2D/3D Deep Registration Along Trajectories With Spatiotemporal Context: Application To Prostate Biopsy Navigation

Tamara Dupuy, Clément Beitone, Jocelyne Troccaz, Sandrine Voros

2D/3D Deep Registration Along Trajectories With Spatiotemporal Context: Application To Prostate Biopsy Navigation

Tamara Dupuy, Clément Beitone, Jocelyne Troccaz *Fellow, IEEE* and Sandrine Voros

***Abstract*— *Objective:* The accuracy of biopsy targeting is a major issue for prostate cancer diagnosis and therapy. However, navigation to biopsy targets remains challenging due to the limitations of transrectal ultrasound (TRUS) guidance added to prostate motion issues. This article describes a rigid 2D/3D deep registration method, which provides a continuous tracking of the biopsy location w.r.t the prostate for enhanced navigation. *Methods:* A spatiotemporal registration network (SpT-Net) is proposed to localize the live 2D US image relatively to a previously aquired US reference volume. The temporal context relies on prior trajectory information based on previous registration results and probe tracking. Different forms of spatial context were compared through inputs (local, partial or global) or using an additional spatial penalty term. The proposed 3D CNN architecture with all combinations of spatial and temporal context was evaluated in an ablation study. For providing a realistic clinical validation, a cumulative error was computed through series of registrations along trajectories, simulating a complete clinical navigation procedure. We also proposed two dataset generation processes with increasing levels of registration complexity and clinical realism. *Results:* The experiments show that a model using local spatial information combined with temporal information performs better than more complex spatiotemporal combination. *Conclusion:* The best proposed model demonstrates robust real-time 2D/3D US cumulated registration performance on trajectories. Those results respect clinical requirements, application feasibility, and they outperform similar state-of-the-art methods. *Significance:* Our approach seems promising for clinical prostate biopsy navigation assistance or other US image-guided procedure.**[1]

***Index Terms*— 2D/3D registration, prostate biopsy, spatiotemporal context, ultrasound-guided interventions**

I. Introduction

A. Clinical context and motivations

THE accurate localization of anatomical targets and navigation towards them is critical to many clinical tasks, especially for image-guided interventions on soft-tissue. A typical case is prostate biopsy, which is the confirmatory examination to diagnose prostate cancer, one of the most frequent male cancers worldwide. A biopsy session consists in collecting several tissue samples in the prostate, most often under transrectal ultrasound (TRUS) guidance. Most standard protocols consist of 6 or 12 sampling positions spread regularly in the prostate. Additional samples can be taken in specific regions appearing suspicious on a pre-operative MRI. During this TRUS-guided procedure, the navigation to biopsy targets faces several challenges coming from poor image quality and limited 2D anatomical information. The difficulty of mentally representing the 3D impact of US probe motions and prostate motion and deformation makes sample targeting very uncertain. This leads to a low correlation between theoretical biopsy samplings and histological results. Inaccurate sample targeting may thus produce biased diagnosis (about 30% of false negatives [1], [2]), inappropriate therapeutic decision and inaccurate focal therapy application, in addition to a longer and uncomfortable examination procedure for the patient. The smallest significant treatable tumor is mutually agreed among the urology community to a value of 0.5 cm^3 [3]. This corresponds approximately to the volume of a sphere of radius 5mm. Therefore, 5 mm can be considered as the maximum acceptable error between expected and real positions of biopsy cores.

