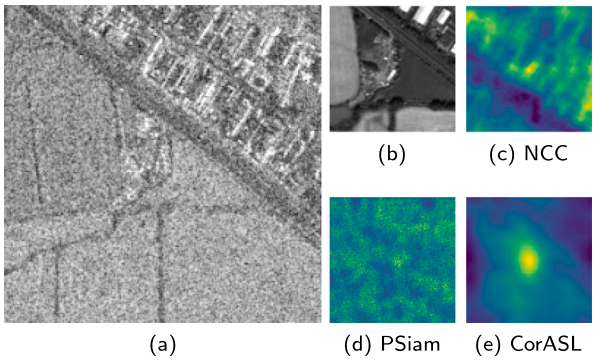

Fig. 13. A positive matching result where (c) NCC, (d) pseudo-siamese (Hughes et al., 2018), and the proposed approach (e), could all find the correspondence of the template patch (b) within the search region (a) with an accuracy of ≤ 1 pixel. The true point of correspondence is located at center point of (a),(c),(d) and (e). For brevity only the *valid* region of the heatmaps in (c–e) is depicted, and the colormap is standardized across heatmaps and covers the range 0-blue to 1-yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

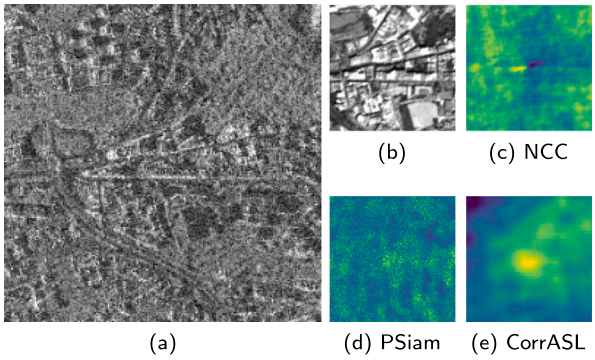

Fig. 14. An inaccurate match, where (c) NCC, (d) pseudo-siamese (Hughes et al., 2018), and the proposed approach (e), all had a matching error of between 3 and 5 pixels when matching the template patch (b) within the search region (a). The expected point of correspondence is in the center of (a), (c), (d) and (e), however, we can see it is slightly offset from center in (c),(d) and (e). For brevity only the *valid* region of the heatmaps in (c–e) is depicted, and the colormap is standardized across heatmaps and covers the range 0-blue to 1-yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We further investigate the quality of the produced correspondence heatmaps through a qualitative process by evaluating a subset of example heatmaps. This subset was selected based on scenes where all three methods obtained a similar matching accuracy. We thus evaluated the correspondence heatmaps in three categories, namely, positive matches (less than 1 pixel error), inaccurate matches (between 3 and 5 pixels error) and unsuccessful matches where by the $L2$ error is larger than 7 pixels. An example result for each category can be seen in Figs. 13, 14 and 15, respectively. For each heatmap the true point of correspondence is the center of the search window, and all heatmaps are computed with valid padding.

Fig. 13 shows the single global peak produced using the correspondence network, compared to the reasonable NCC result, and the very noisy PSiam heatmap. The same trends continue when observing matches with slight inaccuracies, in Fig. 14, although in this case the result achieved with the proposed method looses smoothness and local maxima begin to develop. Finally, in the case of unsuccessful matching, Fig. 15, the heatmap shape for all methods deteriorates to have multiple local maxima, although these all occur along the direction of ambiguity. Figs. 14 and 15 indicate that our method fails in a predictable manner,

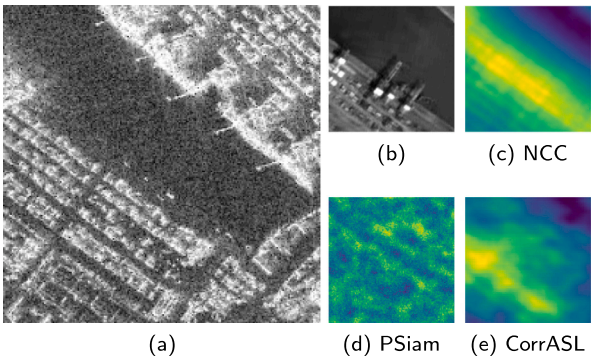

Fig. 15. An unsuccessful match, where (c) NCC, (d) pseudo-siamese (Hughes et al., 2018), and the proposed approach (e), all had a matching error larger than 7 pixels when matching the template patch (b) within the search region (a). The true point of correspondence is located at center point of (a),(c),(d) and (e). For brevity only the *valid* region of the heatmaps in (c–e) is depicted, and the colormap is standardized across heatmaps and covers the range 0-blue to 1-yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
Binary classification performance of the goodness network for the task of predicting which patches within the test dataset lead to accurate matching. The first two rows reflect the classification being performed on single modality goodness only, while the last two rows represent the results for goodness values based on minimum and maximum fusion, respectively.

