International Journal of Electrical and Computer Engineering (IJECE)
Vol. 16, No. 3, June 2026, pp. 1188~1198
ISSN: 2088-8708, DOI: 10.11591/ijece.v16i3.pp1188-1198

Prostate magnetic resonance imaging/transrectal ultrasound registration using vision transformer and convolutional neural network

Hanae Mahmoudi, Hiba Ramadan, Jamal Riffi, Hamid Tairi
LISAC Laboratory, Department of Computer Science, Faculty of Sciences Dhar-Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco

Article Info

Article history:

Received Sep 29, 2024
Revised Jan 24, 2026
Accepted Mar 16, 2026

Keywords:

3D medical images
Affine registration
Convolutional neural networks
Dense displacement field
Multimodal registration
Supervised registration
Vision transformer

ABSTRACT

Multimodal registration of 3D medical images (3D-MReg) plays a key role in several medical applications and remains a very challenging task as it deals with multimodal images and volumetric objects at the same time. Recently, convolutional neural networks (CNNs) based approaches have been proposed to solve 3D-MReg. However, these techniques cannot preserve the global spatial context required for accurate affine registration since they rely on convolution and regional clustering operations. To solve these problems, we propose a supervised approach that combines both CNN and the vision transformer (ViT) to predict a dense displacement field (DDF). In a first step, our method investigates the power of ViT to capture global voxels dependencies for initial rigid alignment. Then we exploit the force of CNNs to focus on local details within pre-aligned concatenated input 3D moving and fixed images and estimate DDF, which is then applied to the moving labels. Our method has been validated in a prostate magnetic resonance imaging/transrectal ultrasound (MRI/TRUS) dataset and achieved promising results compared to previous work based on only CNNs.

Corresponding Author:

Hanae Mahmoudi
LISAC Laboratory, Department of Computer Science, Faculty of Sciences Dhar-Mahraz, Sidi Mohamed Ben Abdellah University
Fez-3000, Morocco
Email: hanae.mahmoudi@usmba.ac.ma

1. INTRODUCTION

Image registration aims to align two images of the same scene—namely a reference image and a moving image acquired at different times, from distinct viewpoints, and/or using different imaging devices. In medical imaging, the objective of registration is to estimate an appropriate spatial transformation that achieves accurate alignment of the corresponding anatomical structures. Medical image registration (MReg) is therefore a fundamental component in many clinical and research applications [1], including organ segmentation [2], atlas construction, image-guided interventions for diagnosis, monitoring, and therapy, as well as tele-surgery and post-operative evaluation.

MReg techniques can be classified according to several criteria. With respect to the imaging modalities involved, such as magnetic resonance imaging (MRI), computed tomography (CT), ultrasound (US), and others, registration methods are commonly divided into unimodal approaches, where both images are acquired with the same modality (*e.g.*, MRI–MRI, CT–CT, US–US), and multimodal approaches, where the images originate from different modalities (*e.g.*, MRI–CT, MRI–US). From a dimensional viewpoint, depending on whether the data are planar or volumetric, registration problems can be formulated as 2D–2D,

2D–3D, or 3D–3D alignment. According to the transformation model, three principal categories are generally considered: rigid, affine, and deformable registration. Rigid registration is parameterized by six degrees of freedom corresponding to translations and rotations. Affine registration generalizes this model by additionally allowing scaling and shearing, leading to twelve degrees of freedom. By contrast, deformable or non-rigid registration estimates spatially varying transformations at the voxel level and typically represents the deformation using a dense displacement field (DDF) to capture complex non-linear anatomical variations.

Multimodal registration of 3D medical images (3D–MReg) is a specific type of registration that involves the alignment of data containing the same 3D object but captured using different modalities. Deformable registration plays an essential role in multimodal registration by offering a more sophisticated and flexible approach to align images from different modalities. Deformable registration makes it possible to model local and non-linear deformations between images, offering a more accurate correspondence. This ability to take account of complex anatomical variations, local distortions, and differences in contrast between modalities makes deformable registration particularly well suited to the challenges posed by multimodal images. In the other hand, applying the affine transformation is necessary for the multimodal registration process, particularly as an initial step before deformable registration. It ensures global alignment and uniform initialization for images from different modalities. The affine transformation creates a solid foundation that accurately handles spatial disparities, accounting for variations in scale and orientation, effectively limiting the search space for subsequent deformable registration. By combining affine and deformable registration, the multimodal registration process becomes more resilient and capable of providing an accurate overlay, taking into account global and local variations between images from different modalities.

Many conventional algorithms have been developed and studied for 3D–MReg [3], [4] based on optimizing a similarity metric, such as a sum of squared differences (SSD) or mutual information (MI) [5], by changing the transformation's settings. The optimization process for traditional registration techniques of 3D multimodal images frequently begins with a preliminary estimation of the transformation parameters, which can lead to convergence to incorrect registration results. In the same way, their processing is time-consuming, particularly for high-resolution or large volumes of images.

Deep learning-based registration methods have recently addressed these limitations and tried to solve the problem of 3D–MReg [6], [7] using the learning capabilities of neural networks. In fact, deep learning-based approaches can automatically estimate transformation parameters while learning complex correspondences between images. Convolution neural networks (CNNs) based approaches are one of deep learning techniques that can extract spatial and intensity data and predict DDF to perform 3D–MReg [8], [9]. Although CNNs can detect local features because of their dependence on regional convolution and pooling procedures, they often need to retain the global spatial context required for accurate 3D–MReg, especially when complex and extensive interactions exist between several components of the image volume.

