FloU-Net: An Optical Flow Network for Multi-modal Self-Supervised Image Registration

Damian Ibañez, *Graduate Student Member, IEEE,* Ruben Fernandez-Beltran, *Senior Member, IEEE,* and Filiberto Pla

***Abstract*—Image registration is an essential task in image processing, where the final objective is to geometrically align two or more images. In remote sensing, this process allows comparing, fusing or analyzing data, specially when multi-modal images are used. In addition, multi-modal image registration becomes fairly challenging when the images have a significant difference in scale and resolution, together with local small image deformations. For this purpose, this paper presents a novel optical flow-based image registration network, named the FloU-Net, which tries to further exploit inter-sensor synergies by means of deep learning. The proposed method is able to extract spatial information from resolution differences and through an U-Net backbone generate an optical flow field estimation to accurately register small local deformations of multi-modal images in a self-supervised fashion. For instance, the registration between Sentinel-2 (S2) and Sentinel-3 (S3) optical data is not trivial, as there are considerable spectral-spatial differences among their sensors. In this case, the higher spatial resolution of S2 result in S2 data being a convenient reference to spatially improve S3 products, as well as those of the forthcoming Fluorescence Explorer (FLEX) mission, since image registration is the initial requirement to obtain higher data processing level products. To validate our method, we compare the proposed FloU-Net with other state-of-the-art techniques using 21 coupled S2/S3 optical images from different locations of interest across Europe. The comparison is performed through different performance measures. Results show that proposed FloU-Net can outperform the compared methods. The code and dataset are available in https://github.com/ibanezfd/FloU-Net.**

***Index Terms*—Image Registration, Convolutional Neural Networks, Inter-sensor, Multi-modal, Multi-spectral, Sentinel-2-3.**

I. Introduction

FOR the last decades, there have emerged new remote sensing missions with a wide variety of instruments to obtain Earth Observation (EO) data. When designing new missions, one of the main limitations of hyper-spectral or multi-spectral sensors is their spatial resolution when a high spectral resolution is also a requirement. EO programmes as the European Spacial Agency (ESA)'s Copernicus [1] already have a number of satellites with different instruments designed to accomplish different tasks. This increase of different source data [2] has an according increase of multi-modal possibilities such in image fusion [3], product mosaicing [4] or spatial-spectral super-resolution to improve land-cover mapping [5] among other applications. However, inter-sensor products can certainly be acquired from different angles, positions, time periods, resolutions or even imaging modalities, affecting the successful exploitation of such data.

To use multi-modal techniques in an accurate manner, image registration is mandatory [6]. This process aims to obtain the spatial correspondence between two images which represent a common area. Roughly speaking, every image registration process [7] follows the same steps to align the input image (slave) to its corresponding reference image (master): (1) the image pre-processing and the selection of search space; (2) the feature detection if the pixel intensity value is not used; (3) the matching of these features between the slave and master images; (4) the search strategy where a correspondence metric is maximized or minimized; (5) the acquisition of the transformation parameters and (6) the wrapping of the slave image with the transformation values to the master image.

In the literature, many different approaches [8] have been proposed to solve the spatial correspondence between images. These image registration techniques can be classified by the similarity of their correspondence metrics and the features they use to register images. Other details of the methods can differ as well, such as the transformation model (rigid, affine, deformable...), the domain (local or global) or even the image modality (optical, synthetic-aperture radar, laser imaging detection and ranging...). Among the different similarity metrics, the cross-correlation is the basic statistic technique to evaluate the similarity between to images, for example in [9] Sarvaiya *et al.* used the normalized cross-correlation to template matching medical images. Related to correlation methods, the maximization of mutual information (MI) is a common metric in multi-modal image registration. MI indicates the statistic dependency degree between images, estimating the joint probability of the pixels' intensity. For example, ELASTIX [10] uses a multilevel pyramidal registration to obtain the affine transformation parameters maximizing the MI. Nevertheless, the most popular methods used in remote sensing are feature-based methods. These methods match edges, lines, local maximum or minimum points or even regions between the slave and master images. From the local feature methods proposed in remote sensing, the Scale-Invariant Feature Transform (SIFT) is one of the most used [11]. Learning methods which are able to perform the complete registration process or extract features from decision trees, convolutional neural networks (CNN) or other regression algorithms are gaining popularity as well, inspired by the recent success of CNN and deep learning in other fields. Yang *et al.* designed in [12] a CNN to obtain robust features to register satellite and UAV (Unmanned Aerial Vehicle) images

This work was supported by Ministerio de Ciencia e Innovación (PID2021-128794OB-I00) and Generalitat Valenciana (ACIF/2021/215).