Modality	Accuracy	Precision	Recall
SAR	63.6	68.9	69.0
Optical	65.1	69.8	71.3
Cross-Min	62.1	**75.1**	61.6
Cross-Max	**67.0**	66.4	**88.7**

and thus the hypothesis, that correspondence heatmaps can be used directly for the detection of outliers, holds true.

5.3. Goodness results

To gain an understanding for the performance of the domain specific goodness networks, as well as the effects of minimum or maximum fusion on the cross-domain goodness, we assess the binary classification performance with respect to the test dataset. This assessment is done in terms of the classification accuracy, precision and recall, which are defined as,

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{9}$$

$$precision = \frac{TP}{TP + FP}, \tag{10}$$

$$recall = \frac{TP}{TP + FN}, \tag{11}$$

where TP, TN, FN and FP represent the total number of true positives, true negatives, false negatives and false positives results, respectively. The results for this investigation are described in Table 6.

The results presented in Table 6 represent the binary classification accuracy when predicting the likelihood of a SAR-optical patch pair for generating an accurate correspondence. Therefore, the results do not reflect the correspondence accuracy under these patches, but rather how well the goodness network can identify good patches under various configurations. The overall and relatively low accuracy of the goodness network, see Table 6, highlights the complexity of determining matchable regions across vastly heterogeneous domains, such as SAR and optical. However, by comparing the cross-domain goodness results we see an improvement in the precision of the goodness network when using minimum fusion, and a large improvement in recall when using maximum fusion. These results highlight the effect of various fusion operators on the final set of candidate points which are later used for matching. By analyzing the results further, it is clear that the minimum fusion operator biases the goodness network towards a stricter classification, which in turn leads to a decreased false positive rate, but at the cost of a reduced set of candidate patches. Conversely, the opposite is true for maximum fusion, which is biased towards selecting a higher number of candidate patches, at the cost of more false positive classifications.

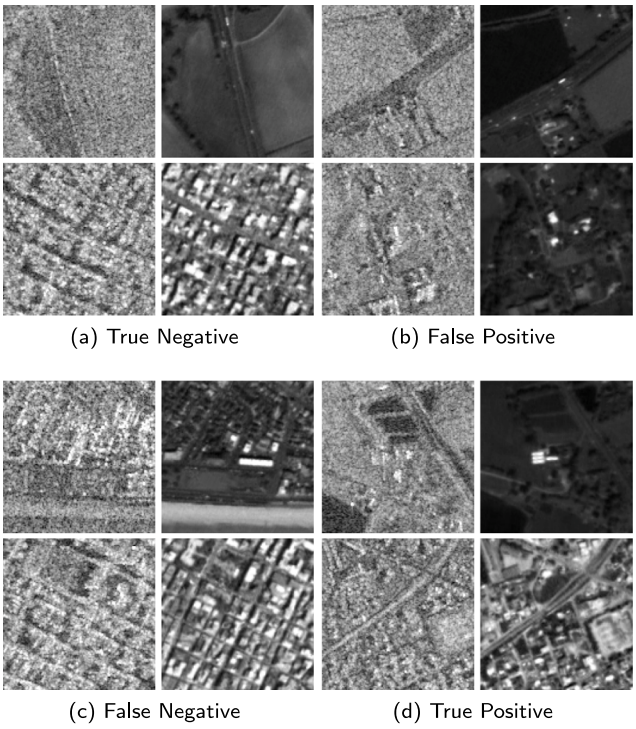

Fig. 16. Examples of regions of low and high goodness (a) and (d) respectively, along with misclassified regions (b) and (c). The SAR patch is shown on the left, and optical on the right for each of the patch pairs.

Fig. 16 depicts examples of regions with high and low goodness, as well as regions which were incorrectly classified. These example regions were drawn from the cross-domain goodness results generated using minimum fusion.