Thanks to the great success of transformer [10] in natural language processing (NLP), a big focus is given actually to self-attention (SA) mechanism-based architectures in many computer vision tasks and medical image applications [11] to improve the nonlocal modeling capability. Since prioritizing global picture information is necessary for registration tasks, vision transformers [12] may be employed to learn the overall image representations and gather semantic elements crucial for 3D-MReg using attention mechanisms [13].

Inspired by the work in [14], which proposes mono-modal 3D image registration based on the vision transformer (ViT), and the works [8], [9], where the authors develop a CNN-based multimodal 3D-MReg, we propose a new approach of supervised deformable multimodal image registration based on the combination of ViT and CNN which we name 3D-$MReg_{aff}$ ViT. Our method leverages anatomical labels to estimate voxel-wise transformations. The proposed framework is evaluated on 3D transrectal ultrasound (TRUS) and T2-weighted MRI data from prostate cancer patients and demonstrates encouraging performance for the 3D medical image registration task. The main contributions of this work are summarized:

a. In a first step, we exploit the power of ViT to propose a ViT-based network that we name 3D-$MReg_{aff}$ ViT to achieve accurate affine image registration. The output of this stage will be fed to a second network to perform deformable registration.
b. In the second step, we harness the strength of CNNs to focus on local details. The local-net takes the result of 3D-$MReg_{aff}$ ViT to predict the local DDF. In the final step, we combine the affine matrix and the local DDF to predict the final output DDF which will be applied to the moving labels.
c. Validation on a public prostate cancer MRI/TRUS dataset demonstrates a median target registration error of 1.7 mm at landmark centroids and a median Dice coefficient of 0.95 for the prostate gland.

The remainder of this article is organized as follows. Section 2 reviews related work on multimodal medical image registration. Section 3 introduces 3D-$MReg_{aff}$ ViT, the proposed framework for multimodal registration of volumetric medical images that integrates vision transformers and convolutional neural

networks. Section 4 reports the experimental setup and evaluation results used to assess the effectiveness of the proposed method. Finally, section 5 concludes the paper.

2. RELATED WORK

Multimodal registration plays a central role in numerous medical image analysis pipelines. In recent years, deep learning–based multimodal image registration has attracted considerable attention. In [8], the authors proposed a CNN-based framework to estimate dense voxel correspondences by exploiting multiple identifiable anatomical label types. Their method introduced a dedicated network design to efficiently predict both global and local deformations arising in prostate cancer surgery through MRI/TRUS registration. In a subsequent extension [9], a unified network was proposed to directly predict DDFs at multiple spatial resolutions.

When explicit similarity measures are unavailable, neural network–driven approaches have been employed to infer dense voxel correspondences from a variety of anatomical label shapes, thereby capturing meaningful anatomical features [6]. Haskins *et al.* [15] introduced a deep CNN framework to learn a similarity metric for MRI/TRUS registration and combined it with a hybrid optimization strategy to obtain suitable initializations for second-order optimization.

Simonovsky *et al.* [16] formulated 3D medical image registration as a classification problem to distinguish between well-aligned and misaligned image patches across different modalities using CNNs. In [17], the coherent point drift algorithm was adopted to elastically register surfaces, and the deformation field obtained from thin-plate splines was propagated to the entire gland. A weighted self-similarity structure vector (WSSV) was later applied to perform multimodal registration [18], and this strategy was shown to improve the intraoperative localization of 3D MRI and 2D ultrasound lesions in fusion-guided navigation systems.

Blendowski *et al.* [19] proposed a framework that relies solely on independently obtained segmentation labels from each modality and exploits anatomical shape priors. A shape-constrained encoder–decoder segmentation network was first trained on labeled CT and MRI data without skip connections. Subsequently, an iterative energy-minimization scheme, driven by the network's ability to generate intermediate non-linear shape representations, was introduced to enhance multimodal alignment in the presence of large deformations.

More recently, Song *et al*. [20] presented a cross-modal attention–based approach that correlates feature representations extracted from multimodal image pairs and leverages the resulting correlations to refine the registration process. In [21], a Bayesian framework was developed to construct multi-class brain atlases and to model large inter-subject deformations between multimodal brain images by exploiting modality-specific neuroanatomical information. Finally, a deep reinforcement learning–based multimodal registration method was introduced in [22], which aims to reduce modality discrepancies by automatically learning robust and discriminative feature representations.

3. METHOD

In this section, we present our new method 3D-MReg$_{\text{Def}}$ ViT to perform DDF of multi-modal medical images. Due to the non-linear relationships between input image pair especially when utilizing images from different modalities, the registration procedure remains very challenging. The use of labels with anatomical details helps make the process smoother.

Figure 1 illustrates the overall workflow of the proposed framework. At the beginning, training samples are randomly generated by selecting equal numbers of image pairs together with their corresponding labels. These samples are then processed within a mini-batch optimization scheme. In the next stage, a ViT receives the fixed and moving image pair as input and estimates the affine transformation matrix, which is subsequently used to warp the moving image toward the fixed one. We refer to this ViT-based module for 3D affine registration as 3D-MReg$_{\text{Def}}$ ViT. Subsequently, a local-net takes as input the concatenation of the affinely warped moving image and the fixed image to predict a local non-rigid DDF, which is regularized using a weighted bending energy term [23]. The affine transformation and the predicted local DDF are then combined to form the final displacement field, which is applied to warp the moving labels to align with the fixed labels. Label correspondence is evaluated using the multiscale Dice metric [24]. During training, stochastic gradient descent [25] is employed, relying on an unbiased gradient estimator under the assumption that images and labels are conditionally independent, which enables effective minimization of the loss function. Let L and M denote the fixed and moving training images, and L_f and L_m their corresponding labels. The following subsections describe each component of the proposed architecture in detail.

