Unsupervised historical map registration by a deformation neural network

Unsupervised Historical Map Registration by a Deformation Neural Network

Sidi Wu
sidiwu@ethz.ch
Institute of Cartography and Geoinformation, ETH Zurich
Switzerland

Raimund Schnürer
schnuerer@ethz.ch
Institute of Cartography and Geoinformation, ETH Zurich
Switzerland

Magnus Heitzler
hmagnus@ethz.ch
Institute of Cartography and Geoinformation, ETH Zurich
Switzerland

Lorenz Hurni
lhurni@ethz.ch
Institute of Cartography and Geoinformation, ETH Zurich
Switzerland

ABSTRACT

Image registration that aligns multi-temporal or multi-source images is vital for tasks like change detection and image fusion. Thanks to the advance and large-scale practice of modern surveying methods, multi-temporal historical maps can be unlocked and combined to trace object changes in the past, potentially supporting research in environmental science, ecology and urban planning, etc. Even when maps are geo-referenced, the contained geographical features can be misaligned due to surveying, painting, map generalization, and production bias. In our work, we adapt an end-to-end unsupervised deformation network that couples rigid and non-rigid transformations to align scanned historical map sheets at different time stamps. To the best of our knowledge, we are the first to use unsupervised deep learning to register map images. We address the sparsity of map features by introducing a loss based on distance fields. When aligning the displaced landmark locations by our proposed method, the results are promising both quantitatively and qualitatively. The generated smooth deformation grid can be applied to vector features directly to align them from the source map sheet to the target map sheet.

CCS CONCEPTS

• **Applied computing** → **Cartography**; **Graphics recognition and interpretation**; • **Computing methodologies** → *Matching*.

KEYWORDS

GIS, image registration, historical maps, deep learning, unsupervised neural networks

ACM Reference Format:
Sidi Wu, Raimund Schnürer, Magnus Heitzler, and Lorenz Hurni. 2022. Unsupervised Historical Map Registration by a Deformation Neural Network. In *The 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI '22) , November 1, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3557918.3565871

GeoAI '22, November 1, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9532-8/22/11.
https://doi.org/10.1145/3557918.3565871

1 INTRODUCTION

Image registration is the process of aligning different sources of images that depict the same scene so that the related information can be combined or compared. The images can have different time stamps, modalities, viewing angles, sensors, coordinate systems, location accuracy, and production processes. Image registration can benefit fields like medical imaging in improving the diagnosis accuracy with multiple sources of information [5] or earth observation to allow for change detection [21]. Before the wide use of modern earth observation techniques, historical maps are almost the exclusive resource to depict the Earth's surface. The topographic maps from one or two centuries ago can already be relied on thanks to the advances and large-scale practices of surveying methods. They were usually updated regularly so that maps are often available for the same area at multiple time stamps. Unlocking spatial information at different times enables us to monitor the trajectory of spatio-temporal dynamics of features in the past, which can support studies and analysis in various fields like urban planning, environmental sciences, and ecology. Existing works about geo-referencing historical maps focus on determining their locations by comparing them with other geo-referenced data [14, 15, 1]. However, even after being geo-referenced, features from maps at different time stamps can be spatially misaligned, due to different surveying processes, generalization procedures, production conditions, and map distortions. Therefore, image registration is necessary for letting map features be comparable or combinable.

Generally speaking, image registration can be classified into rigid and deformable methods [26]. Rigid methods usually model transformation with global affine parameters such as rotation and translation, and thus the geometry is transformed uniformly. By contrast, deformable methods estimate non-rigid dense displacement fields, which allows non-homogeneous local geometric transformation. The core procedure of both methods is to find correspondences between the paired images and then estimate the transformation parameters. Common image registration approaches can be categorized into intensity-based methods [10] and feature-based methods that use hand-crafted features [12, 16]. For geospatial images, salient objects like landmarks, triangulation points, road crossings, and street corners are usually manually selected or automatically detected [7, 13] and compared. With the prosperity in deep learning methods, end-to-end neural networks have been used to learn the correspondence between the paired images and to estimate the

transformation parameters simultaneously in a supervised [9, 4] or unsupervised manner [25, 19, 17]. The advantage of applying deep learning methods is obvious: On the one hand, complex multi-level features relevant for the task can be automatically learned without being hand-crafted. On the other hand, the deep-learning-based approach is able to register images in real time at inference, which is computationally more efficient than traditional pairwise optimization that estimates transformation parameters by searching for and aligning pixels with similar intensity/features [8].