D. Ibañez and F. Pla are with the Institute of New Imaging Technologies, University Jaume I, E-12071 Castellón de la Plana, Spain. (e-mail: ibanezd@uji.es; pla@uji.es).

R. Fernandez-Beltran is with the Department of Computer Science and Systems, University of Murcia, 30100 Murcia, Spain. (e-mail: rufernan@um.es).

using a pre-trained VGG16 to extract deep features and then an expectation maximization registration. For images with small deformations, optical flow (OF) methods are commonly used. For example, GEFOLKI [13] uses the Lucas-Kanade local OF algorithm with a multi-scale iterative strategy.

Despite the results achieved by these and many other image registration methods, within the ESA's Copernicus missions, Sentinel-2 (S2) [14] and Sentinel-3 (S3) [15] products registration presents a series of complexities which are still a challenge. Even though both missions belong to the Copernicus programme, their objectives and therefore the instruments carried by each constellation are different enough to make their product registration a complex task. Specifically, the S2 mission formed by two satellites (S2A and S2B) has as main objectives agricultural management and land cover classification. Both S2A and S2B carry the Muti-Spectral Instrument (MSI) sensor. This sensor obtains high-resolution multi-spectral images in 13 bands, from 443 to 2190 nm with a spatial resolution of 10, 20 and 60 meters per pixel (mpp). S3 mission is also composed by two satellites (S3A and S3B), and it has the objective of seas and oceans supervision, seawater quality evaluation, weather forecasting and pollution monitoring. The S3 satellites carry the Ocean and Land Color Instrument (OLCI). This sensor provides multi-spectral images with 21 bands and a spectral range from 400 to 1020 nm with a spatial resolution of 300 mpp. As previously stated, the image registration between S2 and S3 products has significant challenges: the spectral range and bands difference, the resolution difference, sensor local deformations and other multi-modal complications. But those differences are also the reason why S2-S3 synergies are attractive. The large resolution difference between S2-S3 implies that S2 products have lower geo-location errors than S3 images, as the error decreases with the increase of spatial resolution of the sensor. Thus, S2 data becomes a valid ground-truth reference to register S3 images.

Given the aforementioned problems of the S2-S3 product registration, this paper proposes a new image registration network that addresses the four main challenges within the S2-S3 image registration. First, the spatial resolution difference between S2 and S3. Coarse-to-fine registration is a complexity which many image registration methods have to deal with, yet state-of-the-art coarse-to-fine methods are not able to process a difference of tens of times in pixel size between images. Second, the multi-modal spectral differences. Not only the different number and width of spectral bands, but the scale of intensity and contrast between S2 and S3 products. Third, the sensors usually generate local deformations, which makes global geometric transformations unable to perfectly adapt S3 images to S2 information. Fourth, the unavailability of ground-truth data for S2 and S3 operational products, resulting in a limitation in learning and training.

To cope with these complexities, we present a novel deep learning-based optical flow image registration network: the FloU-Net. In contrast to other models constrained to different remote sensing data sources with small spatial changes and affine deformations [16], FloU-Net takes advantage of the so-called U-Net architecture, used in other application domains [17], to provide a new solution for successfully exploiting inter-sensor spatial context with important resolution differences and multi-spectral modalities. The designed model is able to process S2 and S3 image products, extract deep features taking advantage of the U-Net potential to solve multi-modal and local discrepancies, extract the OF field from them, and deform the S3 slave image to adjust it to the S2 master image. Additionally, the loss function used for the training integrates the use of a Local Normalized Cross Correlation (LNCC) metric between the images and a regularization of the OF field to avoid over-deformation, preserving the image borders and maximizing the similarity. Below the rest of the work is organized in three more sections including methodology Section II, experimentation Section III and conclusions Section IV.

