UNSUPERVISED DEEP LEARNING NETWORK FOR DEFORMABLE FUNDUS IMAGE REGISTRATION

Giovana Augusta Benvenuto* *Marilaine Colnago* † *Wallace Casaca*†

* São Paulo State University, Faculty of Science and Technology, Presidente Prudente, Brazil
† São Paulo State University, Department of Energy Engineering, Rosana, Brazil

ABSTRACT

In ophthalmology and vision science applications, the process of registering a pair of fundus images, captured at different scales and viewing angles, is of paramount importance to support the diagnosis of diseases and routine eye examinations. Aiming at addressing the retina registration problem from the Deep Learning perspective, in this paper we introduce an end-to-end framework capable of learning the registration task in a fully unsupervised way. The designed approach combines Convolutional Neural Networks and Spatial Transformation Network into a unified pipeline that takes a similarity metric to gauge the difference between the images, thus enabling the image alignment without requiring any ground-truth data. Once the model is fully trained, it can perform one-shot registrations by just providing as input the pair of *fundus* images. As shown in the validation study, the trained model is able to successfully deal with several categories of *fundus* images, surpassing other recent techniques for retina registration.

Index Terms— Fundus image registration, deep learning.

1. INTRODUCTION

The image registration task consists of finding a geometric transformation that takes a given image to a reference one, by precisely aligning these images. Such a Computer Vision problem is of critical importance in several medical imaging applications, where the use of clinical exams supported by computer-acquired images is recurrent in order to carefully assist in the diagnosis and monitoring of diseases, including eye's pathologies and ocular disorders. In this context, retina (*fundus*) images are routinely taken and compared with images captured at different times and scales, or even by distinct instruments. Manually inspecting possible changes between two or more retina images is arduous, time-consuming, and error-prone, demanding the use of specific computational techniques to computerize the daily examination of the *fundus* images by eye care specialists [1, 2].

In the general context of registration, many interesting methods have been proposed, ranging from iterative approaches [3, 4, 2, 5] to Machine Learning-based models [6, 7, 8, 9, 10, 11, 12, 13]. Although there are several methods in the vast literature of image registration, recently, Litjens et al. [6] and Haskins et al. [12] pointed out that there is no consensus on a specific methodology that best combines the strength of Deep Learning (DL) with high-accuracy registration. Moreover, among the methods that are capable of specifically dealing with retinal *fundus* registration, there are only a few works based on DL paradigm, most of them devoted to coping with supervised learning instead of the unsupervised case. For instance, modern DL architectures such as the ones proposed by Mahapatra et al. [7] and Wang et al. [10] employ an interesting weakly supervised scheme to overcome the need for ground-truth data, by generating the benchmark data synthetically, however, it can lead to a loss of precision when registering low-quality *fundus* images. The method recently presented by Che et al. [8] introduces an unsupervised approach that addresses a particular group of *fundus* images, obtaining high accuracy for optimal disc alignment, however, despite the good results, their method is not able to register the vascularized areas of the retina.

In order to avoid the use of specific mechanisms to generate ground-truth data while still coping with many distinct categories of *fundus* images, in this paper we propose an end-to-end unsupervised learning registration framework that unifies Convolutional Neural Network (CNN) and Spatial Transformer Network (STN). The proposed technique takes advantage of a similarity metric that gauges the difference between the fixed and transformed images, allowing for performing the registration task without any ground-truth data. Once our network is fully trained, it can achieve one-shot registrations by just providing the desired pair of *fundus* images.

In summary, the main contributions of this paper are:

- A fully end-to-end framework for performing retina image registration using Deep Learning techniques.
- A neural network architecture that learns the registration task without using any ground-truth data or artificially created benchmark features.
- A functional and effective registration method capable of operating with distinct classes of *fundus* image pairs.