In a context where prostate cancer diagnosis and therapy are at stakes, several developments have been made over the past two decades to improve biopsy cores targeting. A first kind of approaches focused on improving biopsy localization by using multimodal 3D/3D registration to map preoperative information [3], [4] to the intraoperative context. These methods generally register a preoperative 3D MR reference image to a reference 3D TRUS image, acquired at the beginning of the procedure. This is increasingly done using deep-learning based methods [5]–[9]. However, these approaches do not take into account intra-operative changes (due to probe pressure, patient breathing, etc.), which results in limited accuracy. To include updated information to this fusion process, other studies [10] proposed additional registrations made all along the intra-operative procedure: they register the initial 3D TRUS image to 3D TRUS images acquired prior to each biopsy gun

This work was supported in part by the French Agence Nationale de la Recherche, “Investissement d’Avenir” program (grants MIAI@Grenoble Alpes under reference ANR-19-P3IA-0003 and CAMI Labex under reference ANR-11-LABX-0004), by Région Rhône-Alpes (project ProNavIA) and by the PSPC project DIANA

T. Dupuy, C. Beitone, J. Troccaz and S. Voros are with the Univ. Grenoble Alpes, CNRS, INSERM, TIMC, F-38000, France.

Correspondance to: clement.beitone@univ-grenoble-alpes.fr

firing. Despite their accuracy and robustness, these approaches require 3D acquisition time (generally a few seconds) with a static position of the US probe. This makes real-time navigation assistance impossible.

This has motivated other approaches focused on computing the current localization of the organ and targets from real-time (a.k.a."live") 2D US images. This can be done by registering the live 2D US image either directly to the preoperative 3D MRI by using multimodal 2D/3D registration [11]–[13], or to a reference 3D TRUS image by using mono-modal 2D/3D registration [1], [14]–[19]. Nevertheless, real-time 2D/3D registration is still a challenging task due to image dimension mismatch, the lack of out-of-plane information from the 2D image, in addition to the spherical and symmetric shape of the prostate. For now, traditional (i.e. non deep-learning) methods [1], [11], [15]–[18] still have limited performances especially in terms of computational efficiency. Considering the 2D US image acquisition frequency (10-20 Hz), a computational time below 50 ms is needed to allow real-time navigation. Recent advances in deep learning-based strategies [12], [14], [19] provided a new opportunity to develop robust real-time guidance. Nevertheless, published studies still present limited evaluation strategies in terms of clinical realism. Indeed, most evaluations involved: data from small databases [12], artificially simulated data (resampled from TRUS volume without real-time deformation [14], [19], with a single slice orientation [14]), strongly preprocessed data (segmented [12], centered [14], manually initialized [12], [14]), or with limited ground-truth annotations (2D landmarks [14]). Moreover, these methods do not consider the overall dynamic of prostate biopsy samplings which rely on complete trajectory gestures. Most often, registration is evaluated for standalone and standard slices, without cumulative error consideration over time [14], [19]. Finally, even with poor validation strategies, these studies do not meet clinical accuracy requirements.

B. 2D/3D deep registration strategies for real-time navigation assistance

As only few studies refer to registration for prostate navigation applications, we have extended our literature analysis to any general application involving "2D/3D image registration". This includes both 2D image (referred also as slice) localization inside a volume and volume reconstruction from slices. We exclude registration involving 2D projective images (typically X-ray projections) as they do not share the same properties as our targeted application. The different studies can be organized into 4 categories, described below.

Methods considering only slices inputs: Unlike traditional methods, 2D/3D deep registration is often treated as a problem where only the slice to be registered is provided to the network, without having any information about the volume, even if it exists. Most of the literature [19]–[25] proposes an end-to-end supervised regression task, with convolutional neural networks (CNN) architectures, to predict rigid transformation parameters (rotations and translations) which locate the slice in a common 3D reference frame.

Methods considering spatial context inputs: Adding spatial context information can facilitate network learning especially for 2D/3D registration problems. Providing both slice and volume as inputs is used in few studies [12], [14], [26]. However, the correct combination of these two types of information is made complex due to their different dimensions, the different nature of the features involved, as well as their unequal and unbalanced representation. Most of these studies ultimately transform the problem into a same-dimensional task: (i) either by projecting the 2D features in the 3D space using external tracking (3D/3D problem [12]), (ii) by reslicing the wrapped volume (2D/2D problem [12]), or (iii) by manipulating the dimensions of the convolutions to extend the feature maps (by duplication [26] or transformation [14]) into a volume of the same dimensions as the input volume. One study [14] proposed a dual-branch balanced feature extraction network to make the model equally sensitive to both the frame and volume information. For volume reconstruction problems [27], [28], the addition of very local spatial context is often used through the addition of the previous slice as input. It contributes to the relative localization between successive slices.