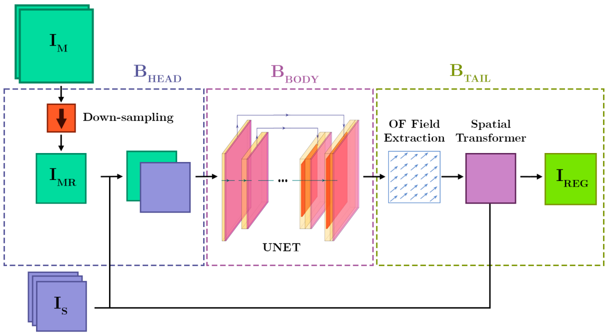

Fig. 1. Diagram of the proposed FloU-Net image registration architecture.

II. Methodology

Let $I_S \in \mathbb{R}^{(X_S \times Y_S \times B_S)}$ be the multi-spectral slave image with B_S spectral bands and $(X_S \times Y_S)$ spatial size. Let $I_M \in \mathbb{R}^{(X_M \times Y_M \times B_M)}$ be the master image with a size of $(X_M \times Y_M)$ and a B_M number of spectral bands. Both I_S and I_M represent the same Earth area surface with a R ratio between their spatial resolutions, $R = X_M/X_S$ or $R = Y_M/Y_S$. Then, the FloU-Net model has the objective of spatial matching the I_S image only using information from I_M to approximate it as much as possible to the corresponding spatial ground-truth image, $I_{GT} \in \mathbb{R}^{(X_S\times\ Y_S\times\ B_S)}$ obtaining the registered image $I_{REG} \in \mathbb{R}^{(X_S\times\ Y_S\times\ B_S)}$. This process can be summarized in the expression $f(I_S, I_M) = I_{REG}$. Therefore, ground-truth information is not used in the model training, following a self-supervised learning scheme thanks to the spatial resolution differences between I_S and I_M.

A. FloU-Net Architecture

To improve modularity and ease the whole process, the FloU-Net model has been divided in three blocks: the head block (B_{HEAD}), whose purpose is to perform the necessary processing of the input images and to extract deep features for the subsequent block, i.e., the body block (B_{BODY}). In the last block, the OF field is obtained from the deep features and the I_S image is wrapped using the OF field into the I_{REG} output. The complete architecture scheme is shown in Figure 1. To balance the deformation from the OF field and the spatial consistency, a composed loss function was designed.

The first block of the FloU-Net, B_{HEAD}, has three main phases. First, a single B_S and B_M band from each multi-spectral product is selected. Then, the input images become $I_S \in \mathbb{R}^{(X_S \times Y_S)}$ and $I_M \in \mathbb{R}^{(X_M \times Y_M)}$. The band selection is followed by a down-sampling process, done over the I_M image to match the spatial resolution of I_S. The down-sampling of I_M is implemented through a 2D-convolution (2DC) with a kernel of size $K = 1 \times 1$ and a stride $S = R$. This configuration is intended to efficiently sub-sample the high-resolution image domain while simulating a sort of spectral PSF (Point Spread Function) to better alleviate inter-sensor multi-modal discrepancies compared to pooling or other straightforward down-sampling operations. After this 2DC layer, we obtain the image $I_{MR} \in \mathbb{R}^{(X_S \times Y_S)}$, which is concatenated to the I_S image and forwarded to the next block.

To obtain the deep features to extract the OF field afterwards, the B_{BODY} is based on a standard U-Net backbone. The U-Net backbone architecture used by the FloU-Net is depicted in Figure 2, composed by two types of encoder layers and three types of decoder layers, with a total of five layer of each kind. On the one hand, the encoder layers are composed by a 2DC layer and a leaky ReLU layer. On the other hand, all the decoder layers are composed by an up-sampling layer, a 2DC and a leaky ReLU layer except for the last one which does not contain an up-sampling layer. The difference between the $E1$ and $E2$ encoder layers lies in the number of filters, as the $E1$ layer has a number of filters $N = 16$ and the $E2$ layers contain $N = 32$ filters to further increase the features representation maps. The kernel size and stride are the same for the both encoding layers, $K = 3 \in 3$ and $S = 2$ to reduce the spatial information while condensing layers' depth information. Meanwhile, the $D1$, $D2$ and $D3$ decoder layers have a similar approach, the three of them having a kernel size of $K = 3 \times 3$, a stride of $S = 1$ and an up-sampling ratio of 2, but with a different number of filters. The $D1$ layers have 32 filters, while $D2$ and $D3$ have 16 to gradually reduce the feature maps depth while increasing the spatial resolution. The difference between the layers $D2$ and $D3$ is the up-sampling layer, which does not exist in the $D3$ layer.