For historical maps, automatic image registration is a vital yet rarely tapped research topic. The most related research is the registration of satellite images [11, 24, 20, 23]. [20] adapts the method from [17] to register satellite images with convolutional neural networks (CNNs), where the transformation network is able to learn both affine and deformable parameters separately. Similarly, the registration of multi-temporal historical maps can be regarded as a mixture of rigid (e.g., systematic surveying or production bias) and deformable transformation (e.g., non-uniform generalization or distortion). Following this line of research, we propose a modified end-to-end deformation network. Different from natural images, map features are sparse — the majority of a map is a plain background that cannot give a progressive signal to the network when it searches for image correspondences and regresses for parameters. To solve this problem, we make use of distance fields (DFs) [6], commonly used in computer graphics and vision to map discrete objects into continuous representations [18, 3].

Our main contributions are: 1) adapting an end-to-end unsupervised deformation network for historical map registration, 2) entangling affine and deformable transformations into a single transformation network which outputs a smooth deformation grid that can be directly applied to vector features 3) introducing DF-based loss to project discrete and sparse objects typical for maps into continuous representations. Supplementary materials (code and data samples) can be found here: *https://github.com/sian-wusidi/mapdeform/*.

2 METHODOLOGY

An overview of our method is illustrated in Figure 1. A pair of images, together with their corresponding DF maps, which are generated offline, are input to a deformation neural network. The network entangles rigid and deformable parameters to predict the warping grid for image registration. The source DF is then warped / deformed to align with the target DF. More details will be introduced in the following sections.

2.1 Distance Fields

A distance field (DF) is defined as a spatial field that represents the minimum scalar distance of a point to a shape of surface geometry or edges of features [6]. To represent the 2D geometry of objects in maps, we extract edges using a Canny detector [2], which applies Gaussian smoothing, calculates intensity gradients, thresholds gradients for edge candidates, and suppresses weak edge candidates that are not connected to strong ones. As historical maps inevitably contain noise from the original drawing, aging, and scanning, a Gaussian filter is necessary to smooth images and extract salient edges. We calculate the distance of each pixel to the closest edge and obtain a DF map. Figure 2 illustrates our derived edge map and DF map.

2.2 Transformation Modelling

The goal of image registration is to estimate an optimal transformation that aligns the paired images the best under specific conditions:

$$\mathcal{M}(T, S \circ W) + \mathcal{R}(W) \tag{1}$$

where S denotes a source image and T denotes the target image. W is a dense warping grid that contains the deformed position to be estimated for each pixel on the source image. $\circ$ indicates a sampling operation to synthesize the deformed source image S'. $\mathcal{M}$ quantifies the level of alignment between the deformed source and target image. $\mathcal{R}$ is the regularization item of W. To obtain S' from the warping grid W, we adopt bilinear interpolation as in [9]:

$$S'(p) = S(p) \circ W(p) = \sum_q S(q) \prod_d max(0, 1 - |[W(p)]_d - q_d|) \tag{2}$$

where p and q are pixel locations and $d \in \{x, y\}$ denotes an axis. This way pixels within one pixel's distance to p along both $x-$ and $y-$axis are used for interpolation. As justified before, rather than original RGB images, we deform their DFs.

Instead of predicting a warping grid directly, similar to [20], we adapt the method proposed by [17] to estimate a spatial gradient map ∇W of the warping grid between consecutive pixels along each axis, which is proven to generate smoother deformation than the direct displacement field / warping grid. The warping grid can then be calculated as the integration of ∇W along $x-$ and $y-$axis starting from the top left corner. In discrete cases, the integration can be simply approximated by the cumulative sum. By constraining gradients to be positive, self-crossings can be avoided between the displacements. The amplitude of gradients controls the degree of squeezing and stretching the distance between consecutive pixels. Concretely, two consecutive pixels will move closer, keep distance and move away from each other when the gradient is smaller than 1, equal to 1, and larger than 1, respectively.