Methods considering spatial context penalization: Another way of adding spatial context information consists in introducing loss-penalization terms during the training phase. Many studies [5], [6], [14], [29] use Normalized Cross Correlation (NCC) or Structural Similarity Index Measurement (SSIM) to estimate the similarity between the predicted localized image and the input slice. Such penalization allows preserving the anatomical structures coherence and realism based on global image content and allows a more relevant registration optimization. This is also referred as "self-context learning", as it relies on weak supervision using internal context, inspired by the iterative optimization of traditional studies. These loss-penalization computations rely on the implementation of differentiable re-sampler modules (Spatial Transform Network [30]), and also require the availability of this 3D space context: either by preoperative acquisition [14] or by immediate predicted reconstruction [29].

Methods with spatiotemporal context: Finally, spatiotemporal context can be given through the input of sequence of images. The predicted localization of each slice is weighted and conditioned by the general context of all other slices. Such slices are usually handled by using 3D convolution on several stacked neighboring slices [31], through dynamical structures like recurrent networks [29], [32] or attention networks [33]. We also consider as spatiotemporal context, any forms of additional information about prior localization. It can be represented either through previous results information [19] or relative probe tracking information [27].

C. Objectives and contributions

To allow navigation assistance during biopsy procedures, the main objective of this work is to present a real-time 2D/3D registration able to localize at each instant the "live" 2D US image relatively to a TRUS reference volume, acquired just before starting the navigation. This registration must satisfy clinical requirements, both in terms of accuracy ($\leq$ 5mm) and computational efficiency ($\leq$ 50ms), as previously mentionned.

Strongly motivated by the intrinsic spatiotemporal nature of prostate biopsy procedures, we developed in a preliminary work [19] a real-time 2D/3D registration with prior trajectory information using a 2D CNN architecture. The temporal part was based on previous registration results and probe tracking, and the spatial part was based on stacked pairs of successive images. We demonstrated that the addition of temporal information significantly improved the registration quality.

A more recent work [14], with the same clinical objectives, proposed to incorporate other forms of spatial context using both spatial input (a sub-volume) and spatial penalty (image similarity loss term). The given sub-volume inputs were centered and oriented using an initialization close to the ground-truth orientation. They proposed a dual-branch features extraction network, using a 3D CNN architecture for both 2D and 3D input. They showed that incorporating such balanced network and similarity loss reduced the registration error.

Given the potential of using other possible forms of spatial context demonstrated in [14], a first objective reported in this paper was to investigate the combination of our preliminary temporal context approach with new possible spatial context informations. For that, we propose a 3D CNN architecture with flexible configurations to conduct a complete ablation study and compare three different forms of spatial context inputs: (i) a neighbor slice, (ii) the reference volume, or (iii) a subpart of this volume. To further reinforce the impact of spatial context information, we also evaluated the benefit of adding a spatial loss penalization.

The second objective of the work was to establish a new protocol to validate an intra-operative navigation assistance for biopsy in a more clinically realistic way. To better reflect the impact that registration errors would have on an overall biopsy gesture, we first propose a cumulative evaluation based on successive registrations over a complete trajectory. Similar to drift quantification in volume reconstruction problems [27], this meaningful evaluation is, however, not provided in most similar previous studies ([14], [19]) despite their ineluctive temporal bias to drifted results. Secondly, while the same simplified data simulation is mostly used in the literature (2D reslicing and registration toward the same volume), we developed a new data generation process with an increased level of difficulty and realism. This new data simulation level permits to evaluate the generalization capabilities of the models while the first level of data simulation allows a fair comparison with published methods.