In the last stage of the FloU-Net model, B_{TAIL} first extracts the OF field from the information obtained through the B_{BODY}. This process is done using an extra 2DC layer, in this case of $N = 2$ filters with a kernel size $K = 3 \times 3$ and stride $S = 1$. The objective of this layer is extracting the displacements in each direction. As only small deformations in the OF field are expected, the weights and the bias values are initialized with low values, specifically 10^{-5} for the weights and 0 for the bias. Finally the OF field Φ obtained in the 2DC is used in the spatial transformer [18] with the original I_S image to obtain the registered image I_{REG}. In the spatial transformer, the new sub-pixel location l for each original pixel location p is calculated as $l = p + \Phi(p)$. For this purpose, a localization network takes the OF, and using hidden layers outputs the predicted spatial transformation that should be applied to the image. Then, the image and a sampling grid generated from the predicted transformation are taken as inputs to the sampler, creating the output registered image. Since pixel locations can only be integers, a bilinear interpolation of the new location l is done using its eight pixel neighborhood.

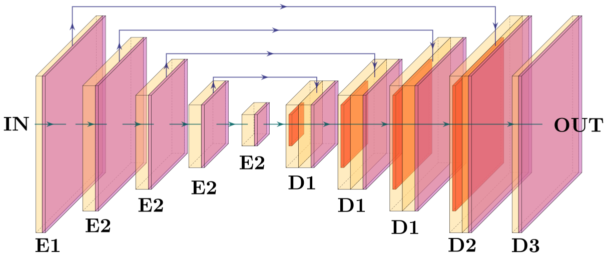

Fig. 2. Diagram of the U-net backbone architecture used in FloU-Net.

In order to train the layers of the different blocks, a composed multi-modal loss function has been used, shown in (Equation (1). The first component of this loss is the LNCC loss, (Equation (2)). This loss allows to asses the multi-modal correlation between the two images. The considered LNCC expression is shown in Equation (3), where $\mathcal{Z}(p)$ is the $n \times n$ neighborhood of the pixel p, $\hat{I}_1$ and $\hat{I}_2$ denote images with local mean intensities subtracted out, and Ω is the image pixel grid in (X, Y) spatial axes. The second loss component is weighted by the hyper-parameter α is a $L2$ regularized Gradient of the Φ OF field to avoid irregular deformations and inconsistencies in the local pixel displacements.

III. Experiments

A. Dataset

To validate the proposed FloU-Net, a dataset of 21 coupled pairs of S2/S3 multi-spectral images with less than one day difference from the year 2020 of different locations in Europe has been used. These images contain a variety of locations: cities, natural parks, coasts and mixtures of them to better show the generalization capabilities of the FloU-Net. Specifically, there are images from Spain, Portugal, France, Italy, Germany, Czech Republic, Poland and Denmark. To generate the dataset, first the study areas were located, and only S2 MSI Level 2 products containing those locations with less than a 1% of cloud coverage were selected. After obtaining the S2 products, the correspondent S3 OLCI Level 1 products with one day of maximum difference were selected and downloaded. Then, the S2 products were re-sampled to a 20 m spatial resolution to obtain a multi-spectral image of $(5490 \times 5490 \times 12)$ size. Meanwhile, the S3 data were Rayleigh corrected and cropped using only the overlapping area with the S2 image, generating images of size $(366 \times 366 \times 21)$. Once the cropping and corrections were finished both S2 and S3 data were normalized.