This project has received funding from São Paulo Research Foundation (FAPESP), grants #2013/07375-0, #2019/26288-7 and #2021/03328-3.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

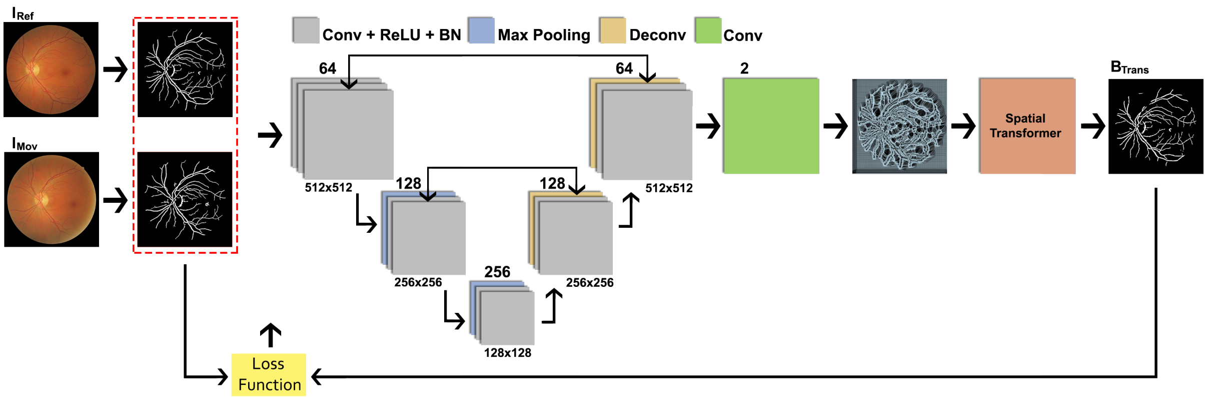

Fig. 1. Pipeline overview of the proposed unsupervised learning framework.

2. PIPELINE OVERVIEW

The proposed framework, which aims at registering a pair of *fundus* images in a fully unsupervised way, combines a neural network architecture based on the U-net [14] with the learning scheme recently proposed by Vos et al [15], where a Convolutional Neural Network estimates a set of matching points from the images, used sequentially by a Spatial Transformer Networks [16] to generate a deformation field which leads to the definitive bilinear interpolation.

Figure 1 illustrates the proposed framework. First, both reference and moving images, I_{Ref} and I_{Mov}, are segmented so that the extracted blood vessels and retina objects, B_{Ref} and B_{Mov}, are then used as input to train our CNN. Next, the network delivers a grid of points, which is subsequently taken by the STN to build a deformation field, used to generate the transformed image B_{Trans} from B_{Mov}. Finally, after the registration procedure, the image goes through a post-processing step, to filter out possible noisy pixels produced during the learning stage. Details of each step are given below.

2.1. Segmentation

First, the target *fundus* images go through a segmentation step that captures and highlights their main structures, such as blood vessels and ocular shape. Such a task is performed by applying the so-called *Isotropic Undecimated Wavelet Transform* (IUWT) [17], which in essence computes and takes the transformation coefficients of the images to binarize I_{Ref} and I_{Mov}. For implementation details, see [17].

2.2. Unsupervised Learning and Training

Our deep learning pipeline, which is designed to determine a correspondence grid between the images, relies on the U-net architecture [14]. As shown in Figure 1, the network gets as input the pair of concatenated segmentations, passing them to a block of convolutional layers. Next, two *dowsample* blocks, composed of two convolutional layers and a layer of *max pooling*, resample the images by halving their resolution while increasing the number of analyzed features per block. The subsequent block of layers, which accounts for the *upsample* process of the pipeline, is formed by a deconvolution layer and two convolutional layers. Each convolutional layer implemented in the network is followed by the ReLU activation function, and a *Batch Normalization* scheme. The outputs of each level from the *downsample* block are then concatenated with the entry of the corresponding level in the *upsample* block. Finally, the last layer, which is formed by two kernels, applies a linear activation function so as to generate the grid of points corresponding to the dimensions of the input images.