From Fig. 16 we can see that the identified regions of high goodness contain strong, unambiguous and discriminable features in both modalities; for instance road intersections, field boundaries, and clear building footprints. While the low goodness regions lack these properties, and contain features only visible in a single modality, or which lack structure. In the case of the false positive regions, strong features do exist in both domains, however, these features are potentially ambiguous or lack discriminability. The same properties can be seen in the false negative results.

While remote sensing imagery is highly calibrated, clouds and varying daylight conditions can lead to illumination variations in the optical imagery. For this reason, we evaluated the repeatability of the jointly identified points of high goodness under differing optical image brightness conditions. To do so, the optical image brightness was adjusted by a factor between [−50%, 50%] in 10% increments. The goodness network was then used to determine points of high goodness between these adjusted optical images and the original SAR image. The repeatability of the detected points, for a specific variation, was then calculated with respect the original set of points detected in the non-brightness adjusted image. It was found that the goodness network can repeatably detect between 45% to 60% of the original points under

Table 7
The percentage of the dataset used for matching (# Regions), and matching performance obtained when pre-selecting search and template patches based on their domain specific and cross-domain goodness scores. The first row represents the matching results when matching without the pre-selection of good patches.

Goodness	# Regions [%]	≤ 1px [%]	avg. L_2 [px]	mAP [px]
CorrASL	100	46.9	2.1	2.62
SAR+CorrASL	55	54.7	1.97	1.69
Opt+CorrASL	62	54.9	1.94	1.58
Cross-Min+CorrASL	48	**59.8**	**1.62**	**1.24**
Cross-Max+CorrASL	75	53.7	2.01	1.87

strong changes in brightness (±20%). However, under extreme changes (larger than 30%) in image illumination, the goodness network breaks down, and the repeatability decreased substantially.

As the purpose of the goodness network is to improve the matching accuracy of the correspondence network, by pre-selecting regions which have a higher probability of being correctly matched, we further evaluate the goodness network through the process of matching. Table 7, presents the matching performance when we match against the test patches which have been identified as having a high domain specific, or cross-modality goodness. The proportion of the original test dataset which was identified to have high goodness is described as the number of regions (# Regions).

From the results presented in Table 7 it is evident that the pre-filtering of regions, based on goodness, leads to improved matching accuracy and precision over the baseline (Table 5). Furthermore, the low percentage of good regions found in the evaluation dataset hints to the non-optimal choice of using optical domain Harris corners to create the dataset.

By comparing the matching results (Table 7) with the goodness classification results (Table 6), the effects of the various fusion approaches are evident. The minimum fusion operator (which biases the goodness network towards favoring a low false positive rate) leads to higher matching accuracy and precision at the cost of fewer identified correspondences. On the contrary, the maximum fusion (which biases the network towards ensuring all good regions are identified) leads to a lower overall matching accuracy, but identifies a higher total number of candidate patches. Based on this it is argued that the selection of fusion operator should be based in knowledge of the final application of the identified correspondences, and the sensitivity of these applications to outliers.

While the use of high goodness regions leads to improved matching performance, it comes at the cost of having fewer overall correspondences as the regions are significantly larger than those used to compute point features. This, however, is deemed to be an acceptable trade-off as many downstream tasks such as co-registration (Müller et al., 2012; Suri and Reinartz, 2010; Merkle et al., 2017) and SAR-optical stereogrammetry (Qiu et al., 2018; Bagheri et al., 2018) favour accuracy and spatial diversity over the number of correspondences.

5.4. Outlier reduction

The final component of the proposed matching pipeline is the outlier reduction network. We evaluate its performance in classifying the correspondence heatmaps of the test dataset. We further investigate the effects the inclusion of the ORN has on matching accuracy and finally we evaluate the full matching framework in an end-to-end manner on the test dataset.

A binary classification accuracy of 81%, with a precision of 76.1% and a recall of 89.5%, was achieved when evaluating the ORN on the test dataset. This shows that the classification of successful matches can be achieved based on the correspondence heatmap alone. Fig. 17, provides visual examples of both positive and negative classification results.