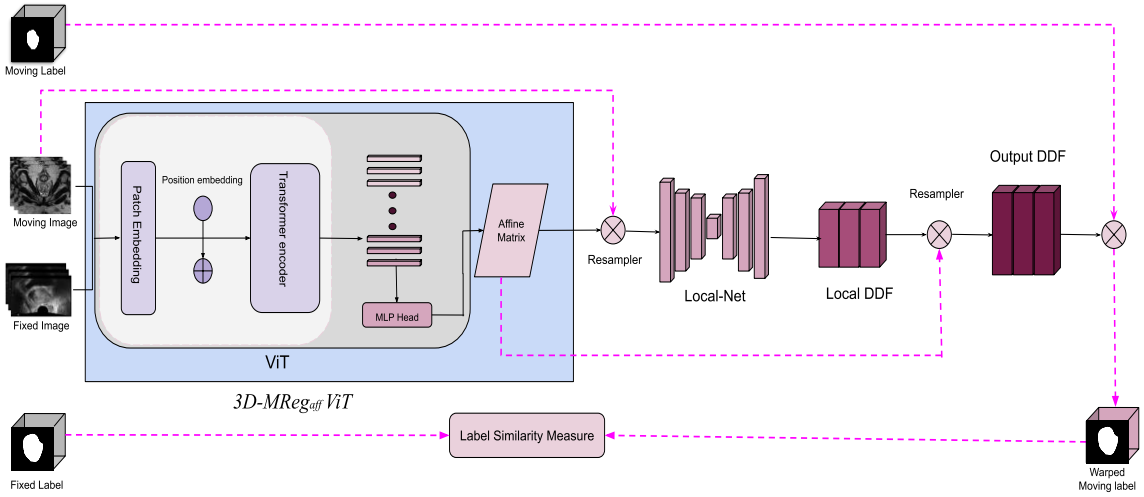

Figure 1. Overview of the proposed approach $3D - MReg_{Def}\ ViT$

3.1. $\mathbf{3D - MReg_{Aff}\ ViT}$

We aim to obtain the optimal affine matrix to align F and M by the transformer encoder [12] as illustrated in Figure 2. After resizing the images to the same size, the concatenation of F and M denoted by C is divided into N vectorized 3D patches $C_p \in \mathbb{R}^{N \times P^3}$ where $= \frac{H \times W \times L}{P^3}$, and (H, W, L) indicate the dimension of C. Then, each patch is mapped to a D vector by applying a learnable linear embedding $C_{pe} = C_p \mathrm{E}$:

$$C_{pe} = [\, C_p^1\, E\, ;\, C_p^2\, E\, ; \ldots \ldots ;\, C_p^N E\,]\, ,\ \mathrm{E} \in \mathbb{R}^{p^3 \times D} \tag{1}$$

Following that, we added a learnable positional embedding E_{pos} to C_{pe} :

$$C_i = C_{pe} + E_{pos}\ ,\ E_{pos} \in \mathbb{R}^{N \times D} \tag{2}$$

The generated patches, called tokens, are then sent to the transformer encoder, which uses them to calculate self-attention (SA). The SA records the dependencies between input sequence tokens C_i. The C_i are translated linearly into the query (Q), key (K), and value (V) matrices by multiplying them with learned weight matrices W:

$$\begin{cases} \mathrm{Q} = C_i\, W_q \\ \mathrm{k} = C_i\, W_k \\ \mathrm{V} = C_i\, W_v \end{cases} \tag{3}$$

The attention weights are then obtained by calculating the attention scores and using the SoftMax function:

$$A_w = softmax\,(\frac{Q\, K^T}{\sqrt{D_k}}) \quad \mathrm{A} \in \mathbb{R}^{N \times N} \tag{4}$$

The scaled dot-product attention is constructed as (5):

$$\mathrm{SA}\,(C_i) = A_w\, V \tag{5}$$

Then, *SA* is calculated several times in parallel with different learned weight matrices, creating multi-headed attention. Therefore, the model can focus on various pieces of input data and capture multiple interactions. Following, the results are transmitted to a feed-forward neural network. At the end of each SA and feed-forward, layer normalization and residual connections are applied. The produced output of the transformer encoder is given as input to the classification head that implements the multi-layer perceptron with the activation function Gaussian error linear unit (GELU). Finally, the resulting is a set of 12-degree-of-freedom affine matrix.