The output generated by the CNN, which gives a deformation grid, serves as input to the STN so that a bilinear interpolation is computed, thus allowing our approach to align the images. Once the image pair is properly registered, the loss function is then calculated via the Normalized Cross-Correlation (NCC) metric (1), which gauges the overlap among both fixed and processed images without the need for ground-truth data:

$$NCC(x,y)=\frac{\sum_{i=0}^{m}\sum_{j=0}^{n}T_{i,j}R_{i,j}}{\sqrt{\left(\sum_{i=0}^{m}\sum_{j=0}^{n}T_{i,j}^{2}\right)\left(\sum_{i=0}^{m}\sum_{j=0}^{n}R_{i,j}^{2}\right)}}, \quad (1)$$

where $T_{i,j}=t(x+i,y+j)-\bar{t}_{x,y}$, $R_{i,j}=r(i,j)-\bar{r}$, $t(i,j)$ and $r(i,j)$ are the pixel values at (i,j) of the matching and reference images, B_{Trans} and B_{Ref}, respectively, and $\bar{r}$ and $\bar{t}$ are the average pixel values w.r.t. B_{Ref}, and B_{Trans} [18].

The learning pipeline is then optimized until convergence by the so-called ADAM algorithm: an optimization technique based on the stochastic descending gradient method [19].

2.3. Post-Processing

The process of applying the deformable transformation to the moving image may eventually cause the presence of noise, especially under circumstances wherein the images to be aligned are very distinct from each other. In order to circumvent this, we add a post-processing denoising step to filter out the noise via Connected Component Analysis (CCA) [20].

The CCA algorithm is then applied to the moving image, generating a collection of objects grouped according to their adjacent pixels. Since *fundus* images are typically constituted by features and objects that imply continuous structures, including blood veins and ocular segments, small pixel portions that may indicate the presence of noise are discarded, resulting in a noise-free image.

3. DATA SETS, RESULTS AND DISCUSSION

Our framework was implemented in Python language using the Computer Vision package *OpenCV*, and the libraries *Tensorflow* and *Keras*. To run our experiments, the well-established database FIRE (*Fundus Image Registration Dataset*) [21] was used. The FIRE dataset is composed of 134 pairs of *fundus* images, which are classified into three categories of images according to their degree of overlapping and anatomical differences. These categories are described as follows [21]: both A and S include images with more than 75% estimated overlap, while P covers images with an estimated overlap value less than 75%. A is also the class of retinal representations that holds anatomical changes.

Our network architecture was trained using images of 512×512 pixels, from category S of FIRE database, with eight batches, for 5000 epochs. The tests were accomplished for all the three FIRE image categories in order to inspect how the network would behave when applied to different groups of *fundus* images.

For the sake of comparison and validation, four very recent image registration methods were taken in our analysis: the algorithms proposed by Hernandez-Matas et al. [5], Wang et al. [4], Motta et al [2] and Vos et al. [15], namely here as Rempe, GFEMR, VOTUS, and DIRNet, respectively.

Next, we present and discuss the obtained results, as well as the comparisons against other methods.

3.1. Quantitative evaluation

In order to quantitatively assess the registration results, the following similarity metrics were taken [8, 15, 10]: *Mean Squared Error* (MSE) [7], *Structural Similarity Index Measure* (SSIM) [22], and *Dice Coefficient* (Dice).

Table 1 presents the mean and standard deviation (between parentheses) for each category of *fundus* images from FIRE, before and after the registration as performed by our approach. From the listed values, one can check that the best scores are achieved after the registration process. Concerning the image categories, the best measurements are obtained when registering the image pairs from collections S and A. Finally, one can note that even in the drastic case of category P, which gathers retina images with low levels of overlaps, our learning scheme was capable of accurately aligning the images, producing satisfactory scores for all the metrics.

Table 1. Quantitative analysis before and after the registration by the proposed framework. The arrows indicate that "lower is better" or "higher is better", while values in bold highlight the best scores.