B. Experimental settings

In order to test the performance of the FloU-Net, it has been compared to other state-of-the-art image registration methods, used in remote sensing and other application fields: SIFT[11], ELASTIX[10], a CNN[12] and GEFOLKI[13]. In this work, we assume that spatial displacements among intra-sensor spectral bands are negligible due to the important resolution differences between S2 and S3 optical sensors. Hence, all the

$$\mathcal{L}(\mathbf{I}_{\mathrm{REG}}, \mathbf{I}_{\mathrm{MR}}, \Phi) = \mathcal{L}_{\mathrm{LNCC}}(\mathbf{I}_{\mathrm{REG}}, \mathbf{I}_{\mathrm{MR}}) + \alpha \mathcal{L}_{\mathrm{GRAD}}(\Phi) \tag{1}$$

$$\mathcal{L}_{\mathrm{LNCC}}(I_1, I_2) = -\sum_{p \in \Omega} \frac{\left(\sum\limits_{q_i \in \mathcal{Z}(p)} (I_1(q_i) - \hat{I}_1(q))(I_2(q_i) - \hat{I}_2(q)) \right)^2}{\left(\sum\limits_{q_i \in \mathcal{Z}(p)} (I_1(q_i) - \hat{I}_1(q)) \right) \left(\sum\limits_{q_i \in \mathcal{Z}(p)} (I_2(q_i) - \hat{I}_2(q)) \right)} \tag{2}$$

$$\mathcal{L}_{GRAD}(\Phi) = \sum_{p \in \Omega} ||\nabla \Phi(p)||^2 \tag{3}$$

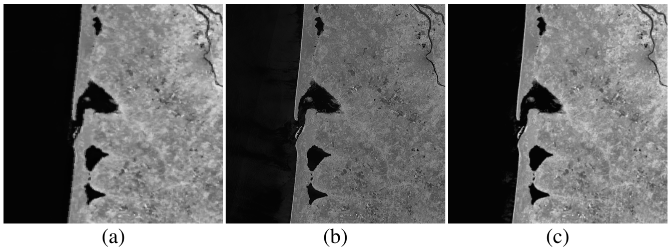
(a) (b) (c)

Fig. 3. Dataset example of Burdeos, France (45.1467528,-1.7279663), (45.1226718,-0.33236694) longitude-latitude: (a) S3 Oa17, (b) S2 B8a, (c) Synthetic ground-truth.

considered methods are used to register only one spectral band of multi-spectral images and then apply the deformation to the remaining bands. Specifically, the Oa17 spectral band of S3 and the B8a band of S2 were chosen as both of them are centered in 865 nm and have a band width of 20 nm. For state-of-the-art methods, the master image I_{S2} B8a band with the complete resolution of (5490×5490) was reduced to the size of I_{S3} simulating the PSF effect using a Lanczos interpolation, obtaining I_{S2R} with a size of (366×366). For the FloU-Net, the full I_{S2} (5490×5490) resolution was used. To asses the results, as no ground-truth exists in this study case, the histogram of I_{S2R} was matched to the S3 Oa17 spectral information using a histogram-based intensity function and histogram equalization, simulating a ground-truth reference image I_{GT} of size (366×366), containing the S3 spectral signature and the low geo-location error of S2. For the proposed FloU-Net the ADAM optimizer with a standard learning rate of 10^{-4}, a batch size of 1 to process each image separately and 8000 iterations were used. This is the minimum number of iterations needed for the model, in some cases the results could improve with more epochs. Also an $\alpha = 0.5$ in the L2 regularized Gradient loss, to give more weight to the multi-modal loss. This experimentation has been done with Python 3.6 PyTorch on a Ubuntu 20.04 x64 machine with Intel(R) Core(TM) i7.6859K, NVIDIA GeForce 2080Ti and 64 Gb of RAM.

C. Results

In Table I, the quantitative results of the conducted experimentation are shown with seven metrics: root mean square error (RMSE), normalized root mean square error (NRMSE), peak signal-to-noise ratio (PSNR) in dB, structural similarity index measure (SSIM), relative dimensionless global error (ERGAS), universal image quality index (Q) and mutual information (MI). The methods are presented in rows and the metrics in columns in the table, being each result the mean of the 21 image pairs of the dataset with their corresponding standard deviation. Note that the images used were represented in unsigned integers of 16 bits, so the RMSE values are according to this dynamic range. The two first metrics RMSE and NRMSE show the ability of each method to reduce the intensity differences between the ground-truth reference and the registered image, the PSNR the noise reduction, MI and SSIM for the similarity between the images and the ERGAS and Q index for image distortions. The optimal values for each of the metrics are the following: RMSE(0), NRMSE(0), PSNR($+\infty$), SSIM(1), ERGAS(0), Q(1), MI($+\infty$).