The next section introduces our proposed method with the different spatiotemporal context addition strategies as well as their related implementation details. Then, data generation and evaluation protocols are described. Finally, we present the experimental results, followed by a discussion.

II. Materials and Methods

A. Spatiotemporal framework for registration

The proposed network estimates the 2D/3D rigid transform to localize continuously the current US slice S_t with respect to the reference volume V_{ref}. This volume is acquired just before starting the navigation towards the next sampling position. The predicted transform is denoted as $\widehat{T}_t = (tx, ty, tz, \theta x, \theta y, \theta z)$ and is composed of translation and rotation parameters along the three axes. Fig.1 illustrates the proposed spatiotemporal framework which includes several blocks: a main backbone branch (block 1), temporal inputs (block 2), a parallel branch with different forms of spatial inputs (block 3) and an additional spatial penalization module (block 4).

1) Main backbone branch: The main input of this network is the US slice S_t which goes through the main branch where a 3D-conv backbone is used. It relies on a 2D convolution layer where the features channel number is chosen using a defined extension parameter E. This parameter has a different role depending on the additional spatial input (see section II-B). Then, successive 3D convolutional blocks are applied to better combine multiple low-level features.

2) Temporal context input: Temporal information consists in prior registration results and relative motion information. Prior trajectory information (T_{prior}) comes from the previous predicted registration result $\widehat{T}_{t-1}$ computed for the localization of S_{t-1}. Probe tracking information (T_{PT}) is related to the relative displacement between S_{t-1} and S_t, e.g. measured by inertial probe sensor attached to the US probe. These optional inputs are concatenated separately as two 6×1 vectors to the 512-vector layer of the network. Section II-C describes in more details how these input are computed or simulated.

3) Different spatial context input forms: We evaluate the benefit of adding 3 different forms of spatial context inputs: local, global, and partial. Each of these spatial inputs will be added separately through different experiments.

Local spatial input: The previous slice S_{t-1} is directly concatenated to S_t in a 2-channel image input. It allows to couple the spatial context in the two slices and explore their 3D dependencies using convolutional layers of the main backbone branch.

Global spatial input: The reference volume is provided as input to the network through an additional and parallel branch (block 3 "Full-volume"). This branch is composed of the same successive 3D-conv blocks as the main branch to obtain a balanced number of features during the later concatenation.

Partial spatial input: The last tested spatial form consists in using a subpart of the reference volume as input through an additional and parallel branch (block 3 "Sub-volume"). This sub-volume is computed and initially oriented using the previous predicted transform $\widehat{T}_{t-1}$. This restricts the search space, centered around the previous slice. In this case, some changes are needed in the rest of the network: as temporal information is already used to extract the new oriented sub-volume input, the additional T_{prior} and probe tracking inputs are removed from block 2.

4) Spatial context penalization: Finally, spatial context information can be strenghened using an additional and differentiable spatial penalization module (block 4). This module consists of a rigid resampler, which first transforms the reference volume (fed as input or not) to the predicted transformation, before reslicing the corresponding image $S(V_{ref}, \widehat{T}_t)$ (see [30] for technical details). Finally, a similarity measure is computed between the predicted resliced image pixels and the input S_t pixels. This SSIM-loss term is then weighted and added as