As mentioned before, registration between historical maps can be regarded as a mixture of affine and deformable transformations. Since the registered maps have the same scale and orientation (i.e., north-up), we exclude rotation and scaling and only consider translation for affine parameters. Actually, modelling translation separately from deformation as [20] is not necessary: it can be approximated by stretching or squeezing the distance between pixels at the border ($\nabla W \neq 1$) while keeping the original distance between pixels in the middle ($\nabla W = 1$). Through the integration process, the displacement from the border will propagate to and shift other pixels globally while not changing their distance from each other. In this way, huge distortions can occur at the border to align middle features that have big displacements. This can increase the alignment loss significantly. To handle this problem, we ease the loss at sheet borders by cropping them out before calculating the alignment loss.

2.3 Network Architecture

We make use of U-Net that integrates atrous spatial pyramid pooling (ASPP) as [22], which achieves state-of-art accuracy for historical

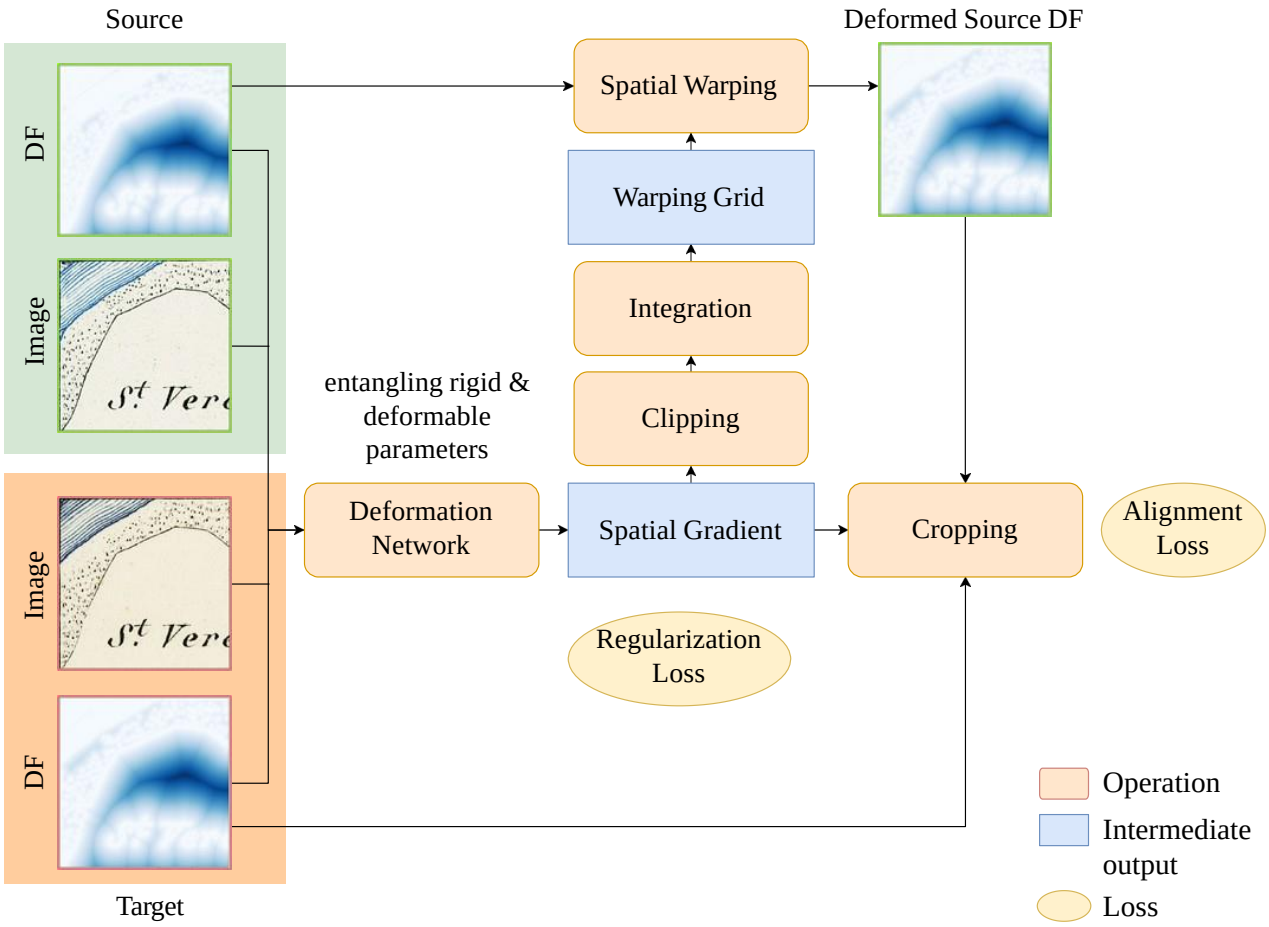

Figure 1: An overview of the proposed methods. The input is the pair of images and their corresponding DF maps, generated offline. A deformation neural network entangles rigid and deformable parameters to predict the warping grid for image registration. The source DF is then warped / deformed to align with the target DF.

(a) (b) (c)

Figure 2: Illustration of an edge map (b) derived from a image (a) using Canny detector and the corresponding DF map (c).

map segmentation both efficiently and effectively. We generate DF maps offline and concatenate the source image, source DF, target image, and target DF together as input to the network. The features extracted by the encoder are passed to an ASPP block to incorporate multi-scale contexts before the decoder. The outputs of our decoder are two spatial gradient maps for both x- and y-axis that have the same spatial dimension as the original images. They are passed to a ReLU layer to have only positive values. We subsequently apply an clipping layer to clip the spatial gradients outside the range (0,4), an integration layer which calculates the warping grid, a bilinear interpolation layer to warp the source image, and a cropping layer to obtain the final deformed image. Our proposed pipeline is indifferent to the network architectures and can be integrated into other algorithms for different applications.