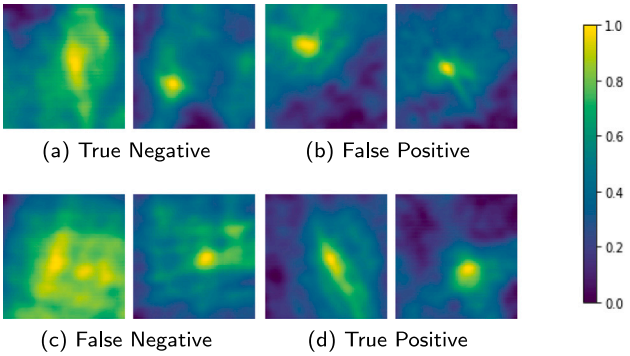

Fig. 17. Examples of heatmaps corresponding to incorrectly (a) and correctly (d) matched regions, along with mis-classified correspondence heatmaps (b) and (c). The ORN only makes use of the *valid* region of the heatmap for classification.

Table 8
Final matching results after removing outliers as classified by the ORN.

Dataset	≤ 1px [%]	avg. L_2 [px]	mAP [px]
CorrASL+ORN	54.1	1.30	1.09
Cross-Min+CorrASL+ORN	65.2	1.71	1.13

The correspondence surface shapes, as shown in Fig. 17, highlight that the network relies on more than just the local characteristics of the peak for classification although these do appear to have a relatively strong effect.

In Table 8 we investigate the effect of the outlier reduction network on matching performance. To do so we apply the ORN to the matching heatmaps of both the test dataset matching results, as shown in Table 5, as well as the minimum fusion (Cross-Min) goodness results, Table 7. The latter resulting in an equivalent end-to-end evaluation of the network.

From Table 8 it is clear that the addition of the outlier reduction network substantially increases the accuracy of the resultant set of correspondences, irrespective of the features or regions used for matching. However, the matching performance using the full framework in an end-to-end manner achieves an overall better result with higher accuracy and improved precision.

5.5. Large-scale scene matching

While we have evaluated the performance the individual sub-components of our framework, as well as the framework as a whole, these investigations have remained limited to the patch-based test dataset. Thus to fully evaluate the end-to-end performance and applicability of our proposed framework, we apply it to the task of determining correspondence on a large-scale test scene (approximately 0.8km × 1.8km) which has not undergone manual co-registration. The example scene is taken from the city of Portsmouth, England and is depicted in Fig. 18 with the final set of correspondences overlaid.

To examine the improvement in co-registration we take the mean shift derived from the final set of correspondences and apply this to the optical scene in order to align it with the SAR image. The checkerboard overlays in Fig. 19a,c depict subsets of the original, non-coregistered scene. While Fig. 19b,d show the same subsets after the alignment has been adjusted using the mean shift of the predicted set of correspondences. The mean shift was found to be $(11.03, -12.74)$ pixels with a standard deviation of $(1.99, 2.20)$ pixels in the x and y dimensions, respectively.

From Fig. 18 it can be seen that our proposed framework does not produce a large set of correspondences. However, Fig. 19 highlights the accuracy and utility of these correspondences in being able

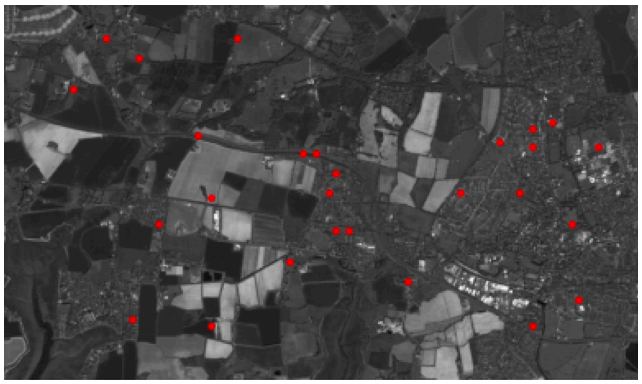

Fig. 18. The final set of correspondences superimposed, as red dots, on the PRISM optical image, taken near the city of Portsmouth, England. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to, seemingly, accurately co-register SAR and optical imagery. While our method for correcting co-registration can be improved by using the correspondences as GCPs to correct the overall optical sensor model (Müller et al., 2012), in the case of our relatively small and flat test scene, such an approach is unlikely to provide a large increase in accuracy over the mean shift method which we followed.

Although these results demonstrate the ability of our proposed methodology to accurately determine correspondences between SAR and optical imagery across spatially diverse European test scenes, it is unlikely that they are reflective of the networks performance in regions which have significantly different natural, and man-made structures. For this to be the case the proposed framework would need to be retrained, or fine-tuned on a globally diverse dataset of corresponding pairs of SAR and optical imagery.