	Process	FIRE Dataset A	 S	 P
MSE (↓)	Before	0.0962 (0.0177)	0.0965 (0.0198)	0.1249 (0.0066)
	After	**0.0068 (0.0015)**	**0.0062 (0.0017)**	**0.0121 (0.0027)**
SSIM (↑)	Before	0.7307 (0.0421)	0.7237 (0.0457)	0.6510 (0.0177)
	After	**0.9731 (0.0055)**	**0.9749 (0.0068)**	**0.9575 (0.0076)**
Dice (↑)	Before	0.2982 (0.1088)	0.3418 (0.1384)	0.1245 (0.0119)
	After	**0.9502 (0.0100)**	**0.9579 (0.0120)**	**0.9103 (0.0238)**

Table 2 summarizes the mean and standard deviation for the registration results produced by each evaluated method when they are applied to the three categories of FIRE database. By numerically checking the tabulated values, the proposed approach was the one that delivered the best scores for all the evaluation metrics in all the analyzed data sets. In addition, it is worth mentioning that since the other registration methods were not able to fully register a few image pairs from category P, for a fairer assessment of the results, we pull off these particular cases of failure, compute scores for success cases only. In contrast, our framework was capable to align the image pairs regardless of the category and overlap level.

3.2. Qualitative evaluation

Aiming at improving the visual readability and interpretation of the results, in Figure 2, we follow [2, 23] so that the aligned images were grouped based on different colorizations. More

Table 2. Quantitative analysis of the registration methods.

	Methods	FIRE Dataset A	 S	 P
MSE (↓)	Proposed	**0.0068 (0.0015)**	**0.0062 (0.0017)**	**0.0121 (0.0027)**
	GFEMR	0.0522 (0.0145)	0.0280 (0.0053)	0.0525 (0.0095)
	Rempe	0.0487 (0.0240)	0.0196 (0.0056)	0.0616 (0.0132)
	VOTUS	0.0525 (0.0229)	0.0189 (0.0052)	0.0514 (0.0119)
	DIRNet	0.0710 (0.0182)	0.0601 (0.0237)	0.1040 (0.0070)
SSIM (↑)	Proposed	**0.9731 (0.0055)**	**0.9749 (0.0068)**	**0.9575 (0.0076)**
	GFEMR	0.8325 (0.0350)	0.8918 (0.0168)	0.8247 (0.0262)
	Rempe	0.8453 (0.0650)	0.9211 (0.0184)	0.8014 (0.0386)
	VOTUS	0.8317 (0.0562)	0.9232 (0.0180)	0.8279 (0.0340)
	DIRNet	0.7852 (0.0459)	0.8099 (0.0611)	0.6816 (0.0173)
Dice (↑)	Proposed	**0.9502 (0.0100)**	**0.9579 (0.0120)**	**0.9103 (0.0238)**
	GFEMR	0.6023 (0.1343)	0.8022 (0.0392)	0.5919 (0.0922)
	Rempe	0.6295 (0.1981)	0.8649 (0.0425)	0.5227 (0.1231)
	VOTUS	0.6105 (0.1802)	0.8702 (0.0388)	0.6149 (0.1004)
	DIRNet	0.4982 (0.1111)	0.6020 (0.1519)	0.2630 (0.0197)