2.4 Loss Functions

The overall loss is defined as:

$$\mathcal{L} = \left\|(\bar{T}_{df} - \bar{S'}_{df}) * (1 - \bar{T}_{df})\right\| + \left\|\frac{\sum^n \nabla W}{n} - 1\right\| + \lambda \left\|\nabla W - I\right\| \quad (3)$$

Let T_{df}, S'_{df} denote the DF map of the target image and deformed source image $S'_{df} = S_{df} \circ W$, respectively. $\bar{X}$ denotes the X after being cropped at borders. The first item is the alignment loss, which is the L1 loss between cropped DFs weighted by the reverse DFs (normalized to $0 - 1$), giving more emphasis on feature edges.

The second and third items are regularization terms. The second item regularizes that the average deformation gradient of an image should be close to 1 to prevent the whole image from being squeezed or stretched where the spatial extent will change. The third item penalizes extreme distortions between consecutive pixels. I is the identity grid and λ is the weighting term. This item is important to avoid large deformations where features change.

3 EXPERIMENTS AND DISCUSSION

3.1 Dataset and Implementation Details

In our experiment, we use two pairs of Siegfried map sheets of Switzerland around year 1880 at the scale of 1:25000[1], each pair is temporally adjacent for approximately 20 years depicting the same area and geo-referenced in a unified coordinate system. Each map sheet is 7000 × 4800 pixels with a spatial resolution of 1.25m. Small patches of 256 × 256 pixels were created for training and testing. After cropping the border, the output size of our network is 200 × 200 pixels. In total, 971 patches were selected randomly for training and 96 for testing. We trained our dataset for around 40K iterations until convergence and we noticed that with image-level loss, the model converged around 20K iterations. We used Adam optimizer with a initial learning rate of 0.002 and a decay of 0.0002 using Keras[2]. For all experiments we use GeForce GTX 2080Ti GPU. The regularization weight λ in equation (3) was set to 10^{-6}.