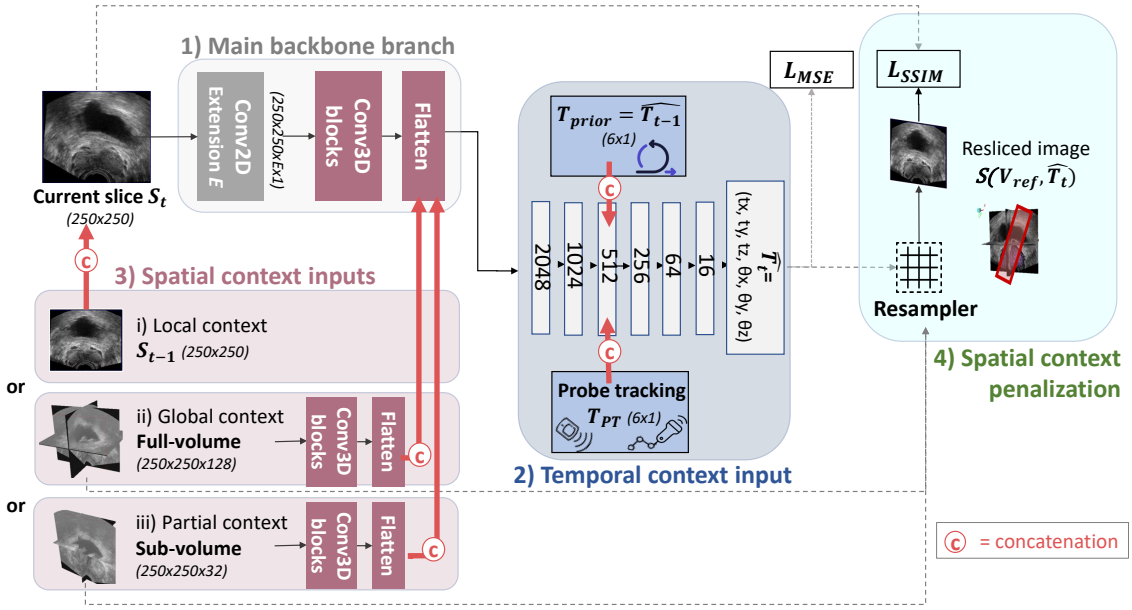

Fig. 1: Proposed method with spatiotemporal context. The framework can be separated in several blocks: a main backbone branch (block 1), a parallel branch with different forms of spatial context inputs (block 3), a spatial context penalization module (block 4), and temporal context inputs (block 2).

loss term, only for experiments adding this SSIM module (see section II-B).

B. Network architecture and training

This section aims at detailing how the network architecture visualized in Fig.1 has been implemented. The main branch is composed of a first extended 2D-convolutional layer (output:250x250xEx1), where the defined "extension parameter" E corresponds to the features channel number. To match the spatial context size of the parallel branch, in partial and global experiments, E must adapted to the input depth size: $E = 128$ for full-volume input or $E = 32$ for sub-volume input. For capturing 3D dependencies between neighbor slices in the local experiment, we chose a default extension parameter $E = 128$ for which we achieved good performance. Then, two 3D-convolutional blocks with increased number of filters are applied. Each one contains two successive 3D-convolutional layers (kernel=5, strides=2, activation=ReLu), followed by max pooling layers (kernel=2, strides=1). Finally, the network consists in 6 fully connected layers, with a decreasing number of neurons to connect with the final output layer.

Two loss functions can be combined for training. A supervised mean squared error (MSE) loss (eq. 1) is computed directly between the network predictions $\widehat{T_t}$ and the ground-truth transform parameters T_t. A weakly supervised image similarity metric (SSIM) is computed between predicted resliced image from the input reference volume $S(V_{ref}, \widehat{T_t})$ and the input slice S_t, as defined in eq. 2. For the linear combination of these two terms, we conducted comparison experiments and we finally favored a fixed weighting. (eq. 3). Contrary to adaptive weighting along the training, this fixed ratio allows focusing on the image similarity term only after several iterations, once MSE has penalized over huge errors, allowing getting two close images, where the similarity comparison is much more meaningful. For all experiments including the SSIM-penalization block, we achieved good performance using $\alpha = 1$ and $\beta = 50$, while for the others, we used $\alpha = 1$ and $\beta = 0$.