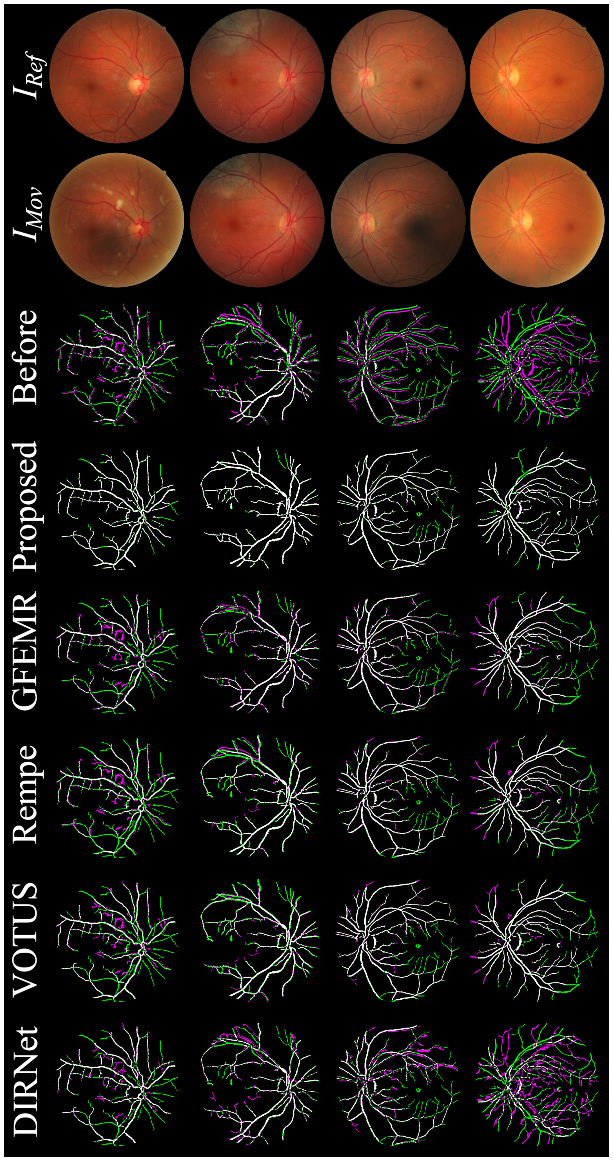

Fig. 2. Qualitative comparison between registration results for a pair of images after registration by all the methods.

specifically, B_{Ref} brings the reference image in green color, while B_{Mov} and B_{Trans} present the moving image before and after the registration, colored in magenta. Finally, the definitive image composition gives the amount of overlap between B_{Ref} and B_{Trans}, highlighted in white, i.e., a large amount of white pixels means better registration outcomes. In general, one can observe from the visual comparisons that our trained model achieves more consistent and pleasant results when compared against other methods for all the cases, mainly w.r.t. the quality of matching refinement as depicted by a large amount of white color in the illustrative montages.

Finally, in order to demonstrate that the registration task is also a learned behavior for the full database, in Figure 3

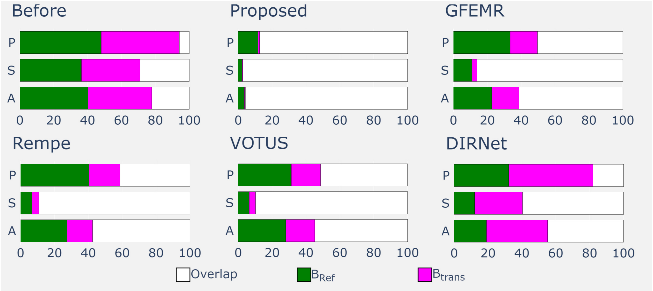

Fig. 3. Quantifying the % of overlap areas in the registrations.

we plot the percentage of pixels in green, magenta and white colors for each category of FIRE dataset. Similar to the experimental findings observed in Table 2 and Figure 2, our trained model was able to preserve consistency and accuracy while performing well in terms of overlap fitting capability, producing more white-filled areas than other registration algorithms.

4. CONCLUSIONS

This paper proposed an end-to-end framework for deformable registration of retinal images that relies on Deep Learning. Our approach was designed to operate in an unsupervised manner so that it does not require any specific apparatus to induce artificially created ground-truth data for training purposes. Once the model is fully trained, it allows for one-shot registrations by just providing the pair of *fundus* images.

In contrast to other modern image registration methods, our approach was capable of producing a definitive registration regardless of the overlap degree as well as the anatomical changes present in the images. Also, as verified by the experiments with three different categories of retina images from a well-established benchmark on image registration, our framework was able to outperform the others, both in qualitative as well as quantitative aspects. In summary, all those properties render the proposed framework a useful and compelling unsupervised registration technique for *fundus* images, achieving a high level of accuracy even in the absence of ground-truth data or large labeled data sets to train a definitive model.