3.2 Results

We conducted experiments on both image-based loss and DF-based loss. We compared the performance of our proposed method with a state-of-art model for satellite image registration [20] that learns affine and deformable parameters in separate networks simultaneously. As mentioned before, since only translation is relevant in our approach, we only estimated two translation parameters using their method instead of six affine parameters. In [20], they mention that their proposed method is independent of the specific network architecture. To make experiments comparable, we ran their experiment using the same architecture as ours (i.e., ASPP-integrated U-Net). To evaluate the performance, we randomly selected 63 landmark locations that are unlikely to change within around two decades: road intersections, road/stream – contour line intersections, centroids of waterbodies, forest boundaries, inflection points of contour lines and triangulation points. We group the landmarks according to their original shifts and calculate the average misalignment (Table 1). For qualitative results (Figure 3), only configurations with DF-based loss are presented.

	Average Misalignment (m)		
Method	Small(<=10)	Medium (10,20]	Large (>20)
Unregistered	5.89	15.15	26.43
Deformation + Translation [20] (image-based loss)	5.21	15.15	26.43
Deformation + Translation [20] (DF-based loss)	13.45	10.93	21.51
Proposed (image-based loss)	5.55	15.15	26.43
Proposed (DF-based loss)	**2.72**	**7.73**	**16.81**

Table 1: Efficacy of different deformation configurations on registering unaligned landmark locations. We group the landmarks into three categories based on their initial displacements.

3.3 Discussion

In Table 1, we can see that using image-based loss only slightly reduces displacements for small misalignment. This corresponds to our reckoning that sparse historical maps per se can hardly give a continuous and progressive signal for registration. By contrast, DF-based loss can guide the network meaningfully, leading to a remarkable improvement of accuracy. Interestingly, modelling deformation and translation explicitly [20] helps to refine the registration of medium and large misalignment but distorts small alignment drastically on the other hand. This might be explained by a much larger spatial variance of misalignment in historical maps than in satellite images [20], where a global translation parameter is not suitable. Our proposed method with DF-based loss has achieved the best alignment accuracy, correcting misalignment for around 50%. For qualitative evaluation in Figure 3, we can see our proposed method can align contour lines (a,b), roads (a,b,c), streams (c) and forest boundaries (d) much more plausibly than [20]. For smaller objects like buildings and texts in (c,d), both deformation methods tend to distort them greatly when aligning elongated/linear objects. One possible reason is that the network is apt to favor the registration of elongated objects which occupy a much larger proportion than small structures. Since the spatial integration layer propagates deformation between neighboring pixels, small structures close to the deformed elongated features will be displaced even when they should stay unchanged. A multi-layer deformation network can be investigated in the future to separate objects into layers with different deformations. In our experiment, we observe that putting a small limitation on spatial gradients allows the network to align largely-displaced features, but generates unreasonable deformation in the regions of change on the other hand. An example could be found in (c) — where contour lines deform into the shape of buildings. To handle this dilemma, methods like automatic change detection can be applied to distinguish between changed and unchanged features. As the generated deformation grid is smooth in our approach, we can directly apply it to vector features, if available, from a source image to generate vectors that align with the target image without additional training. To achieve this, we first convert the feature vector into raster, warp / deform the raster with the predicted deformation grid, and convert the raster prediction back to the vector. Figure 4 shows an example of the result.

[1] https://www.swisstopo.admin.ch/en/geodata/maps/historical/siegfried25.html
[2] https://keras.io/