As can be observed in Table I, the FloU-Net outperforms every other image registration method tested for multi-modal image registration between S2 and S3 images. Followed by GEFOLKI, and SIFT having the worst performance. GEFOLKI's high performance can be explained due to being the only one designed to deal with small local deformations through optical flow. ELASTIX was able to obtain competitive MI and SSIM results, but the affine transformation with B-spline was not able to solve properly the local deformations, which also happened to SIFT's affine transformation. On the other hand, the CNN performed better with RMSE and NRMSE metrics using a thin plate spline transformation, but lacked in similarity. Nonetheless, the proposed FloU-Net is able to obtain better results in all the considered metrics with the smallest deviation as well, except for the PSNR. As an example of qualitative results, Figure 3 shows the results of the four best methods (ELASTIX, CNN, GEFOLKI and FloU-Net) for the images in Figure 4. These images contain two areas that have been magnified to appreciate the details of the checkerboard image comparing each image registered (lighter) and the corresponding ground-truth reference (darker). In the first magnified area, the coastline and a near inland water body are shown. Note how the proposed method is able to better correct the coastline and define the border of the water body reducing the noise and with sharper borders. The second magnified area contains the bifurcation of the Garona and Dordoña rivers. Again, the difference with the noise reduction and the sharpness is noticeable between the FloU-Net and the other methods, in the riverbanks and in other small details.

As it has been shown with quantitative and qualitative

Methods	RMSE	NRMSE	PSNR	SSIM	ERGAS	Q	MI
SIFT [11]	6531±5413	0.12±0.12	21.9±5.26	0.68±0.19	1.33±1.20	0.62±0.21	1.22±0.59
ELASTIX[10]	4202±3973	0.07±0.11	25.8±5.06	0.79±0.09	0.95±1.05	0.74±0.13	1.41±0.32
CNN [12]	3939±1318	0.07±0.02	24.9±3.26	0.61±0.09	0.84±0.21	0.56±0.11	0.98±0.35
GEFOLKI[13]	2808±944	0.05±0.02	27.9±3.17	0.80±0.06	0.58±0.13	0.75±0.11	1.41±0.33
FloU-NET	**2500±876**	**0.04±0.01**	**28.9±3.46**	**0.80±0.06**	**0.52±0.10**	**0.75±0.11**	**1.42±0.33**

TABLE I
QUANTITATIVE ASSESSMENT FOR RMSE, NRMSE, PSNR, SSIM, ERGAS, Q AND MI METRICS.

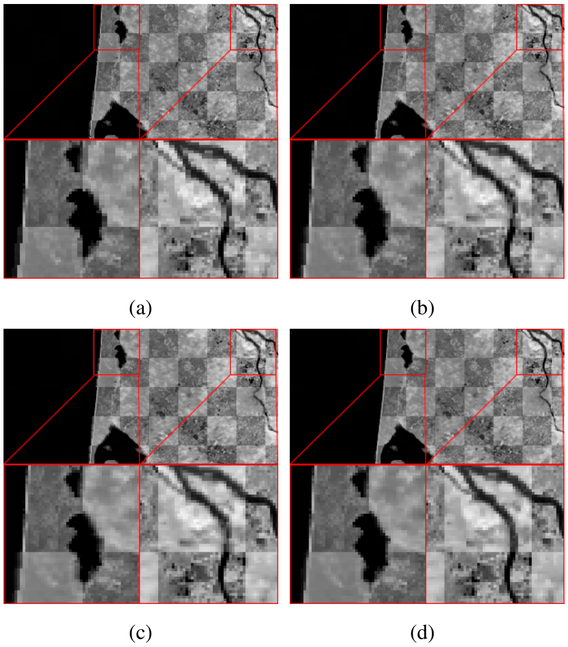
(a) (b) (c) (d)

Fig. 4. Registration qualitative assessment, comparison between ground-truth and: (a) CNN, (b) ELASTIX, (c) GEFOLKI, (d) FloU-Net.

results, the proposed FloU-Net has three main advantages: (i) the exploitation of inter-sensor resolution differences, (ii) the capacity to register multi-modal images, (iii) the ability to solve local deformations. The proposed FloU-Net is able to use the complete spatial information contained in the full resolution S2 images (15 times bigger than S3 data), allowing the network to adjust with sub-pixel accuracy. This comes at the cost of a higher computational cost compared with the other methods. However, FloU-Net is the only multi-modal registration method able to deal with such resolution differences. In addition, the multi-modal loss is able to optimize the registration in spite of the intensity and contrast differences between the spectral information of S2 and S3. Finally, the OF field solution makes possible to exploit the sub-pixel displacements obtained coping with the resolution difference between the images. Thanks to these advantages, we expect the proposed FloU-Net to perform in a similar fashion with other multi-modal registration tasks, even in higher resolutions.