6. Conclusion

In this paper we proposed an end-to-end framework for the sparse matching of SAR and optical imagery. The framework consists of three sub-components, each of which were trained to perform a specific task within the standard proposal, matching, outlier detection pipeline. The goodness network proposes candidate patches with a high chance of being matchable in both domains. The correspondence network performs cross correlation on a multi-scale, feature space to produce a correspondence heatmap, which is finally filtered by the outlier reduction network in order to reduce the number of false positive correspondences.

We demonstrated that, individually, each of these sub-components improves the matching accuracy and precision achieved on a test dataset in comparison to existing SAR-optical matching approaches, namely NCC (Burger and Burge, 2009) and pseudo-siamese (Hughes et al., 2018). We further evaluated the pipeline in and end-to-end manner and showed that it was able to achieve and average L_2 (distance to ground truth correspondence) of 1.71 pixels with a precision of 1.13 pixels. Finally, we demonstrated the effectiveness of our framework in producing an accurate set of correspondences which can be applied to the task of improving the overall geo-localization accuracy of optical imagery.

Still, there is room for improvement: the size of the final correspondence set is mostly limited by the goodness and outlier reduction networks. Thus, in future work we will investigate alternative architectures for the goodness network which can operate on the full scale image while still accounting for the offsets between domains. Furthermore, recent research has shown success in progressive training strategies, whereby multiple sub-components are trained in an iterative and alternating manner (Karras et al., 2017; Shaham et al., 2019). The application of such an approach to training the goodness and correspondence network could reduce the effects of the non-optimally selected training points, by allowing the network to refine these locations iteratively, and thus potentially lead to improved performance.

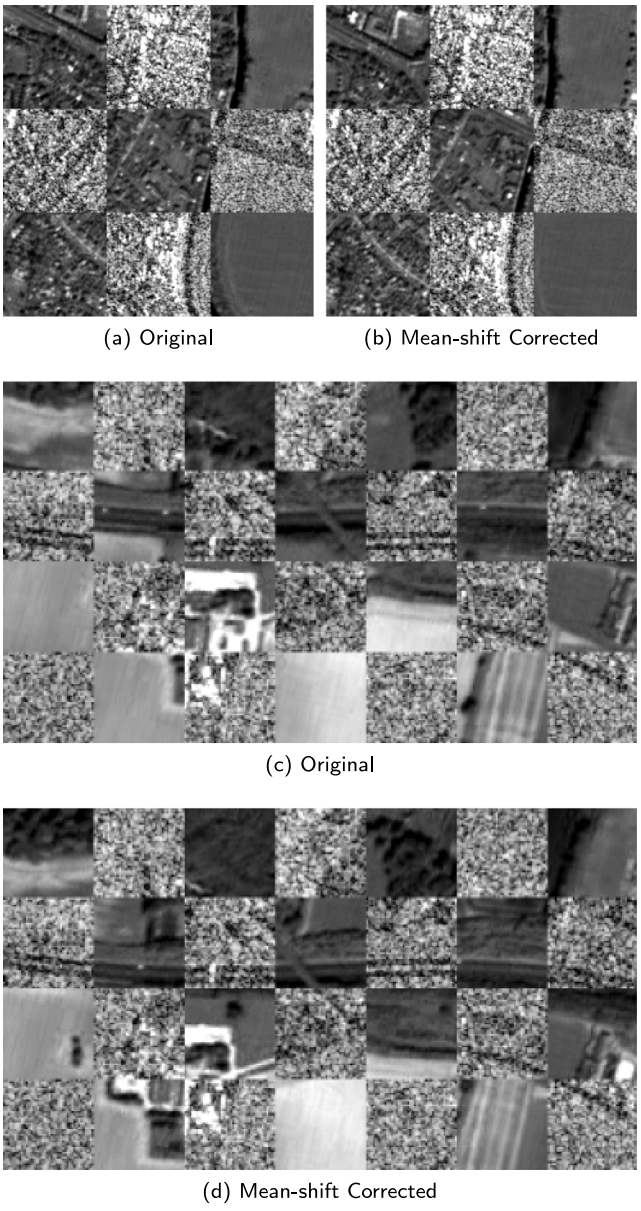

Fig. 19. Checkerboard overlays comparing the alignment of a TerraSAR-X image to the original (non-coregistered), and mean-shifted optical imagery for two subsets of the Portsmouth, England test scene. The original imagery is depicted in (a) and (c), while the mean-shift, correct imagery is shown in (b) and (d). All images have a pixel spacing of 2.5 m.