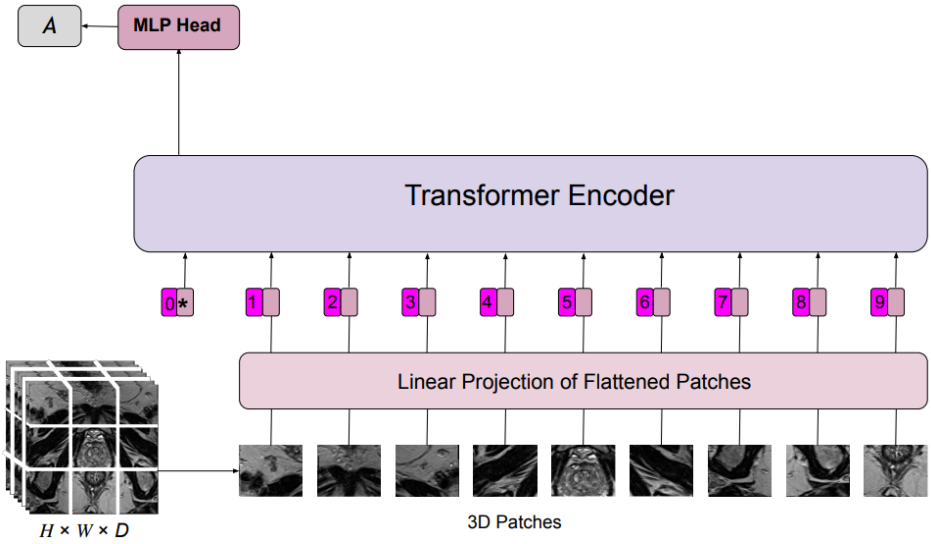

Figure 2. VIT architecture applied to 3D medical images used in our approach

3.2. Local-net for deformable registration in $3D-MReg_{Def}$ ViT

The predicted affine transformation matrix resulting by $3D-MReg_{Aff}$ ViT is used to deform the M. The warped moving image is represented by $\widehat{M}$. After being concatenated, F and $\widehat{M}$ are fed into the local network to predict the local DDF. The local-net [8] comprises four subsampling blocks and four up-sampling blocks, incorporating shortened connections at different levels of resolution. By beginning with an initial 32 channels, the network strategically doubles and halves the number of the channels during subsampling and Up-sampling, respectively. Each block uses residual network units (ResNet) with shortened connections, improving the network's training capabilities. Subsampling is performed through convolutional layers with steps of two while up-sampling uses transposed convolutional layers. The output layer of the network initializes with an additional convolution operation and a bias term, deliberately lacking batch normalization or non-linear activation to allow random initialization with zero mean and minimal variation. At last, local DDF is computed. The final output DDF results from combining the affine transformation, produced by $3D-MReg_{Aff}$ ViT, and the local DDF, generated by local-net. Using this combined DDF, the moving label L_m is deformed into the fixed label L_f one by a trilinear image resampler. To avoid over-fitting issues, binary masks of labelled images have been preprocessed using a one-sided smoothing mechanism following the same strategy in [8]. For the inference part, only the unlabeled image pairs are used as input data for the model.

3.3. Multi-scale dice

The soft probabilistic Dice metric proposed by Milletari *et al.* [24] forms the foundation of the multi-scale Dice measure. It is defined a:

$$Dice\ (L_F\ ,L_M\ ) = \frac{2\ \sum_{i-1}^{I} L_{F_i}\ L_{M_i}}{\sum_{i-1}^{I} L_{F_i} + \sum_{i-1}^{I} L_{M_i}} \tag{6}$$

where $L_F\ = \{L_{F_i}\}$, $L_M\ = \{L_{M_i}\}$ denote the fixed and moving label maps, respectively, with $\{L_{F_i}, L_{M_i}\} \in [0,1]$ and $i = 1, \dots \dots, I$, over I voxels in the image. The multi-scale Dice metric is then defined as:

$$MSD_k = \frac{1}{s}\sum_{\sigma} Dice\ (f_\sigma(L_k^M), f_\sigma(L_k^F)) \tag{7}$$

where $L_k^M\ = \{(L_k^M)_i\}$, $L_k^F\ = \{(L_k^F)_i\}$ are a pair of binary label maps. The operator f_σ denotes a 3D Gaussian smoothing filter with isotropic standard deviation σ. In this work, σ ∈ {0, 1, 2, 4, 8, 16} corresponding to $S = 6$ different scales. The adoption of the multi-scale Dice metric enhances sensitivity to spatial relationships between anatomical labels, in contrast to conventional measures such as Dice, Jaccard, and cross-entropy, which do not explicitly account for spatial information.

4. EXPERIMENTS

4.1. Datasets and configuration

We assessed the proposed method using prostate MRI/TRUS data from the smart target® clinical trials [26]. The dataset includes 108 matched pairs of T2-weighted MRI and TRUS volumes acquired from 76 patients. Because patients underwent different clinical interventions, each subject could have up to three separate imaging sessions, including biopsy procedures, therapeutic interventions, or repeated echography acquisitions [27].

Both MRI and TRUS volumes were resampled to an isotropic spatial resolution of 1.0 $mm^3$and subsequently standardized to zero mean and unit variance. With respect to the anatomical annotations, prostate segmentations were first manually delineated by a medical student in three orthogonal views axial, sagittal, and coronal and then reviewed and corrected by an experienced urologist. For the TRUS modality, gland contours were manually refined based on automatically estimated prostate profiles on the original ultrasound slices [28]. Prostate segmentations on MRI were provided directly as part of the clinical study protocols [26]. The anatomical labels were encoded as binary masks and smoothed using a normalized inverse distance transform after resampling to match the spatial dimensions of the corresponding MRI or TRUS volumes. The proposed framework was implemented in Python with the TensorFlow library and trained on a workstation equipped with a 24 GB NVIDIA Quadro P6000 GPU. Optimization was carried out using the Adam algorithm with a mini-batch size of 2, a learning rate of 10^{-5}, and a total of 1000 training iterations.