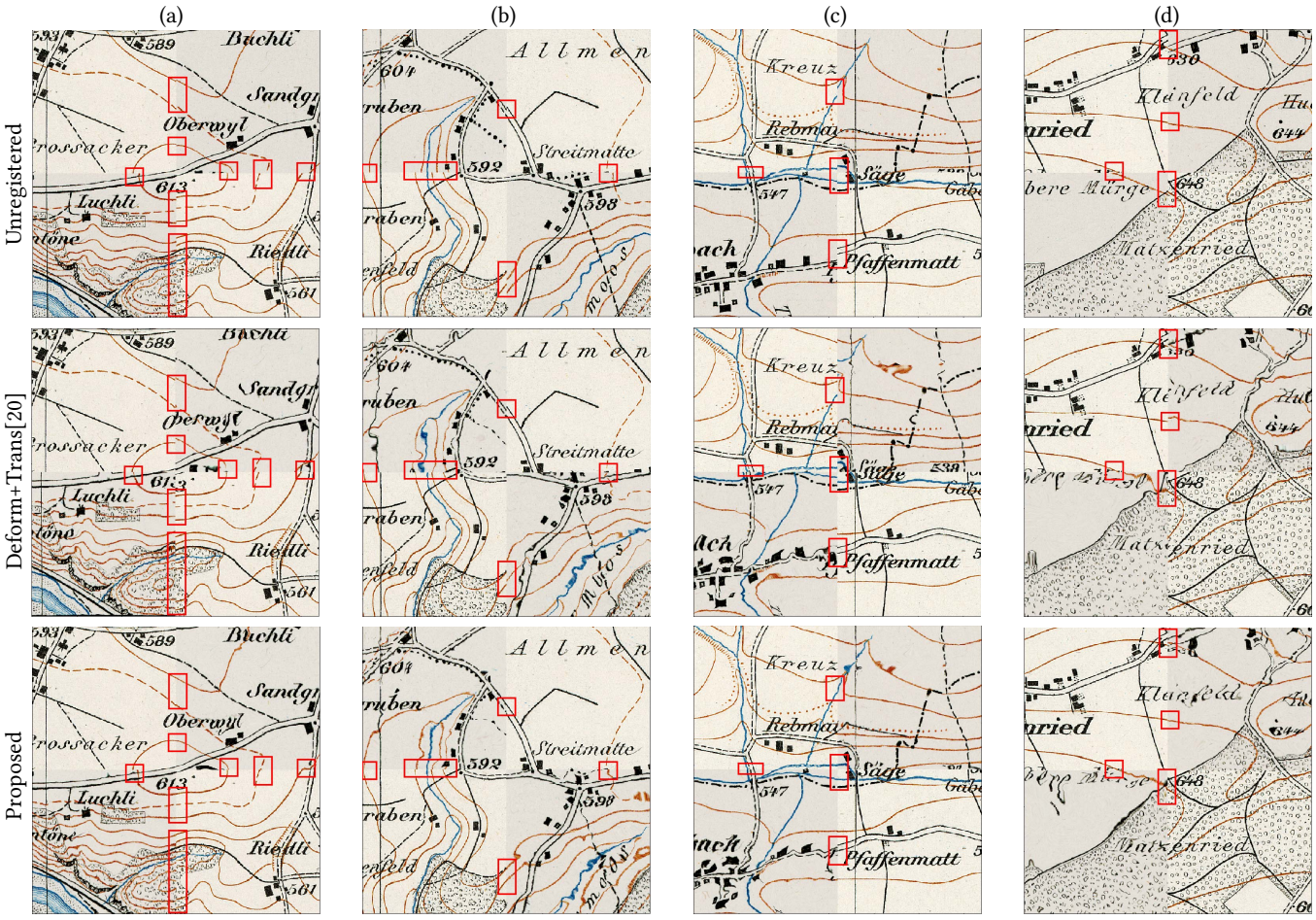

Figure 3: Qualitative comparison for four different areas. We darken the deformed tiles and their original images for visualization. Only configurations with DF-based loss are presented. Regions of interest are marked with red rectangles.

