MARU-Net: Multiscale Attention Gated Residual U-Net With Contrastive Loss for SAR-Optical Image Matching

Michele Gazzea, *Student Member, IEEE*, Oscar Sommervold, and Reza Arghandeh, *Senior Member, IEEE*

***Abstract*—Accurate synthetic aperture radar-optical matching is essential for combining the complementary information from the two sensors. However, the main challenge is overcoming the different heterogeneous characteristics of the two imaging sensors. In this article, we propose an end-to-end machine learning pipeline inspired by recent advances in image segmentation. We develop a siamese multiscale attention-gated residual U-Net for feature extraction from satellite images. The siamese architecture shares weights and transforms the heterogeneous images into a homogeneous feature space. Fast Fourier transform is used to compute the cross-correlation between the feature maps and produce a similarity map. A contrastive loss is introduced to aid the training procedure of the model and maximize the discriminability of the model. The experimental results on a benchmark dataset show that the proposed method has superior matching accuracy and precision compared to other state-of-the-art methods.**

***Index Terms*—Deep learning, optical, synthetic aperture radar (SAR), template matching.**

I. Introduction

In RECENT years, the availability and accessibility of high-quality remote sensing data have skyrocketed [1]. Numerous sensors continuously monitoring the Earth unfold the possibility of synergistic use of data in various applications, such as disaster management (forest fires [2], hurricane impact [3], flood and drought monitoring [4]), agriculture (soil moisture monitoring [5], vegetation monitoring [6], [7], [8]), climate monitoring (deforestation [9], pollution [10], weather forecasting [11], ocean ecosystem [12]), urban planning [13], marine traffic monitoring [12].

Optical and synthetic aperture radar (SAR) are some of the most common sensors for Earth observation. Optical sensors are passive sensors that measure the reflected sunlight from objects on the Earth's surface, making them susceptible to atmospheric conditions, such as cloud coverage, time of day, and other weather conditions, such as mist, fog, and smoke. However, they provide useful multispectral information. Instead, SAR is an active sensor that transmits radio wave pulses and measures the back-scattered signals, making them operational without the sun needing to illuminate the surface and in every weather condition. Furthermore, SAR can capture the surface properties (such as roughness) of objects. However, it provides no spectral information, resulting in noisy black-and-white imagery. Image matching is the process of aligning two or more images. The SAR-optical matching is especially problematic due to the significant radiometric and geometric differences and the visual disparities introduced by different remote sensing sensors. Consequently, combining data from various sensor types remains one of the major challenges in remote sensing [14]. The distinct characteristics of the different imaging principles cause the imagery to reveal different aspects of the Earth's surface. As a result, objects on the surface appear inherently dissimilar from active and passive sensors' viewpoints. Therefore, locating salient features in both images is complex, especially in areas with fewer distinct features. Combining the two sensors will increase the information content and possibly open new use cases in remote sensing applications, given the complementary information from SAR and optical imagery. However, the two images must align accurately before combining them.

Traditionally, feature-based matching methods, such as SIFT [15], SAR-SIFT [16], optical-to-SAR SIFT (OS-SIFT) [17], and (RIFT) [18] have been used for SAR-optical matching. These methods calculate feature descriptors from images and match them together, evaluating the feature correspondence. Affine correction is performed by selecting noncollinear matched features as control points. However, the extraction of such features requires export knowledge and handcrafted procedures. Furthermore, they cannot handle well the heterogeneous characteristics caused by the SAR and optical imaging mechanisms on scenes with few salient features. Feature similarity is calculated using the sum of squared differences (SSD), normalized cross-correlation (NCC), or mutual information (MI) [19], [20]. However, MI-based approaches suffer from high computational costs.

Several studies have recently suggested deep learning methods to overcome the shortcomings of nonlearning methods. The availability of paired SAR-optical datasets, such as SpaceNet-6 [21] and SEN1-2 [22], help the development of machine learning-based SAR-optical matching. Zhang et al. [23] and Merkle et al. [24] proposed fully convolutional siamese networks to extract features in SAR-optical imagery. Both methods rely on a time-consuming pixel-by-pixel search to perform matching

Manuscript received 24 February 2023; revised 5 May 2023; accepted 9 May 2023. Date of publication 18 May 2023; date of current version 7 June 2023. *(Corresponding author: Michele Gazzea.)*

The authors are with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, 5036 Bergen, Norway (e-mail: mgaz@hvl.no; scar.os@hotmail.com; arghandehr@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2023.3277550

and use shallow convolutional neural networks (CNN) with few parameters as feature extractors. Hughes et al. [25] implemented a component-based framework using three separate networks to extract patches suitable for matching, perform template matching, and remove outliers, respectively. However, the framework downsamples the produced feature maps due to time complexity constraints, thus losing matching precision when interpolating the similarity score. More recently, Zhou et al. [26] proposed a machine learning modification of channel features of orientated gradients (CFOG) proposed by [27] called multiscale convolutional gradient features (MCGF). Like CFOG, MCGF achieves fast matching since it also performs similarity evaluation in the frequency domain using NCC. Zhang et al. [28] proposed the deep dense feature network (DDFN), also inspired by the CFOG method. DDFN extracts a 9-D feature vector for each pixel and uses the SSD for similarity computation. The experiments show that the deep siamese network outperforms the state-of-the-art handcrafted CFOG descriptor. Fang et al. [29] introduced the fast Fourier transform (FFT) U-Net, using the image segmentation model U-Net [30] as a feature extractor with an FFT accelerated NCC layer to perform matching. Similarly, [31] demonstrated the superiority of using the U-Net as a feature extractor in SAR-optical matching. However, SAR-optical matching remains a challenging problem due to the inherent geometric and radiometric differences between the two sensors. A review on recent methods and current research trends can be found in [32].

In this article, we tackle the problem of SAR-optical matching as a multiclass classification task. We use a siamese architecture to extract shared features, mapping different multimodal images (e.g., optical and SAR into the same space) into a common feature space. As the core of the model, we chose a UNet-based architecture because it is one of the most effective deep learning architectures for image classification and image segmentation. We enhanced the classic architecture with additional components, such as the attention mechanism and residual blocks, which were initially developed for image semantic segmentation tasks but not fully exploited yet in the context of SAR-optical template matching. Moreover, we compute the feature maps at different scales. Computing the features at different scales is a well-known method that can improve the representation ability of features in many tasks, such as standard object detection [33] and image segmentation [34]. The multiscale feature map makes the network more robust and improves the pixel-level matching accuracy. At the same time, the attention mechanism helps the model to locate and focus on salient regions in the SAR-optical imagery. Furthermore, we combine the standard cross-entropy with an additional contrastive loss to build a combined loss function tailored for the SAR-optical matching problem. The loss function reduces false positive matching locations and increases the discriminability of the proposed framework.