4.2. Evaluation metrics

In this work, a 10-fold cross-validation strategy was conducted at the patient level. In each fold, data from 7 to 8 patients were held out for testing, while the remaining patient scans were used for training. The performance of the proposed method was evaluated using two quantitative measures: the dice similarity coefficient (DSC) and the target registration error (TRE). DSC is determined between the binary warped and fixed labels representing the prostate glands:

$$DSC\ (L_F\ , L_M\ ) = \frac{2\ |L_F \cap L_M\ |}{|\ L_F\ | + |L_M|} \tag{8}$$

where L_F is fixed label and L_M is the warped moving label. TRE is computed as the root mean square of the spatial distance between the centers of mass of the warped labels and those of the corresponding fixed labels.

$$TRE\ (L_F\ , L_M\ ) = \sqrt{\frac{1}{N}\ \sum_{i=1}^{N} \|L_{Fi} - \ L_{Mi}\|^2} \tag{9}$$

where L_{Fi} the coordinates of a point in fixed label, L_{Mi} is the coordinates of a point in the warped moving label and N is the total number of corresponding points or landmarks.

4.3. Results and discussion

Figure 3 illustrates the registration results produced by $3D - MReg_{aff}$ ViT and $3D - MReg_{def}$ ViT using image slices from two representative test cases (Cases 1 and 2). Rows a–d correspond to slices from the original MRI images, the fixed TRUS images, the MRI images warped by $3D - MReg_{def}$ ViT and the MRI images warped by $3D - MReg_{def}$ ViT, respectively.

In our study, we started the training with $3D - MReg_{aff}$ ViT to predict the affine transformation. Then, we used $3D - MReg_{Def}$ ViT to predict the DDF. We compared our approach including both of $3D - MReg_{aff}$ ViT and $3D - MReg_{def}$ ViT with CNN-based methods proposed in [8], [9] using the Dice scores and the TRE measure as evaluation metrics. In the original work of Hu *et al.* [8], they used the global-net, local-net and composite-net methods. The global-net method predicts an affine transformation and the local-net method estimates a non-rigid local DDF, while the Composite-net method is composed of the Global-net followed by the local-net. Later, in their extended study [9], they integrated these networks into a single framework that predicts the DDF as final output. All these methods are tested in the same conditions as our approach, using the same dataset and the same numbers of image pairs. Moreover, we employed an adaptive gradient descent optimization process and similar execution environment as in [9].

Table 1 and Figure 4 summarize the quantitative results in term of dice metric and Table 2 presents the TRE measures for all the compared methods. Our proposed network for $3D - MReg_{aff}$ViT , yielded a median DSC of 0.90, with 5th and 95th percentiles of 0.80 and 0.94, respectively. In parallel, the median TRE for landmark centroids obtained from the same networks measured 6.0, with 5th and 95th percentiles of 2.9 and 10.9, respectively. Moreover, $3D - MReg_{def}$ ViT gave a median DSC of 0.95, with 5th and 95th percentiles of 0.83 and 0.97, respectively. Simultaneously, the median TRE for landmark centroids obtained

from the same networks measured 1.7, with 5th and 95th percentiles of 0.5 and 6.6, respectively. We deduce that our approach achieves the highest dice score and lower TRE compared to the other approaches (with p-value < 0.001). By the following, we will discuss these results in details.

Firstly, we explore the power of the ViT integration in the affine registration step. We compare $3\mathrm{D}-\mathrm{MReg}_{\mathrm{aff}}$ ViT (the first step of our network) with Global-Net, which is the closest method to our work since both of them tackle the affine transformation task. As reported in Table 1, our approach produces 90% in term of DSC value outperforming Global-Net by 13%. In addition, $3\mathrm{D}-\mathrm{MReg}_{\mathrm{aff}}$ ViT outperforms local-net, Composite-Net and those generated by Hu *et al.* [9] with 8%, 6% and 2% respectively, even these networks have been proposed to handle deformable registration. This performance of $3\mathrm{D}-\mathrm{MReg}_{\mathrm{aff}}$ ViT is due to ViT's known ability to model long-term data dependencies using SA mechanisms [10], [11], especially in affine registration, which involves the correction of large-scale spatial transformations such as translations, rotations and scaling. On the other hand, CNNs have been very effective in registering deformable medical images [27], but are less suited in modeling and learning affine registration. When it comes to multimodal registration of medical images, which requires detailed local attention as well as global understanding, affine registration proves insufficient. This is why, in our study, we opted for the application of affine transformation as an initial step using ViT ($3\mathrm{D}-\mathrm{MReg}_{\mathrm{aff}}$ ViT), followed by a CNN network to achieve deformable registration. This combination represented by our $3\mathrm{D}-\mathrm{MReg}_{\mathrm{def}}$ ViT led to significantly better results than the exclusive use of CNN or ViT alone.

As depicted in Table 1 and Table 2, the Dice and TRE scores of $3\mathrm{D}-\mathrm{MReg}_{\mathrm{def}}$ ViT, reached 95% and 17.8% respectively. This represents a 5% improvement of the results obtained by local-net, a 7% improvement of those produced by Hu *et al.* [9], in terms of Dice value. Concerning the TRE score, a 5% improvement on the results obtained by local-net, and a 7% improvement on those generated by Hu *et al.* [9] have been reached. The error produced is the lowest, which indicates that our method is more effective at minimizing registration errors than other approaches. A minimal TRE indicates a better spatial match between the warped image and the reference image, underlining the superior performance of our method in terms of registration accuracy compared to other approaches. The obtained results of $3\mathrm{D}-\mathrm{MReg}_{\mathrm{def}}$ ViT improve significantly the DDF compared to deformable CNN-based methods. This fact can be explained by the focus of CNN architectures on local informations and neglect global image context. In contrast, the use of ViT in our model in the preprocessing step solves the limitations of convolutional operations and led to capture global anatomical structures dependencies to be processed in the second network and generate accurate DDF. We notice that it is difficult to conduct safe comparison with other works proposed in the literature of 3D-MReg because of the lack of available benchmarks and most of published papers in the field deal with specific medical applications and private datasets.