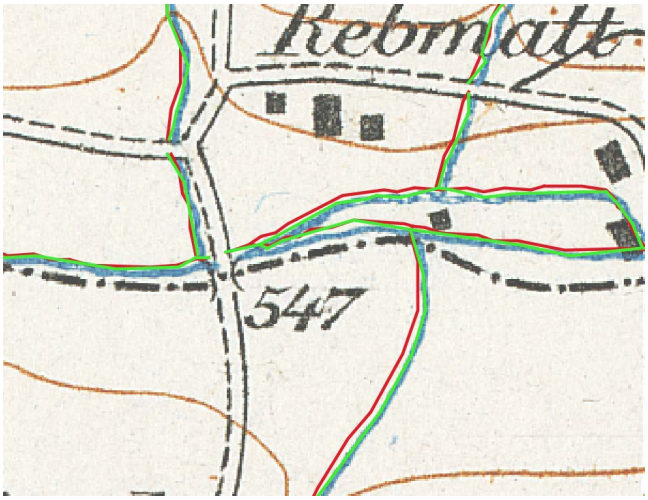

Figure 4: Applying a predicted deformation grid to the stream vector. The red and green line represents the vector of a source image and the deformed vector aligning with the target image (displayed as the base map), respectively.

4 CONCLUSIONS

In this work, we have proposed a novel image registration pipeline for historical maps. It is a purely unsupervised deep learning approach that tangles rigid and non-rigid transformations. We have introduced DF-based loss to tackle feature sparsity when image-based loss fails. Since we are the first to investigate unsupervised image registration for maps, we compare our model with a state-of-art approach for satellite imagery — a closely related field. Unlike their approach, our model does not need to decouple rigid and non-rigid transformation parameters and has performed significantly better. Smooth and continuous deformation grids generated by the proposed method show a good potential when being directly applied to vector features. In the future, we are going to investigate multi-layer deformation to separate features with different deformations and change detection methods to determine features to be/not be changed.

ACKNOWLEDGEMENTS

This paper is based on research works supported by the Swiss National Science Foundation (SNSF) under Grant No.188692 as a part of the project "HistoRiCH: Historical river change 547 - Planning for the future by exploring the mapped past".

REFERENCES

[1] Enrique J. Arriaga-Varela and Toru Takahashi. 2019. Automatic Georeferencing of Heterogeneous Historic and Illustrated Maps. In *Abstracts of the ICA*. Vol. 1. (July 2019). DOI: 10.5194/ica-abs-1-15-2019.

[2] John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, 6, 679–698. DOI: 10.1109/TPAMI.1986.4767851.

[3] Dengfeng Chai, Shawn Newsam, and Jingfeng Huang. 2020. Aerial image semantic segmentation using dcnn predicted distance maps. *ISPRS Journal of Photogrammetry and Remote Sensing*, 161, 309–322. DOI: https://doi.org/10.1016/j.isprsjprs.2020.01.023.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 764–773. DOI: 10.1109/ICCV.2017.89.

[5] Fatma El-Zahraa Ahmed El-Gamal, Mohammed Elmogy, and Ahmed Atwan. 2016. Current trends in medical image registration and fusion. *Egyptian Informatics Journal*, 17, 1, 99–124. DOI: https://doi.org/10.1016/j.eij.2015.09.002.

[6] Sarah F Frisken Gibson. 1998. Using distance maps for accurate surface representation in sampled volumes. In *IEEE Symposium on Volume Visualization*. IEEE, 23–30. DOI: 10.1109/SVV.1998.729581.

[7] Stefan Growe and Ralf Tonjes. 1997. A knowledge based approach to automatic image registration. In *Proceedings of International Conference on Image Processing*. Vol. 3. IEEE, 228–231. DOI: 10.1109/ICIP.1997.632067.

[8] Mattias P. Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V. Gleeson, Sir Michael Brady, and Julia A. Schnabel. 2012. Mind: modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical Image Analysis*, 16, 7, 1423–1435. Special Issue on the 2011 Conference on Medical Image Computing and Computer Assisted Intervention. DOI: https://doi.org/10.1016/j.media.2012.05.008.

[9] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. *Advances in neural information processing systems*, 28. DOI: https://doi.org/10.48550/arXiv.1506.02025.

[10] Hans J. Johnson and G. Christensen. 2002. Consistent landmark and intensity-based image registration. *IEEE Transactions on Medical Imaging*, 21, 450–461. DOI: 10.1109/TMI.2002.1009381.

[11] Konstantinos Karantzalos, Aristeidis Sotiras, and Nikos Paragios. 2014. Efficient and automated multimodal satellite data registration through mrfs and linear programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 329–336. DOI: 10.1109/CVPRW.2014.57.