II. Methodology

The approach used in this work is based on template matching, which consists of finding the most likely position of a small image (template) within a larger image (reference). As such, the starting point is acquiring an optical image (at this moment also called reference) with dimensions $R_x \times R_y$ and an SAR image (at this moment also called template) with dimensions $T_x \times T_y$. Fig. 1 outlines the structure of the proposed pipeline. Details of each component, including the architecture, FFT NCC layer, and loss function, are described as follows.

A. MARU-Net Architecture

As a preprocessing step, the two input images (i.e., optical and SAR) are downscaled, reducing their original size by half. The resulting four images (optical, SAR, and their corresponding downscaled versions) are passed through a siamese CNN composed of four units. Each unit has the same architecture and shares the same weights with the others. This way, they work in tandem on different input vectors to compute comparable feature maps [35].

Each unit is a CNN that consists of a U-Net backbone where standard convolution blocks were replaced by residual convolution blocks [36]. The architecture has four layers with channel dimensions of {32, 64, 128, 256} in the contracting path and similarly four layers of {256, 128, 64, 32} in the expanding path. We inserted attention gates [37] in the expanding path instead of the standard direct skip-connections. Each residual block consists of two iterations of 3×3 2-D convolutions, followed by batch normalization and an ELU activation. The shortcut path consists of one convolutional layer. Instead of using traditional transposed convolutions with learnable parameters to upscale the encoded feature maps, we use upsampling layers with bilinear interpolation to increase the resolution of the feature maps and thus preserve the initial details of the encoded features. This is because CNN architectures employing transposed convolutions from lower to higher resolution are prone to checkerboard artifacts [38].

Each network produces a 4-channel feature tensor ψ with the same dimensions (height and width) of the corresponding input. The feature map extracted from the downscaled optical and SAR images, ψ^d_{opt} and ψ^d_{SAR}, are upscaled and concatenated with the corresponding feature map extracted from the original images ψ^o_{opt} and ψ^o_{SAR}. Thus, $\psi_{\text{opt}} = \psi^o_{\text{opt}} \otimes \psi^d_{\text{opt}}$ and $\psi_{\text{SAR}} = \psi^o_{\text{SAR}} \otimes \psi^d_{\text{SAR}}$, resulting in a 8-channel feature tensor.

B. FFT NCC Layer

Comparing the feature maps pixelwise is time-consuming and drastically increases the training and inference time of the model. To speed up the process, we compute the NCC in the frequency domain to evaluate the similarity map S of the derived feature maps using the FFT as

$$S = F^{-1}_{2\,d}\left[F_{2\,d}\left(\psi_{\text{opt}}\right) \cdot F_{2\,d}\left(\psi_{\text{SAR}}\right)\right] \tag{1}$$

where, S denotes the derived similarity map, ψ_{opt} and ψ_{SAR} are the optical and SAR feature maps produced by the network presented in the previous section, "·" is the elementwise product operation, and $F_{2\,d}$ and $F^{-1}_{2\,d}$ denote the 2-D forward and inverse FFT, respectively. In the FFT layer, the dimensions of S corresponds to the dimensions of the search space and is $S_x \times S_y \times 8$, where $S_x = R_x - T_x + 1$, and $S_y = R_y - T_y + 1$. The similarity map S is then normalized into $\tilde{S}$ according to [39]. As a result,