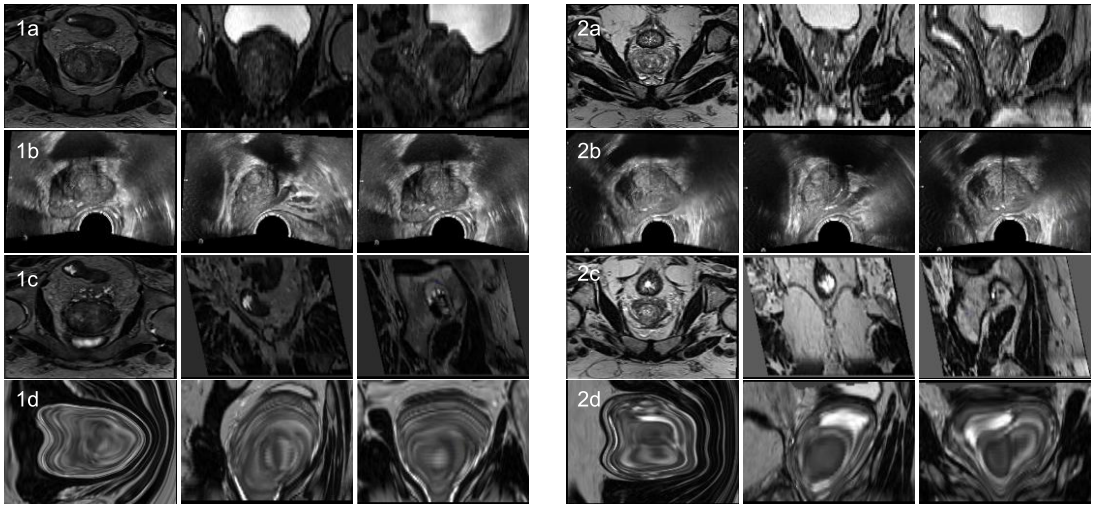

Figure 3. Example of image slices from two test cases, 1 and 2. Rows a, b, c, and d show the slices extracted from the original MRI images, the corresponding fixed TRUS images, the slices generated by $3D-MReg_{aff}\ ViT$ from the warped moving MRI images, and the slices generated by $3D-MReg_{def}\ ViT$ from the warped moving MRI images, respectively

Table 1. The results of the current study in terms of DSC and their comparison with another recent research

	DSC	
	Median	Percentiles [5th, 95th]
Global-net [8]	0.77	[0.46, 0.87]
Local-net [8]	0.82	[0.46, 0.91]
Composite-net [8]	0.84	[0.80, 0.94]
Hu *et al.* [9]	0.88	[0.77, 0.91]
$3D-MReg_{aff}ViT$	0.90	[0.80, 0.94]
$3D-MReg_{def}ViT$	0.95	[0.83, 0.97]