IV. CONCLUSIONS

In this paper a new image registration method for multi-modal images with resolution differences and local deformations is proposed, the FloU-Net. This self-supervised CNN model is able to exploit the significant image resolution differences, adapt to the multi-modal discrepancies and solve local displacements directly from the full resolution original data. An experimental study using 21 couples of different Europe locations comparing the FloU-Net to other state-of-the-art methods validates its performance and results. The main conclusion of this work is the relevance of taking advantage of the complete spatial resolution in remote sensing registration and deep learning techniques for achieving registration accuracy in the described conditions. Future plans are directed to further exploit spectral information while extending the experiments to other forthcoming tandem platforms like FLEX.

REFERENCES

[1] J. Aschbacher and M. Milagro-Pérez, "The european earth monitoring (gmes) programme: Status and perspectives," *Remote Sens. Environ.*, vol. 120, pp. 3–8, 2012.

[2] M. Chi, A. Plaza, J. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," *Proc. IEEE*, vol. 104, no. 11, pp. 2207–2219, 2016.

[3] R. Fernandez-Beltran, J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "Remote sensing image fusion using hierarchical multimodal probabilistic latent semantic analysis," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 11, no. 12, pp. 4982–4993, 2018.

[4] D. Ibañez, R. Fernandez-Beltran, J. M. Sotoca, R. A. Mollineda, J. Moreno, and F. Pla, "Multitemporal mosaicing for sentinel-3/flex derived level-2 product composites," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 5439–5454, 2020.

[5] S. S. Rwanga, J. M. Ndambuki *et al.*, "Accuracy assessment of land use/land cover classification using remote sensing and gis," *International Journal of Geosciences*, vol. 8, no. 04, p. 611, 2017.

[6] X. Dai and S. Khorram, "The effects of image misregistration on the accuracy of remotely sensed change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 5, pp. 1566–1577, 1998.

[7] J. Le Moigne, N. Netanyahu, and R. Eastman, *Image registration for remote sensing*. Cambridge University Press, 2011.

[8] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image Vision Comput.*, vol. 21, no. 11, pp. 977–1000, 2003.

[9] J. Sarvaiya, S. Patnaik, and S. Bombaywala, "Image registration by template matching using normalized cross-correlation," in *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*, 2009, pp. 819–822.

[10] S. Klein, M. Staring, K. Murphy *et al.*, "Elastix: a toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[11] W. Ma, Z. Wen, Y. Wu *et al.*, "Remote sensing image registration with modified sift and enhanced feature matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 1, pp. 3–7, 2017.

[12] Z. Yang, T. Dan, and Y. Yang, "Multi-temporal remote sensing image registration using deep convolutional features," *Ieee Access*, vol. 6, pp. 38 544–38 555, 2018.

[13] G. Brigot, E. Colin-Koeniguer, A. Plyer, and F. Janez, "Adaptation and evaluation of an optical flow method applied to coregistration of forest remote sensing images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 9, no. 7, pp. 2923–2939, 2016.

[14] M. Drusch *et al.*, "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services"," *Remote Sens. Environ.*, vol. 120, no. Supplement C, pp. 25 – 36, 2012, the Sentinel Missions - New Opportunities for Science.

[15] C. Donlon *et al.*, "The Global Monitoring for Environment and Security (GMES) Sentinel-3 mission"," *Remote Sens. Environ.*, vol. 120, no. Supplement C, pp. 37 – 57, 2012.

[16] Y. Ye, T. Tang, B. Zhu, C. Yang, B. Li, and S. Hao, "A multiscale framework with unsupervised learning for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[17] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, “An unsupervised learning model for deformable medical image registration,” in *CVPR*, 2018, pp. 9252–9260.

[18] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, “Spatial transformer networks,” *Adv. in neur. inf. proc. sys.*, vol. 28, pp. 2017–2025, 2015.