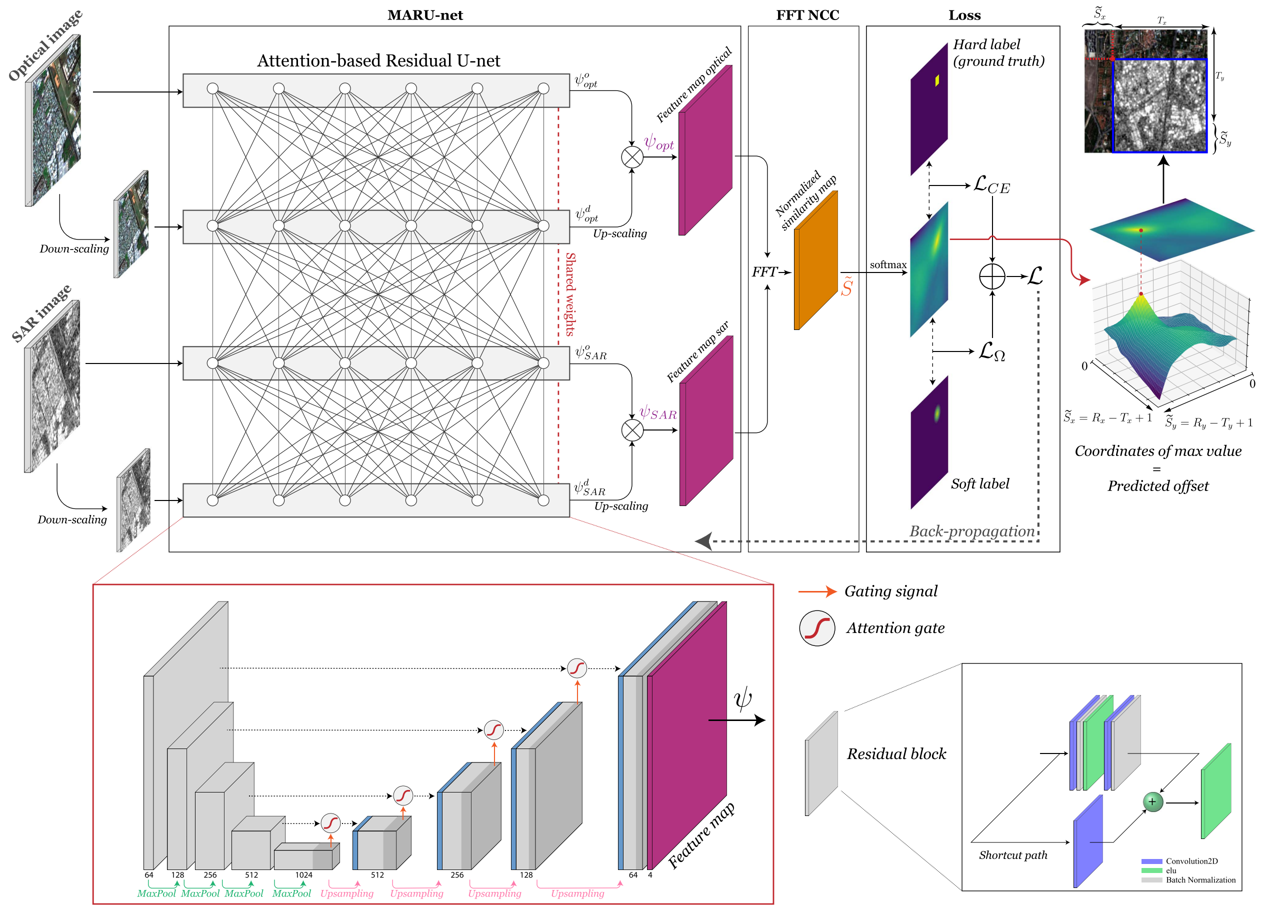

Fig. 1. Overview of the proposed architecture. Optical and SAR images are downscaled, and their resolution is halved. The four images (e.g., optical and SAR with original resolution and optical and SAT halved resolution) are fed into a 4-unit siamese neural network. Each unit has the same architecture and shares the same weights. A feature map ψ is then produced for each image. The feature map computed from the original resolution and halved resolution are concatenated together for both optical and SAR. These are then compared in the FFT NCC layer. The coordinates of the maximum value in the normalized similarity map are the predicted shift of the smaller SAR template within the larger optical reference image.

every value in $\widetilde{S}$ can be interpreted as the observed similarity of the template (i.e., SAR image) within the reference (i.e., optical image).

C. Loss Function

If a *softmax* function is applied to $\widetilde{S}$ each value in the derived similarity map can be interpreted as the probability of a specific shift between the reference and the template. That is, the coordinates in the similarity map $\widetilde{S}$ corresponding to the maximum value indicate the predicted shift of the template with respect to the reference. Given the discrete dimensions of the search space and having the ground truth with the correct shift, locating the 2-D pixel shift between the reference (e.g., optical) and the template (e.g., SAR) image can be formulated as a multiclass classification problem, where the classes denote the shift coordinates of the SAR template within the larger optical image. As such, we adopt the cross-entropy loss function $\mathcal{L}_{\text{CE}}$ as

$$\mathcal{L}_{\text{CE}} = -\sum_{i}^{S_x}\sum_{j}^{S_y} y_{i,j}\log\left(S_{i,j}\right) + \left(1-y_{i,j}\right)\log\left(1-S_{i,j}\right) \tag{2}$$

where, $y_{i,j}$ is the ground truth value at position (i, j), while $\widetilde{S}_{i,j}$ is the similarity score at position (i, j).

However, for such a classification task, the size of $\widetilde{S}$ yields $\widetilde{S}_x \times \widetilde{S}_y$ different classes where the correct matching location (1 class) is considered as the correct class, while the all the rest are considered wrong. Therefore, this formulation results in a heavily imbalanced distribution of classes, which negatively impacts the training. Inspired by [28], we include a new term in the loss function to reduce the impact of the imbalanced distribution of the classes and improve the discriminability of the network. We apply the discrete approximation of the Gaussian function G on the area around the correct matching position (c_i, c_j) obtained from the ground truth to construct a soft ground truth map as

$$G_{ij} = \begin{cases} \frac{1}{2\pi\sigma}\cdot e^{-\frac{\left\|(i,j)-(c_i,c_j)\right\|_2^2}{2\sigma}}, & \text{if}\left\|(\mathrm{i,j})-(\mathrm{c_i,c_j})\right\|_2 < 2 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where, σ is set to 1, $\|\cdot\|_2$ is the L_2 (Euclidean) distance. L_1 normalization is applied to G to render it as a probability distribution. G has the same size of the similarity map $\widetilde{S}$.

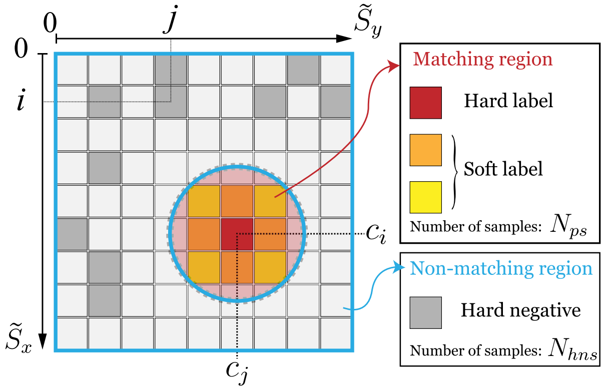

Fig. 2. Grid is the generated similarity map $\widetilde{S}$. The hard label (red square) is provided by the ground truth and corresponds to the correct match between the satellite and SAR. The Gaussian weighting function G is applied to the hard label to produce the soft labels (orange and yellow squares). The union of hard and soft labels is called the matching region. Points outside the matching region fall inside the nonmatching region. The N_{hns} points with the most negative nonzeros values are selected within this region. These points are called hard negative samples (gray squares).

The similarity map space is then divided into two nonoverlapping regions, namely the matching region $\text{MR} = \widetilde{S} \times G$ and the nonmatching region $\text{NMR} = \widetilde{S} \times (1 - \lceil G \rceil)$, where $\lceil G \rceil$ is the ceiling function applied to G (nonzeros values are mapped into 1, zeros values remain zero). The matching region consists of pixels corresponding to the true class provided by the ground truth (hard label) and the nearby classes, weighted by the Gaussian function G (soft labels). We denote the number of these classes by N_{ps}. Inside the NMR region, we select the N_{hns} pixels with the lowest nonzero values in the similarity map (hard negative samples). Fig. 2 shows the two regions and the different types of labels graphically.