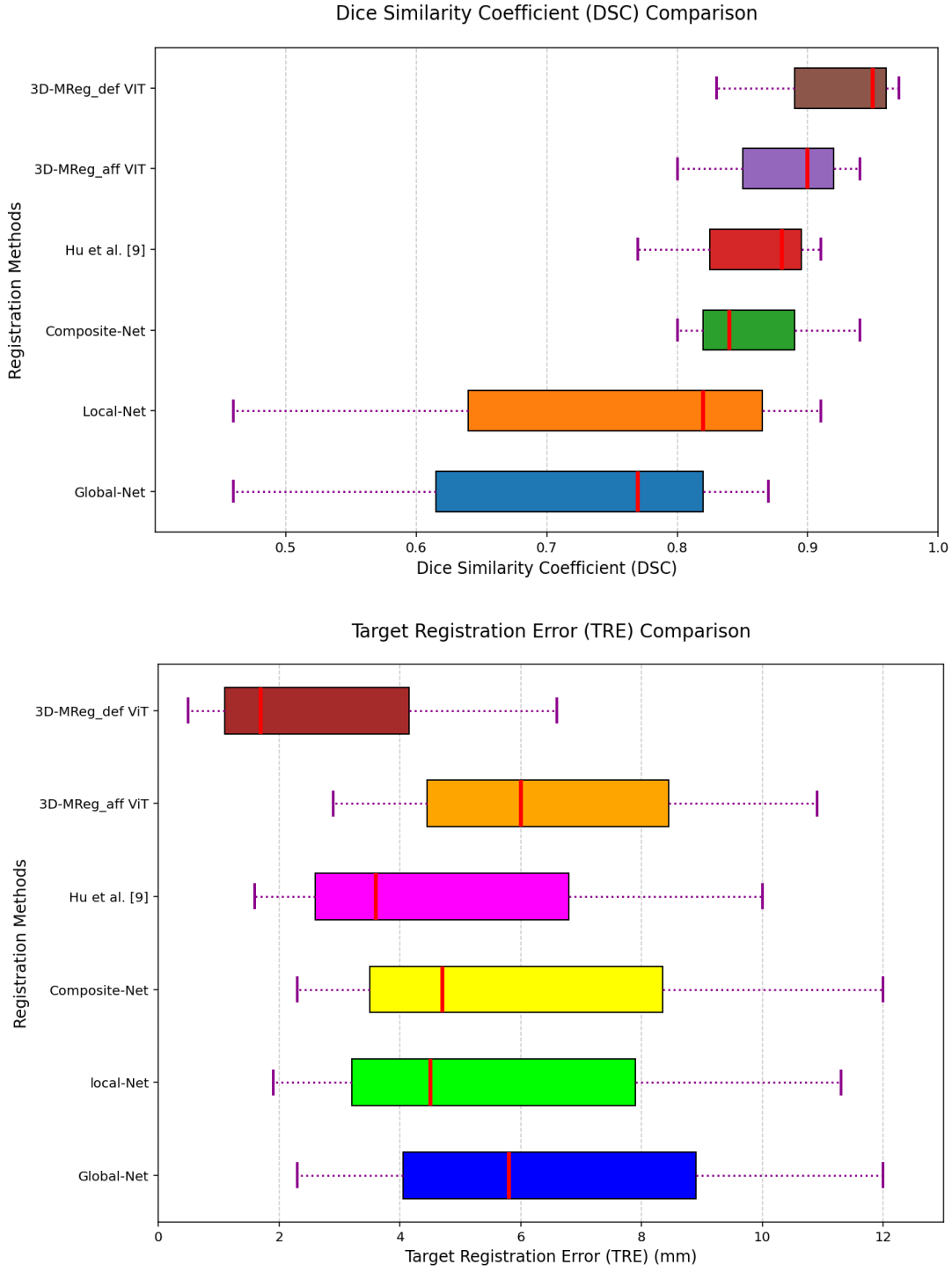

Figure 4. Tukey boxplots displaying the cross-validation results for the different methods

Table 2. The results of the current study in term of TRE and their comparison with other recent research

	TRE	
	Median	Percentiles [5th, 95th]
Global-net [8]	5.8	[2.3, 12.0]
Local-net [8]	4.5	[1.9, 11.3]
Composite-net [8]	4.7	[2.3, 12.0]
Hu *et al.* [9]	3.6	[1.6, 10.0]
$\mathbf{3D-MReg_{aff}ViT}$	6.0	[2.9, 10.9]
$\mathbf{3D-MReg_{Def}Vi}$	1.7	[0.5, 6.6]

5. CONCLUSION

This paper presents deformable multimodal registration of medical images based on ViT and CNN ($3D-\mathrm{MReg_{def}}$ ViT). It uses 3D medical images including MRI and TRUS, considered respectively as moving and reference images, and their labels. In this process, images and labels are preprocessed before use: MRI and TRUS images are resized and normalized, then their labels are re-dimensioned to the same size as images and smoothed using inverse distance transform. The whole architecture takes the concatenation of MRI and TRUS images as input to the ViT architecture to predict the affine transformation matrix that will be applied to the moving image. Then, the resulting image will be concatenated with the fixed image and fed to the local network to predict the local DDF. The final step consists of combining the transformation matrix with the local DDF to predict the output of the DDF, which is then applied to the moving labels. The results are evaluated using the dice metric calculated between the moving deformed and fixed targets. Our proposed framework achieved promising results compared to previous works using pure CNN architectures. As future work, we will explore other variants of ViTs and test other datasets in the field.

ACKNOWLEDGEMENTS

The authors sincerely thank the anonymous reviewers for their constructive comments and insightful suggestions, which significantly improved this manuscript.

FUNDING INFORMATION

Authors state no funding involved.

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] Q. Zheng, C. Liu, and J. Chang, "Non-rigid registration of medical images based on $S_2^1(\Delta_{mn}^{(2)})$ non-tensor product B-spline," *Visual Computing for Industry, Biomedicine, and Art*, vol. 5, no. 1, p. 5, Dec. 2022, doi: 10.1186/s42492-022-00101-8.

[2] J. Ying *et al.*, "Two fully automated data-driven 3D whole-breast segmentation strategies in MRI for MR-based breast density using image registration and U-Net with a focus on reproducibility," *Visual Computing for Industry, Biomedicine, and Art*, vol. 5, no. 1, p. 25, Oct. 2022, doi: 10.1186/s42492-022-00121-4.

[3] G. Hermosillo, C. Chefd'Hotel, and O. Faugeras, "Variational methods for multimodal image matching," *International Journal of Computer Vision*, vol. 50, no. 3, pp. 329–343, Dec. 2002, doi: 10.1023/A:1020830525823.

[4] J. B. A. Maintz, P. A. van den Elsen, and M. A. Viergever, "3D multimodality medical image registration using morphological tools," *Image and Vision Computing*, vol. 19, no. 1–2, pp. 53–62, Jan. 2001, doi: 10.1016/S0262-8856(00)00051-2.

[5] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, Apr. 1997, doi: 10.1109/42.563664.

[6] S. Bharati, M. Mondal, P. Podder, and V. B. Prasath, "Deep learning for medical image registration: a comprehensive review," *arXiv preprint arXiv:2204.11341*, 2022.

[7] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," in *International Workshop on Deep Learning in Medical Image Analysis*, 2017, pp. 204–212, doi: 10.1007/978-3-319-67558-9_24.

[8] Y. Hu *et al.*, "Label-driven weakly-supervised learning for multimodal deformarle image registration," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE, Apr. 2018, pp. 1070–1074, doi: 10.1109/ISBI.2018.8363756.