[12] Qiaoliang Li, Guoyou Wang, Jianguo Liu, and Shaobo Chen. 2009. Robust scale-invariant feature matching for remote sensing image registration. *IEEE Geoscience and Remote Sensing Letters*, 6, 2, 287–291. DOI: 10.1109/LGRS.2008.2011751.

[13] Stan Z Li, Josef Kittler, and Maria Petrou. 1992. Matching and recognition of road networks from aerial images. In *European conference on computer vision*. Springer, 857–861. DOI: https://doi.org/10.1007/3-540-55426-2_99.

[14] Jonas Luft and Jochen Schiewe. 2021. Automatic content-based georeferencing of historical topographic maps. *Transactions in GIS*, 25, 6, 2888–2906. DOI: https://doi.org/10.1111/tgis.12794.

[15] Jonas Luft and Jochen Schiewe. 2021. Content-based Image Retrieval for Map Georeferencing. *Proceedings of the ICA*, 4, (Dec. 2021), 69, 69. DOI: 10.5194/ica-proc-4-69-2021.

[16] Wenping Ma, Zelian Wen, Yue Wu, Licheng Jiao, Maoguo Gong, Yafei Zheng, and Liang Liu. 2017. Remote sensing image registration with modified sift and enhanced feature matching. *IEEE Geoscience and Remote Sensing Letters*, 14, 1, 3–7. DOI: 10.1109/LGRS.2016.2600858.

[17] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. 2018. Deforming autoencoders: unsupervised disentangling of shape and appearance. In *Proceedings of the European Conference on Computer Vision*, 650–665. DOI: https://doi.org/10.48550/arXiv.1806.06503.

[18] Dmitriy Smirnov, Matthew Fisher, Vladimir G Kim, Richard Zhang, and Justin Solomon. 2020. Deep parametric shape predictions using distance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 561–570. DOI: https://doi.org/10.48550/arXiv.1904.08921.

[19] James Thewlis, Hakan Bilen, and Andrea Vedaldi. 2017. Unsupervised object learning from dense equivariant image labelling. In *Neural Information Processing Systems*. DOI: https://doi.org/10.48550/arXiv.1706.02932.

[20] Maria Vakalopoulou, Stergios Christodoulidis, Mihir Sahasrabudhe, Stavroula Mougiakakou, and Nikos Paragios. 2019. Image registration of satellite imagery with deep convolutional neural networks. In *IGARSS IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 4939–4942. DOI: 10.1109/IGARSS.2019.8898220.

[21] Maria Vakalopoulou, Konstantinos Karantzalos, Nikos Komodakis, and Nikos Paragios. 2016. Graph-Based Registration, Change Detection, and Classification in Very High Resolution Multitemporal Remote Sensing Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9, 2940–2951. DOI: 10.1109/JSTARS.2016.2557081.

[22] Sidi Wu, Magnus Heitzler, and Lorenz Hurni. 2022. Leveraging uncertainty estimation and spatial pyramid pooling for extracting hydrological features from scanned historical topographic maps. *GIScience & Remote Sensing*, 59, 1, 200–214. DOI: 10.1080/15481603.2021.2023840.

[23] Yuanxin Ye, Tengfeng Tang, Bai Zhu, Chao Yang, Bo Li, and Siyuan Hao. 2022. A multiscale framework with unsupervised learning for remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–15. DOI: 10.1109/TGRS.2022.3167644.

[24] Junhao Zhang, Masoumeh Zareapoor, Xiangjian He, Donghao Shen, Deying Feng, and Jie Yang. 2018. Mutual information based multi-modal remote sensing image registration using adaptive feature weight. *Remote Sensing Letters*, 9, 7, 646–655. DOI: https://doi.org/10.1080/2150704X.2018.1458343.

[25] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. 2016. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 117–126. DOI: https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.20.

[26] Barbara Zitová and Jan Flusser. 2003. Image registration methods: a survey. *Image and Vision Computing*, 21, 11, 977–1000. DOI: https://doi.org/10.1016/S0262-8856(03)00137-9.