We define a new term Ω, formulated as the difference between the observed similarity scores within the two regions, which we aim to maximize. Mathematically

$$\Omega = \frac{1}{N_{ps}} \sum_{k=1}^{N_{ps}} \text{MR}(k) - \frac{1}{N_{\text{hns}}} \sum_{k=1}^{N_{\text{hns}}} \text{NMR}(k) + 1. \tag{4}$$

Like in [28], we add a margin of 1 to prevent significantly low values in Ω and increase the separability of the positive and negative samples. Experimentally, we find that setting $N_{\text{hns}} = 16$ yields the best results. Finally, the contrastive loss $\mathcal{L}_\Omega$ is defined as $\mathcal{L}_\Omega = -\Omega$ to make it compatible with the cross-entropy term $\mathcal{L}_{\text{CE}}$ (e.g., minimize $-\Omega$ is equivalent to maximize Ω). We construct the combined loss function, which is the sum of the cross-entropy term and the contrastive term

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_\Omega. \tag{5}$$

The training process aims to minimize (5) through backpropagation. In practice, without the contrastive term Ω, we note that the simple cross-entropy loss $\mathcal{L}_{\text{CE}}$ does not distinguish well between a mismatch close to the ground truth or far from it. With the term Ω, the model is penalized less if the maximum value in $\widetilde{S}$ is close to the ground truth. At the same time, the model is trained to distinguish better between positive samples and hard negative samples.

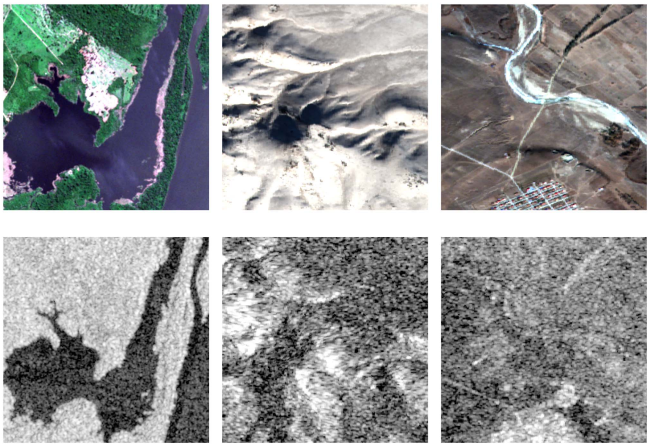

Fig. 3. Sample images taken from the SEN1-2 dataset. First row: images taken from Sentinel-2 (optical). Second row: images taken from Sentinel-1 (SAR).

III. Experiments

A. Dataset

We use the SEN1-2 open benchmark dataset [22] to examine the performance of our approach. The dataset consists of 282 384 coregistered SAR-optical image patches, including all four seasons and environments (e.g., urban, rural, deserts, mountains, etc.). The image patches are 256 × 256 pixels in size and have a 10-m spatial resolution. Fig. 3 shows some examples of the images taken from the SEN1-2 dataset.

We select 100 random patches from every folder in the dataset across all four seasons. The selected subsection of the dataset is then split into training and test sets with a ratio of 70:30, yielding 18.060 image pairs as our training data and 7.740 for testing. To generate the ground truth, the SAR patches are randomly cropped to a size of 192 × 192, where the row and column offset is stored as the ground truth value. The RGB optical images are converted to grayscale, and the SAR images are denoised using the Lee filter [40].

B. Evaluation Metrics and Implementation Details

The L_2 distance between the peak of the generated similarity map $\widetilde{S}$ and the ground truth is used to determine the matching accuracy. With a 256 × 256 reference image and a 192 × 192 template, the resulting similarity map is a 65 × 65 matrix. We select MI, siamese CNN, DDFN, and FFT U-Net to compare against our proposed MARU-Net method. Models are trained for five epochs with a batch size of 4 using the Adam optimizer with a learning rate of $5e^{-4}$. All methods are implemented using the Keras API of TensorFlow 2, and training is performed on a machine with an Intel Core i5-7600 k CPU and a GeForce GTX 1080 GPU.

TABLE I
MATCHING RESULTS ON THE SEN1-2 DATASET

Methods	Accuracy				Precision	Avg time
	$\leq 1px$	$\leq 2px$	$\leq 3px$	$\leq 5px$	avg L_2	ms
Standard NCC [42]	0.05	0.11	0.16	0.23	31.6	**57**
DDFN [28]	0.30	0.42	0.50	0.55	17.5	239
MI [20]	0.30	0.45	0.54	0.62	11.8	5450
Siamese CNN [24]	0.42	0.61	0.72	0.78	8.27	629
FFT U-Net [29]	0.44	0.63	0.74	0.80	6.92	437
Proposed	**0.55**	**0.75**	**0.82**	**0.87**	**4.94**	396

The bold values represent the best score column-wise.

TABLE II
ABLATION STUDY: COMPONENTWISE COMPARISON

Methods	Accuracy (%)				Precision
	$\leq 1px$	$\leq 2px$	$\leq 3px$	$\leq 5px$	avg L_2
U-Net with cross-entropy	0.44	0.63	0.73	0.79	6.92
+ Contrastive loss	0.46	0.67	0.77	0.83	5.79
+ Residual and Attention Gate	0.48	0.70	0.80	0.85	5.21
+ Multiscale (MARU-Net)	**0.57**	**0.75**	**0.82**	**0.87**	**4.94**

The bold values represent the best score column-wise.

C. SAR and Optical Matching Performance Validation

We compute the results using our proposed approach as well selected state-of-the-art methods [32]. The obtained results of the testing dataset are shown in Table I. Besides the more advanced learning methods, we also evaluate the standard cross-correlation method, using the implementation provided by [41]. To evaluate the performance, we select the percentage of image pairs with a L_2 pixel distance from the ground truth smaller than a given threshold as the correct matching rate (CMR). We also compare the average L_2 distance value as a measure of precision and the time complexity, measuring the average time to perform a single matching.

As shown in Table I, the proposed MARU-Net method obtains the best accuracy across all CMR thresholds. In addition, it exhibits the best precision compared to the usual methods. We note that the standard NCC, although faster than the more advanced methods, shows very poor performance. This is because the similarity map is performed directly on the input images, which are too diverse. MI shows acceptable results on the coarse 10 m/px resolution imagery (which is the resolution of the SEN1-2 dataset used in this study), although other studies have shown that MI performs noticeably worse when it comes to pixel-level accuracy on very high-resolution imagery [29], [43]. Nevertheless, it is also the most time-consuming among the tested methods. The DDFN achieves the fastest performance due to its shallow seven-layer CNN structure with only 300 000 parameters. However, the shallow nature of DDFN yields low precision scores due to poor matching accuracy on imagery with few salient features. The siamese CNN is also a shallow network with few parameters but employs a new architecture with shared weights and cross-entropy as a loss function, yielding considerably better performance compared to DDFN. Still, the Siamese CNN is the slowest among the machine learning methods due to a time-consuming dot-product computation of the extracted feature vectors. The state-of-the-art FFT U-Net utilizes a deep classic U-Net with cross-entropy as a loss function, producing the best results among the selected methods, as shown in Table I. The experimental results show that the proposed MARU-Net architecture yields significant improvements in matching performance compared to other state-of-the-art methods. In addition, our method is also computationally efficient, in line with the other methods.