Declaration of competing interest

One or more of the authors of this paper have disclosed potential or pertinent conflicts of interest, which may include receipt of payment, either direct or indirect, institutional support, or association with an entity in the biomedical field which may be perceived to have potential conflict of interest with this work. For full disclosure statements refer to https://doi.org/10.1016/j.isprsjprs.2020.09.012.Devis Tuia is associate editor in the journal (also in the special issue the paper is being submitted to). He has instructed the other AEs not to disclose in any

way information about the reviewing process and has not and will not put any pressure for acceptance of the paper.

Acknowledgment

This work is supported by the German Research Foundation (DFG) as grant SCHM 3322/1-2.

References

Bagheri, H., Schmitt, M., d'Angelo, P., Zhu, X.X., 2018. A framework for SAR-optical stereogrammetry over urban areas. ISPRS Journal of Photogrammetry and Remote Sensing 146, 389–408.

Balntas, V., Johns, E., Tang, L., Mikolajczyk, K., 2016a. PN-Net: Conjoined triple deep network for learning local image descriptors. arXiv preprint arXiv:1601.05030.

Balntas, V., Riba, E., Ponsa, D., Mikolajczyk, K., 2016. Learning local feature descriptors with triplets and shallow convolutional neural networks. In: Proc. British Machine Vision Conference. British Machine Vision Association, p. 3.

Burger, W., Burge, M.J., 2009. Principles of Digital Image Processing, Vol. 54. Springer London.

Bürgmann, T., Koppe, W., Schmitt, M., 2019. Matching of TerraSAR-X derived ground control points to optical image patches using deep learning. ISPRS Journal of Photogrammetry and Remote Sensing 158, 241–248.

Citak, E., Bilgin, G., 2019. Visual saliency aided SAR and optical image matching. In: Proc. Innovations in Intelligent Systems and Applications Conference. IEEE, pp. 1–5.

Dellinger, F., Delon, J., Gousseau, Y., Michel, J., Tupin, F., 2015. SAR-SIFT: A SIFT-like algorithm for SAR images. IEEE Transactions on Geoscience and Remote Sensing 53 (1), 453–466.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T., 2019. D2-Net: A trainable CNN for joint description and detection of local features. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 8092–8101.

Fischer, P., Dosovitskiy, A., Brox, T., 2014. Descriptor matching with convolutional neural networks: A comparison to SIFT. arXiv preprint arXiv:1405.5769.

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24 (6), 381–395.

Gong, M., Zhao, S., Jiao, L., Tian, D., Wang, S., 2014. A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information. IEEE Transactions on Geoscience and Remote Sensing 52 (7), 4328–4338.

Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C., 2015. MatchNet: Unifying feature and metric learning for patch-based matching. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3279–3286.

Hariharan, B., Arbelaez, P., Girshick, R., Malik, J., 2015. Hypercolumns for object segmentation and fine-grained localization. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 447–456.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proc. IEEE International Conference on Computer Vision. IEEE, pp. 1026–1034.

Hoffmann, S., Brust, C.-A., Shadaydeh, M., Denzler, J., 2019. Registration of high resolution SAR and optical satellite imagery using fully convolutional networks. In: Proc. IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 5152–5155.

Hughes, L.H., Merkle, N., Burgmann, T., Auer, S., Schmitt, M., 2019. Deep learning for SAR-optical image matching. In: Proc. IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 4877–4880.

Hughes, L.H., Schmitt, M., 2019. A semi-supervised approach to SAR-optical image matching. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences IV-2/W7, 71–78.

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

Iglovikov, V., Shvets, A., 2018. Ternausnet: U-Net with VGG11 encoder pre-trained on imagenet for image segmentation. arXiv preprint arXiv:1801.05746.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2017. Progressive growing of GANs for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kuppala, K., Banda, S., Barige, T.R., 2020. An overview of deep learning methods for image registration with focus on feature-based approaches. International Journal of Image and Data Fusion 1–23.

Li, J., Hu, Q., Ai, M., 2020. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. IEEE Transactions on Image Processing 29, 3296–3310.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Internationl Journal of Computer Vision 60 (2), 91–110.

Ma, W., Wen, Z., Wu, Y., Jiao, L., Gong, M., Zheng, Y., Liu, L., 2017. Remote sensing image registration with modified SIFT and enhanced feature matching. IEEE Geoscience and Remote Sensing Letters 14 (1), 3–7.