[9] Y. Hu *et al.*, "Weakly-supervised convolutional neural networks for multimodal image registration," *Medical Image Analysis*,

vol. 49, pp. 1–13, Oct. 2018, doi: 10.1016/j.media.2018.07.002.
[10] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[11] Y. Salini and J. HariKiran, "ViT: quantifying chest x-ray images using vision transformer & XAI technique," *SN Computer Science*, vol. 4, no. 6, p. 754, Sep. 2023, doi: 10.1007/s42979-023-02204-2.
[12] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: transformers for image recognition at scale," in *arXiv preprint arXiv:2010.11929*, 2021.
[13] H. Ramadan, D. El Bourakadi, A. Yahyaouy, and H. Tairi, "Medical image registration in the era of transformers: a recent review," *Informatics in Medicine Unlocked*, vol. 49, p. 101540, 2024, doi: 10.1016/j.imu.2024.101540.
[14] T. C. W. Mok and A. C. S. Chung, "Affine medical image registration with coarse-to-fine vision transformer," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2022, pp. 20803–20812, doi: 10.1109/CVPR52688.2022.02017.
[15] G. Haskins *et al.*, "Learning deep similarity metric for 3D MR–TRUS image registration," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 3, pp. 417–425, Mar. 2019, doi: 10.1007/s11548-018-1875-7.
[16] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, and N. Komodakis, "A deep metric for multimodal registration," in *International conference on medical image computing and computer-assisted intervention*, 2016, pp. 10–18, doi: 10.1007/978-3-319-46726-9_2.
[17] O. Zettinig *et al.*, "Multimodal image-guided prostate fusion biopsy based on automatic deformable registration," *International Journal of Computer Assisted Radiology and Surgery*, vol. 10, no. 12, pp. 1997–2007, Dec. 2015, doi: 10.1007/s11548-015-1233-y.
[18] Y. Wang *et al.*, "Multimodal registration of ultrasound and MR images using weighted self-similarity structure vector," *Computers in Biology and Medicine*, vol. 155, p. 106661, Mar. 2023, doi: 10.1016/j.compbiomed.2023.106661.
[19] M. Blendowski, N. Bouteldja, and M. P. Heinrich, "Multimodal 3D medical image registration guided by shape encoder–decoder networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 2, pp. 269–276, Feb. 2020, doi: 10.1007/s11548-019-02089-8.
[20] X. Song *et al.*, "Cross-modal attention for multi-modal image registration," *Medical Image Analysis*, vol. 82, p. 102612, Nov. 2022, doi: 10.1016/j.media.2022.102612.
[21] P. Lorenzen, M. Prastawa, B. Davis, G. Gerig, E. Bullitt, and S. Joshi, "Multi-modal image set registration and atlas formation," *Medical Image Analysis*, vol. 10, no. 3, pp. 440–451, Jun. 2006, doi: 10.1016/j.media.2005.03.002.
[22] K. Ma *et al.*, "Multimodal image registration with deep context reinforcement learning," in *International Conference on Medical image computing and computer-assisted intervention*, 2017, pp. 240–248, doi: 10.1007/978-3-319-66182-7_28.
[23] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 712–721, 1999, doi: 10.1109/42.796284.
[24] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, IEEE, Oct. 2016, pp. 565–571, doi: 10.1109/3DV.2016.79.
[25] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1, no. 2. Cambridge: MIT press Cambridge, 2016.
[26] I. Donaldson *et al.*, "MP33-20 the smart target biopsy trial: a prospective paired blinded trial with randomisation to compare visual-estimation and image-fusion targeted prostate biopsies," *Journal of Urology*, vol. 197, no. 4S, pp. e425--e425, Apr. 2017, doi: 10.1016/j.juro.2017.02.1016.
[27] L. University College, "SmartTarget - a magnetic resonance image to ultrasound fusion system for targeted prostate intervention: biopsy." [Online]. Available: https://clinicaltrials.gov/study/NCT02341677
[28] N. Ghavami *et al.*, "Automatic slice segmentation of intraoperative transrectal ultrasound images using convolutional neural networks," in *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, SPIE, Mar. 2018, doi: 10.1117/12.2293300.

BIOGRAPHIES OF AUTHORS

Hanae Mahmoudi obtained a master's degree in business intelligence and intelligent vision from Sidi Mohamed Ben Abdellah University, Morocco, in 2020. Now she is a Ph.D. student at Sidi Mohammed Ben Abdellah University, Morocco, 2020. Her research interests include artificial intelligence and medical imaging. She can be contacted at email: hanae.mahmoudi@usmba.ac.ma.

Hiba Ramadan received her Ph.D. degree in 2018 from the University Sidi Mohamed Ben Abdellah (USMBA), Fez, Morocco. Currently, she is an assistant professor at Sidi Mohamed Ben Abdellah University Morocco. Her research interests are in video analysis, in medical imaging, in visual information retrieval and pattern recognition. She can be contacted at email: hiba.ramadan@usmba.ac.ma.

Jamal Riffi is a professor of computer science at University Sidi Mohamed Ben Abdellah, Fez, Morocco. He is a member of the LISAC Laboratory. He specializes in data mining and deep learning. His main research fields are image mining and medical image analysis, neural network architectures and text mining and cybersecurity. He can be contacted at email: riffi.jamal@gmail.com.

Hamid Tairi received his Ph.D. degree in 2001 from USMBA, Morocco. In 2002 he has been a postdoc in the image processing group of the laboratory (LE2I) in France. Since 2003, he has been an associate professor at USMBA, where he obtained his HDR in 2009. He is currently the head of the laboratory (LISAC). His current research interests include advanced medical information processing, and spatial and functional information modeling and analysis in medical images, with applications in various medical domains including gastroenterology and ophthalmology, machine learning, biometric authentication and data fusion. He can be contacted at email: hamidtairi@gmail.com.