The approach used in this work, and the other cited works, is based on template matching. However, this approach will lead to difficulties when the two images are heavily warped one with respect to the other. Most of the presented methods cannot effectively deal with significant rotation and scale differences between the two image and further research is needed to address these issues.

D. Visual Comparison of Matching Results

In Fig. 4, we show qualitatively two samples and the produced similarity maps using different methods. The chosen scenes consist of a nonurban and an urban image pair. Low response values in the similarity map are represented by a dark blue color, while high values are bright yellow. Ideally, an optimal matching result would be a single sharp yellow peak overlapping the ground truth value (red dot).

In the first scene, a snowy mountain scene, we observe no distinct features and few details, making the SAR-optical matching a challenging task. As such, the selected methods exhibit an unfocused response pattern with a low response in the correct matching region. MI, DDFN, and Siamese CNN particularly fail, with numerous peaks in incorrect areas. The FFT U-Net also yields an unfocused similarity map with a moderate response in the matching area. The proposed method shows a comparable similarity map with a singular sharp peak close to the red dot. The attention gates and combined loss function encourage the network to focus on a single region resulting in more focused similarity maps with sharper peaks.

The second scene depicts an urban area with detailed structures (river, buildings). Compared to the first scene, the matching is considerably more manageable, and all methods increase the matching performance since features, such as the river, are salient in both images. Still, the siamese CNN appears relatively unfocused. The main benefit of our proposed method is a more robust matching performance in scenes like in $S1$ where the level of details is low and few features are present.

E. Ablation Study

We perform an ablation study to verify how the components in the proposed method contribute to increasing the matching performance. The FFT U-Net method serves as a baseline, and the different components of our method are gradually added to verify the performance gain. The experimental results of the componentwise comparison are shown in Table II.

Including contrastive loss improves the matching precision, ensuring improved training compared to solely cross-entropy.

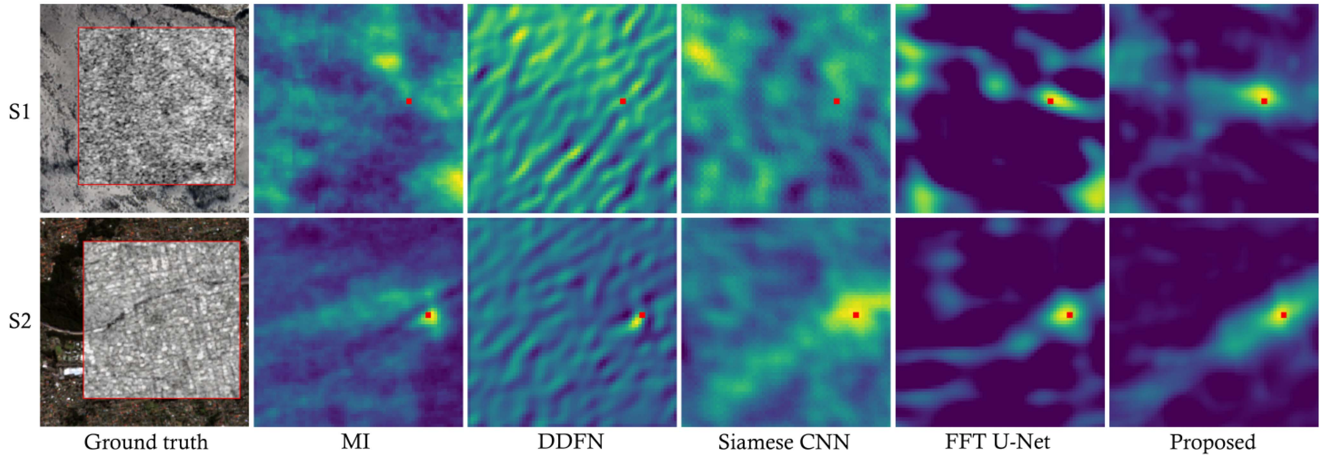

Fig. 4. Matching results for different methods on two sample scenes in the SEN1-2 dataset: S1 (rural area), S2 (urban area). The generated similarity heatmap is color coded from blue to yellow. Low values are represented by a dark blue color, while high values are bright yellow. The red dot denotes the true ground truth offset coordinates. Qualitatively, a better performance corresponds to having a single yellow peak around the red dot.