Ma, W., Zhang, J., Wu, Y., Jiao, L., Zhu, H., Zhao, W., 2019. A novel two-step registration method for remote sensing images based on deep and local features. IEEE Transactions on Geoscience and Remote Sensing 57 (7), 4834–4843.

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

Mishchuk, A., Mishkin, D., Radenoviundefined, F., Matas, J., 2017. Working hard to know your neighbor's margins: Local descriptor learning loss. In: Proc. International Conference on Neural Information Processing Systems. pp. 4829–4840.

Mou, L., Schmitt, M., Wang, Y., Zhu, X.X., 2017. A CNN for the identification of corresponding patches in SAR and optical imagery of urban scenes. In: Proc. Joint Urban Remote Sensing Event. IEEE, Dubai, pp. 1–4.

Müller, R., Krauß, T., Schneider, M., Reinartz, P., 2012. Automated georeferencing of optical satellite data with integrated sensor model improvement. Photogrammetric Engineering and Remote Sensing 78 (1), 61–74.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 8024–8035.

Qiu, C., Schmitt, M., Zhu, X.X., 2018. Towards automatic SAR-optical stereogrammetry over urban areas using very high resolution imagery. ISPRS Journal Photogrammetry Remote Sensing 138, 218–231.

Revaud, J., Weinzaepfel, P., de Souza, C.R., Humenberger, M., 2019. R2d2: repeatable and reliable detector and descriptor.. In: Proc. Neural Information Processing Systems.

Schmitt, M., Tupin, F., Zhu, X.X., 2017. Fusion of SAR and optical remote sensing data – Challenges and recent trends. In: IEEE International Geoscience and Remote Sensing Symposium. IEEE, Fort Worth, TX, USA, pp. 5458–5461.

Schneider, M., Müller, R., Krauss, T., Reinartz, P., Hörsch, B., Schmuck, S., 2010. Urban Atlas – DLR processing chain for orthorectification of PRISM and AVNIR-2 images and TerraSAR-X as possible GCP source. Internet Proceedings 1–6.

Shaham, T.R., Dekel, T., Michaeli, T., 2019. SinGAN: Learning a generative model from a single natural image. In: Proc. IEEE/CVF International Conference on Computer Vision. IEEE, pp. 4570–4580.

Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., Moreno-Noguer, F., 2015. Discriminative learning of deep convolutional feature point descriptors. In: Proc. IEEE International Conference on Computer Vision. IEEE, pp. 118–126.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition.. In: Proc. International Conference on Learning Representations.

Smith, L.N., 2017. Cyclical learning rates for training neural networks. In: Proc. IEEE Winter Conference on Applications of Computer Vision. IEEE, pp. 464–472.

Suri, S., Reinartz, P., 2010. Mutual-information-based registration of TerraSAR-X and Ikonos imagery in urban areas. IEEE Transactions on Geoscience and Remote Sensing 48 (2), 939–949.

Suri, S., Schwind, P., Uhl, J., Reinartz, P., 2010. Modifications in the SIFT operator for effective SAR image matching. International Journal of Image and Data Fusion 1 (3), 243–256.

Vargas-Muñoz, J.E., Lobry, S., Falcão, A.X., Tuia, D., 2019. Correcting rural building annotations in OpenStreetMap using convolutional neural networks. ISPRS Journal of Photogrammetry Remote Sensing 147, 283–293.

Wang, S., Quan, D., Liang, X., Ning, M., Guo, Y., Jiao, L., 2018. A deep learning framework for remote sensing image registration. ISPRS Journal of Photogrammetry Remote Sensing 145, 148–164.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. CBAM: Convolutional block attention module. In: Proc. European Conference on Computer Vision. Springer International Publishing, Cham, pp. 3–19.

Xiang, Y., Wang, F., You, H., 2018. OS-SIFT: A robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas. IEEE Transactions on Geoscience and Remote Sensing 56 (6), 3078–3090.

Ye, Y., Shen, L., 2016. HOPC: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 3, 9.

Yi, K.M., Trulls, E., Lepetit, V., Fua, P., 2016. LIFT: Learned invariant feature transform. In: Proc. European Conference on Computer Vision. Springer, pp. 467–483.

Zagoruyko, S., Komodakis, N., 2015. Learning to compare image patches via convolutional neural networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 4353–4361.