5. ACKNOWLEDGMENTS

This research was supported by São Paulo Research Foundation (FAPESP), grants #2013/07375-0, #2019/26288-7 and #2021/03328-3, and the National Council for scientific and Technological Development (CNPq).

6. REFERENCES

[1] E. Karali, P. Asvestas, K. S. Nikita, and G. K. Matsopoulos, "Comparison of different global and local automatic

registration schemes: An application to retinal images," in *Int. Conference on Medical Image Computing and Computer-Assisted Intervention*, 2004, pp. 813–820.

[2] D. Motta, W. Casaca, and A. Paiva, "Vessel optimal transport for automated alignment of retinal fundus images," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6154–6168, 2019.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[4] J. Wang, J. Chen, H. Xu, S. Zhang, X. Mei, J. Huang, and J. Ma, "Gaussian field estimator with manifold regularization for retinal image registration," *Signal Processing*, vol. 157, pp. 225–235, 2019.

[5] C. Hernandez-Matas, X. Zabulis, and A.A. Argyros, "Rempe: Registration of retinal images through eye modelling and pose estimation," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3362–3373, 2020.

[6] G. Litjens et al, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[7] D. Mahapatra, B. Antony, S. Sedai, and R. Garnavi, "Deformable medical image registration using generative adversarial networks," in *IEEE International Symposium on Biomedical Imaging*, 2018, pp. 1449–1453.

[8] T. Che, Y. Zheng, J. Cong, Y. Jiang, Y. Niu, W. Jiao, B. Zhao, and Y. Ding, "Deep group-wise registration for multi-spectral images from fundus images," *IEEE Access*, vol. 7, pp. 27650–27661, 2019.

[9] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: A learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[10] Y. Wang, J. Zhang, C. An, M. Cavichini, M. Jhingan, M. J. Amador-Patarroyo, C. P. Long, D. G. Bartsch, W. R. Freeman, and T. Q. Nguyen, "A segmentation based robust deep learning framework for multimodal retinal image registration," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 1369–1373.

[11] Yuntong Tian, Yan Hu, Yuhui Ma, Huaying Hao, Lei Mou, Jianlong Yang, Yitian Zhao, and Jiang Liu, "Multi-scale u-net with edge guidance for multimodal retinal image deformable registration," in *International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 1360–1363.

[12] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: A survey," *Machine Vision and Applications*, vol. 31, no. 8, pp. 1–18, 2020.

[13] M. Hoffmann, B. Billot, J. E. Iglesias, B. Fischl, and A. V. Dalca, "Learning mri contrast-agnostic registration," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 899–903.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv e-prints*, p. arXiv:1505.04597, May 2015.

[15] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Medical Image Analysis*, vol. 52, pp. 128–143, 2019.

[16] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, vol. 28.

[17] P. Bankhead, C. N. Scholfield, J. G. McGeown, and T. M. Curtis, "Fast retinal vessel detection and measurement using wavelets and edge location refinement," *PloS One*, vol. 7, no. 3, pp. e32435, 2012.

[18] J. P. Lewis, "Fast normalized cross-correlation," *Industrial Light & Magic*, vol. 10, 10 2001.

[19] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[20] L. He, X. Ren, Q. Gao, X. Zhao, B. Yao, and Y. Chao, "The connected-component labeling problem: A review of state-of-the-art algorithms," *Pattern Recognition*, vol. 70, pp. 25–43, 2017.

[21] C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, S. Douma, and A.A. Argyros, "Fire: Fundus image registration dataset," *Journal for Modeling in Ophthalmology*, vol. 1, no. 4, pp. 16–28, 2017, source code: http://www.ics.forth.gr/cvrl/fire/.

[22] A. Kori and G. Krishnamurthi, "Zero Shot Learning for Multi-Modal Real Time Image Registration," *arXiv e-prints*, p. arXiv:1908.06213, Aug 2019.

[23] D. Motta, W. Casaca, and A. Paiva, "Fundus image transformation revisited: Towards determining more accurate registrations," in *IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, 2018, pp. 227–232.