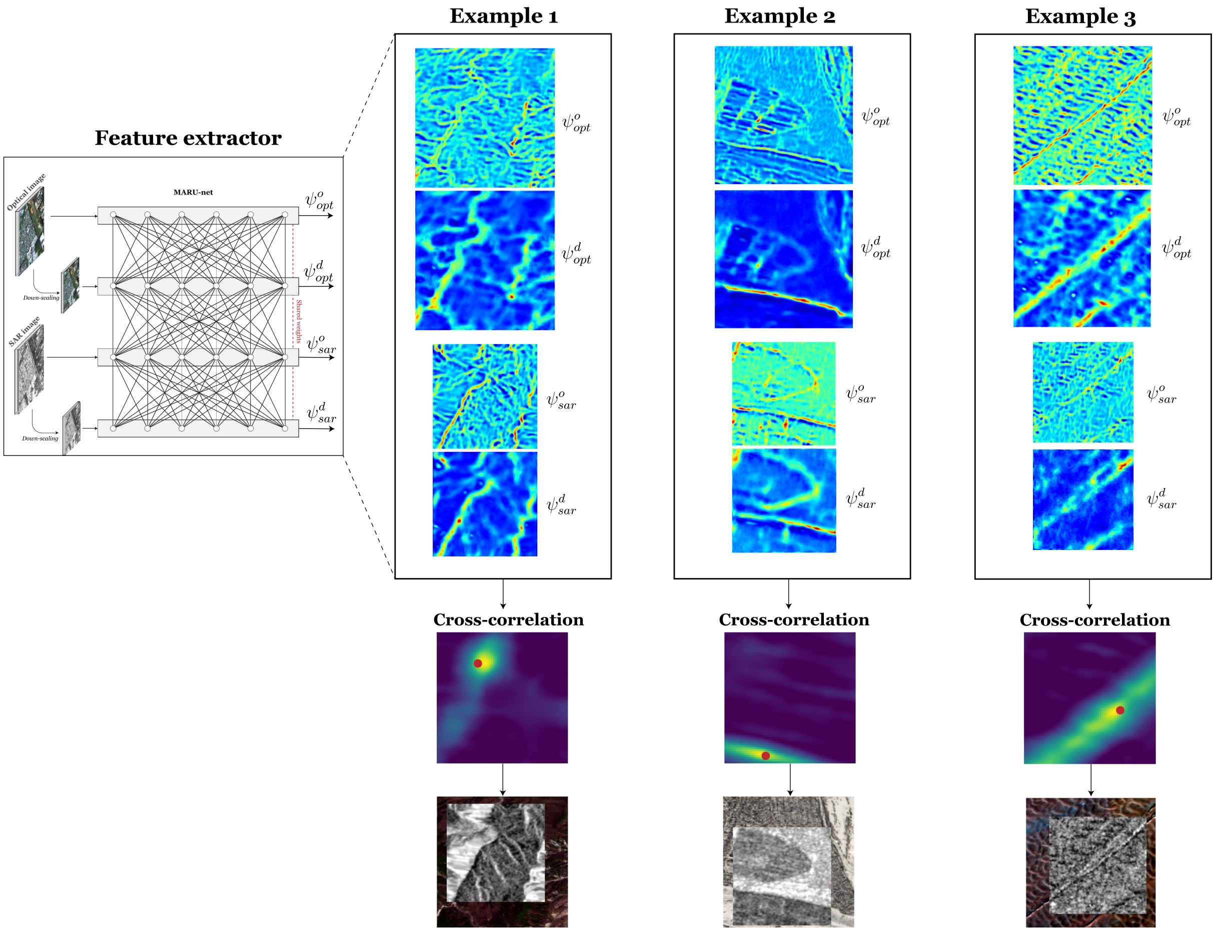

Fig. 5. Learned feature maps $\psi^o_{\text{opt}}, \psi^o_{\text{SAR}}, \psi^d_{\text{opt}}, \psi^d_{\text{SAR}}$. The feature maps are extracted from the MARU-Net network before calculating the cross-correlation block.

Fig. 5 shows the feature maps extracted from the feature extractor. For clarity, the feature extractor from Fig. 1 is visualized here in the left part. The learned feature maps are $\psi^o_{\text{opt}}, \psi^o_{\text{SAR}}, \psi^d_{\text{opt}}, \psi^d_{\text{SAR}}$. We notice that the downscale feature maps $\psi^d_{\text{opt}}, psi^d_{\text{SAR}}$, differently from the original resolution feature maps $\psi^o_{\text{opt}}, \psi^o_{\text{SAR}}$ suppress the details, keeping only the most significant features and thus making the match more robust. On the other hand, due to the decreased resolution, small details are canceled out. Such missing information is retrieved by combining the feature maps from the original resolution, which improves the pixel-level matching accuracy. In our study, we select two scales (original resolution and halved). Testing with more scales did not improve the results, as further decreasing the resolution would delete too much information and details,

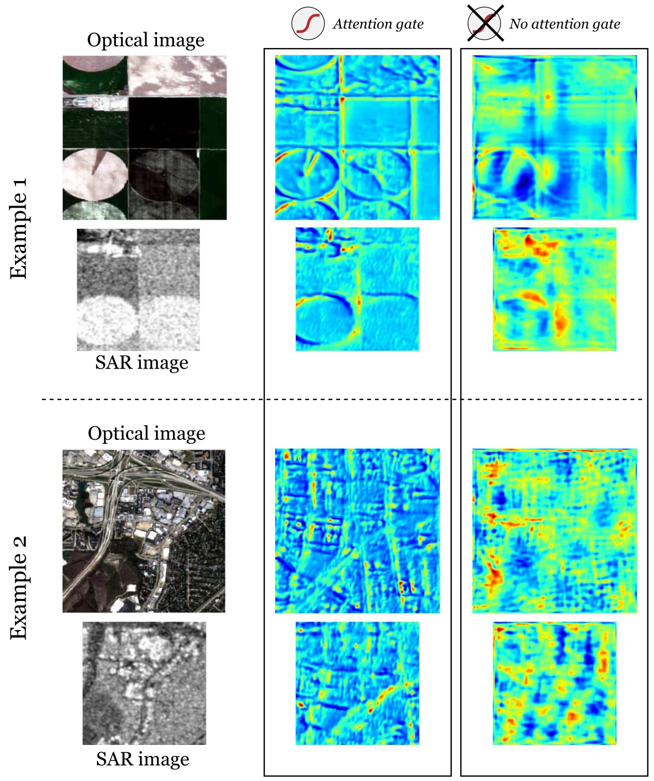

Fig. 6. Comparison of the learned feature maps with and without attention mechanism.

while making the model even more difficult to train (due to more parameters). However, choosing multiple scales, for example, pyramidal approaches [44], can be implemented when dealing with larger images and a finer resolution.

Fig. 6 shows the difference in the feature maps with and without the attention mechanism. Adding the attention mechanism yields much sharper and more focused feature maps. This is because, as presented in [37], the attention mechanism calculates attention coefficients that are multiplied during the concatenation in the expanding path of the architecture (decoder), identifying salient image regions, and prune feature responses to preserve only the relevant activations.

Finally, we present some limitations in the dataset that significantly affect the model's performance, and in general, all remote sensing-based approaches. Fig. 7 shows some problematic scenes with the dataset we used. A scene can be ambiguous, for example, if there are straight lines and the template can match all the positions along the line. In Fig. 7, first column, we notice that the regular patterns of the fields in the SAR image can be matched in different locations. Another example is when the scenes are unclear, such as in the middle of the ocean (7, second column). Finally, there might be temporal differences between optical and SAR acquisition times. Even if the difference is small, this can be significant if moving objects are in the scene, as are the boats in Fig. 7, third column.

Finally, when we plot the distribution of the error (see Fig. 8), we notice a large peak around the 0- and 1-pixel error with few outliers with a very high pixel error. These outliers are due to difficult scenes present in the dataset, as described in Fig. 7.

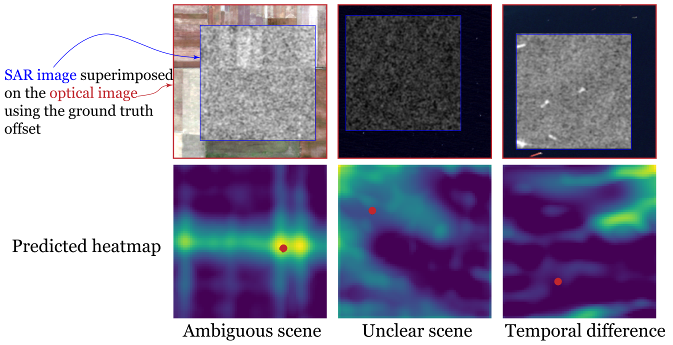

Fig. 7. Practical limitations in the dataset that affects the model performances. A scene can be ambiguous, not clear scenes (for example, in the middle of the ocean), or there might be temporal differences between the acquisition times of optical and SAR.

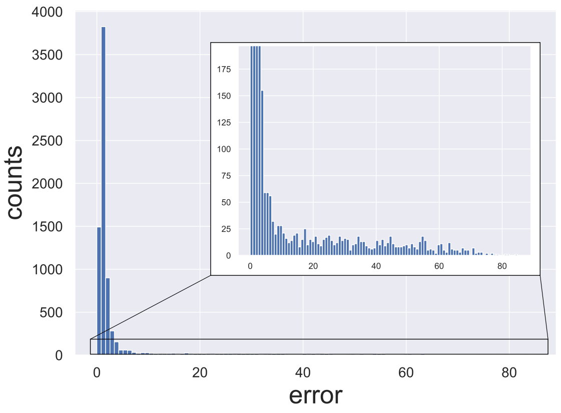

Fig. 8. Error distribution for the SAR-optical matching using our dataset and the proposed model.

Our approach falls in the category of template matching. Therefore, it does not work well if the images are warped (i.e., involving a nonaffine transformation) one to the other or if significant rotations are involved. Future works are toward addressing these challenges. We clarified this better in the results section and in the conclusions.

IV. Conclusion

In this article, we propose an SAR-optical image matching method to increase the matching accuracy and precision at the pixel-level. We extend the classical U-Net with attention mechanisms to improve the feature extraction capabilities of the encoded–decoder architecture. We incorporate a multiscale strategy to produce feature maps from original and downscaled imagery to increase robustness. In addition, we propose a loss function consisting of the combination of cross-entropy and a contrastive loss, tailored particularly to the SAR-optical matching problem. Experiments show that our method outperforms other state-of-the-art methods while still being computationally efficient. Future works are toward exploring the potential of unsupervised or semisupervised methods in SAR-optical matching to overcome the inherent shortcomings of relying on large datasets. Also, another direction is to research how to make the network more robust to scales and rotations.

Acknowledgment

The authors would like to thank Andreas Hay Kaljord and Charlotte Bishop from Kongsberg Satellite Services AS (KSAT), Norway, for their insight and discussion.

References

[1] P. Liu, "A survey of remote-sensing Big Data," *Front. Environ. Sci.*, vol. 3, 2015, Art. no. 00045. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fenvs.2015.00045

[2] L. B. Lentile et al., "Remote sensing techniques to assess active fire characteristics and post-fire effects," *Int. J. Wildland Fire*, vol. 15, pp. 319–345, 2006.

[3] M. Gazzea et al., "Automated satellite-based assessment of hurricane impacts on roadways," *IEEE Trans. Ind. Inform.*, vol. 18, no. 3, pp. 2110–2119, Mar. 2022.

[4] T. Lopez, A. Al Bitar, S. Biancamaria, A. Güntner, and A. Jäggi, "On the use of satellite remote sensing to detect floods and droughts at large scales," *Surv. Geophys.*, vol. 41, no. 6, pp. 1461–1487, Nov. 2020, doi: 10.1007/s10712-020-09618-0.

[5] E. Babaeian, M. Sadeghi, S. B. Jones, C. Montzka, H. Vereecken, and M. Tuller, "Ground, proximal, and satellite remote sensing of soil moisture," *Rev. Geophys.*, vol. 57, no. 2, pp. 530–616, 2019, doi: 10.1029/2018RG000618.

[6] Y. Xie, Z. Sha, and M. Yu, "Remote sensing imagery in vegetation mapping: A review," *J. Plant Ecol.*, vol. 1, no. 1, pp. 9–23, Mar. 2008, doi: 10.1093/jpe/rtm005.

[7] M. Gazzea, L. M. Kristensen, F. Pirotti, E. E. Ozguven, and R. Arghandeh, "Tree species classification using high-resolution satellite imagery and weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4414311.

[8] M. Gazzea, M. Pacevicius, D. O. Dammann, A. Sapronova, T. M. Lunde, and R. Arghandeh, "Automated power lines vegetation monitoring using high-resolution satellite imagery," *IEEE Trans. Power Del.*, vol. 37, no. 1, pp. 308–316, Feb. 2022.

[9] Y. Gao, M. Skutsch, J. Paneque-Gálvez, and A. Ghilardi, "Remote sensing of forest degradation: A review," *Environ. Res. Lett.*, vol. 15, no. 10, Sep. 2020, Art. no. 103001, doi: 10.1088/1748-9326/abaad7.

[10] J. Fishman et al., "Remote sensing of tropospheric pollution from space," *Bull. Amer. Meteorological Soc.*, vol. 89, no. 6, pp. 805–822, 2008. [Online]. Available: https://journals.ametsoc.org/view/journals/bams/89/6/2008bams2526_1.xml

[11] S. Dewitte, J. P. Cornelis, R. Müller, and A. Munteanu, "Artificial intelligence revolutionises weather forecast, climate monitoring and decadal prediction," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3209. [Online]. Available: https://www.mdpi.com/2072-4292/13/16/3209

[12] L.-L. Fu, T. Lee, W. T. Liu, and R. Kwok, "50 years of satellite remote sensing of the ocean," *Meteorological Monographs*, vol. 59, pp. 5.1–5.46, 2019. [Online]. Available: https://journals.ametsoc.org/view/journals/amsm/59/1/amsmonographs-d-18-0010.1.xml

[13] N. Kadhim, M. Mourshed, and M. Bray, "Advances in remote sensing applications for urban sustainability," *Euro-Mediterranean J. Environ. Integration*, vol. 1, no. 1, Oct. 2016, Art. no. 7, doi: 10.1007/s41207-016-0007-4.

[14] O. Dubovik et al., "Grand challenges in satellite remote sensing," *Front. Remote Sens.*, vol. 2, 2021, Art. no. 619818. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frsen.2021.619818

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94.

[16] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, "SAR-SIFT: A SIFT-like algorithm for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 453–466, Jan. 2015.

[17] Y. Xiang, F. Wang, and H. You, "OS-SIFT: A robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3078–3090, Jun. 2018.

[18] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2020.

[19] A. Cole-Rhodes, K. Johnson, J. LeMoigne, and I. Zavorin, "Multiresolution registration of remote sensing imagery by optimization of mutual information using a stochastic gradient," *IEEE Trans. Image Process.*, vol. 12, no. 12, pp. 1495–1511, Dec. 2003.

[20] J. Inglada and A. Giros, "On the possibility of automatic multisensor image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 10, pp. 2104–2120, Oct. 2004.

[21] J. Shermeyer et al., "SpaceNet 6: Multi-sensor all weather mapping dataset," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 768–777, 2020.

[22] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The SEN1-2 dataset for deep learning in SAR-optical data fusion," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. IV-1, pp. 141–146, 2018. [Online]. Available: https://doi.org/10.5194/isprs-annals-IV-1-141-2018

[23] H. Zhang et al., "Registration of multimodal remote sensing image based on deep fully convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 3028–3042, Aug. 2019.

[24] N. Merkle, W. Luo, S. Auer, R. Müller, and R. Urtasun, "Exploiting deep matching and SAR data for the geo-localization accuracy improvement of optical satellite images," *Remote Sens.*, vol. 9, no. 6, 2017, Art. no. 586. [Online]. Available: https://www.mdpi.com/2072-4292/9/6/586

[25] L. H. Hughes, D. Marcos, S. Lobry, D. Tuia, and M. Schmitt, "A deep learning framework for matching of SAR and optical imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 169, pp. 166–179, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271620302598

[26] L. Zhou, Y. Ye, T. Tang, K. Nan, and Y. Qin, "Robust matching for SAR and optical images using multiscale convolutional gradient features," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4017605.

[27] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, and Q. Zhu, "Fast and robust matching for multimodal remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9059–9070, Nov. 2019.

[28] H. Zhang et al., "Optical and SAR image matching using pixelwise deep dense features," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6000705.

[29] Y. Fang, J. Hu, C. Du, Z. Liu, and L. Zhang, "SAR-optical image matching by integrating Siamese U-Net with FFT correlation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4016505.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland, Springer, 2015, pp. 234–241.

[31] W. Wu, Y. Xian, J. Su, and L. Ren, "A Siamese template matching method for SAR and optical image," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4017905.

[32] O. Sommervold, M. Gazzea, and R. Arghandeh, "A survey on SAR and optical satellite image registration," *Remote Sens.*, vol. 15, no. 3, Feb. 2023, Art. no. 850, doi: 10.3390/rs15030850.

[33] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[34] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.

[35] D. Chicco, "Siamese neural networks: An overview," *Methods Mol. Biol.*, vol. 2190, pp. 73–94, 2021.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[37] J. Schlemper et al., "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, 2018. [Online]. Available: https://doi.org/10.1016/j.media.2019.01.012

[38] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, 2016, Art. no. e3. [Online]. Available: http://distill.pub/2016/deconv-checkerboard

[39] J. Yoo and T. Han, "Fast normalized cross-correlation," *Circuits Syst. Signal Process.*, vol. 28, pp. 819–843, 2009.

[40] J.-S. Lee, "Digital image enhancement and noise filtering by use of local statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-2, no. 2, pp. 165–168, Mar. 1980.

[41] S. van der Walt et al., "Scikit-image: Image processing in python," *PeerJ*, vol. 2, 2014, Art. no. e453.

[42] K. Briechle and U. D. Hanebeck, "Template matching using fast normalized cross correlation," in *Proc. SPIE*, vol. 4387, pp. 95–102, 2001.

[43] L. Li, L. Han, H. Cao, and H. Hu, "Joint self-attention for remote sensing image matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4511105.

[44] L. Tang, W. Tang, X. Qu, Y. Han, W. Wang, and B. Zhao, "A scale-aware pyramid network for multi-scale object detection in SAR images," *Remote Sens.*, vol. 14, no. 4, Feb. 2022, Art. no. 973, doi: 10.3390/rs14040973.

Michele Gazzea (Student Member, IEEE) received the bachelor's degree in information engineering and the master's degree in automation and control engineering, both from the University of Padova, Padua, Italy, in 2014 and 2017, respectively. He is currently working toward the Ph.D. degree with Western Norway University of Applied Sciences, Bergen, Norway.

He was a Research and Development Engineer designing machine learning-based diagnostic techniques on milling and engraving CNC machines. His research interests include data analytics, machine and deep learning, computer vision, and remote sensing applications.

Oscar Sommervold received the bachelor's degree in cognitive science and informatics from the University of Bergen, Bergen, Norway, in 2017, and the master's degree in software engineering from Western Norway University of Applied Sciences, Bergen, Norway, and the University of Bergen, in 2022.

He is currently a Machine Learning/Artificial Intelligence Consultant. His research interests include computer vision and large language models.

Reza Arghandeh (Senior Member, IEEE) received the bachelor's degree in electrical engineering and the master's degree in mechanical engineering from the K. N. Toosi University of Technology, Tehran, Iran, 2005 and 2008, respectively, and the master's degree in industrial engineering and the Ph.D. degree in electrical engineering from Virginia Tech, Blacksburg, VA, USA, in 2013.

He was a Postdoctoral Scholar with EECS Department, University of California, Berkeley, Berkeley, CA, USA, from 2013–2015, and an Assistant Professor with ECE Department, FSU, Tallahassee, FL, USA, from 2015–2018. He is a Full Professor with the Department of Computing, Mathematics, and Physics and Department of Electrical Engineering, Western Norway University of Applied Sciences (HVL), Bergen, Norway. He is a Director of Collaborative Intelligent Infrastructure Lab (CI2), Bergen, Norway. He is also a Lead Data Scientist with StormGeo AS, Bergen, Norway. His research has been supported by U.S. National Science Foundation, U.S. Department of Energy, the European Space Agency, and the European Commission.