$$L_{MSE} = \frac{1}{n}\sum_{t=1}^{n} ||(T_t - \widehat{T_t})||^2 \quad (1)$$

$$L_{SSIM} = \frac{1}{n}\sum_{t=1}^{n} ||(S_t - S(V_{ref}, \widehat{T_t})||^2 \quad (2)$$

$$L_{tot} = \alpha * L_{MSE} + \beta * L_{SSIM} \quad (3)$$

C. Data generation process

1) Available clinical data: All the data were collected during routine prostate biopsy exams, assisted by a US-based guidance platform (Urostation® and Trinity® from Koelis SAS), performed by urologists from the Grenoble University Hospital (agreement MR2711140520 from CNIL, French Authority for Data Management). Several elements are available for each examination as illustrated in Fig. 2: (i) The "panorama" US volume (V_{pano}) taken at the beginning of the examination with its associated prostate surface mesh, (ii) intra-operative biopsy US volumes (V_i) taken at each biopsy site $i = (1, j, ...N)$ all along the procedure, and (iii) the rigid transform ($T_{V_i \to V_{pano}}$) obtained through 3D/3D organ-based rigid registration between each V_i and the V_{pano} [10]. Finally, rigid registration between any pair of 3D biopsy volumes ($T_{V_i \to V_j}$) can thus be deduced from these available global localizations toward the panorama volume, as described in Fig. 2.

2) Data simulation: To simulate the 2D US flow used to navigate during a clinical procedure, series of successive 2D US images are resliced from these biopsy volumes. For that,

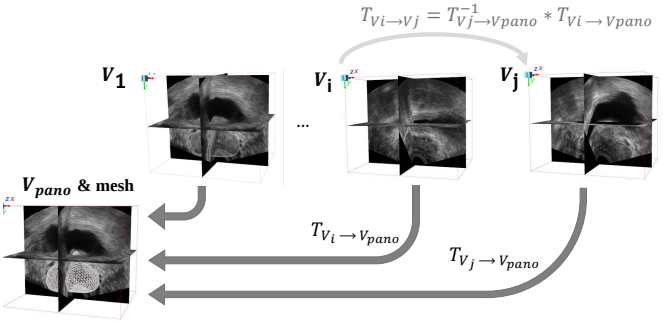

Fig. 2: Volume localizations from available clinical data.

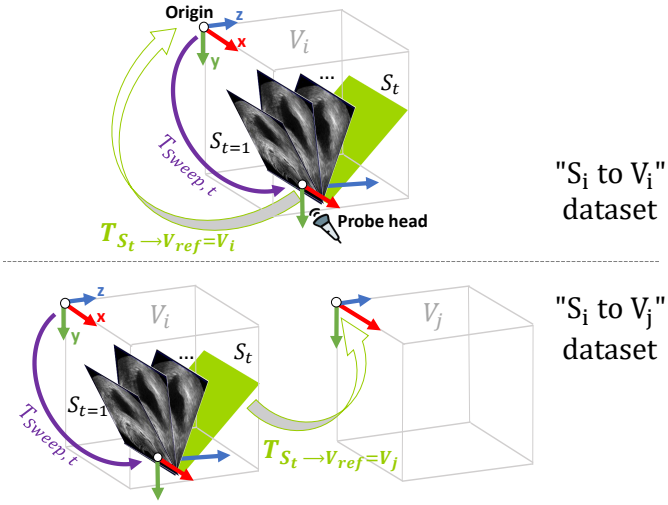

Fig. 3: Data generation process. Slices are always referred to be extracted from V_i. Two datasets are created according to the ground-truth slice localization w.r.t either V_i or V_j.

sweep trajectories ($T_{Sweep,t}$), corresponding to several back-and-forth base-to-apex sweeps (from upper to lower prostate extremities), are mimicked to generate the successive slices ($S_{t=1}, S_{t=2}, \dots S_{t=n}$). These trajectories are obtained by virtually rotating the image plane around the probe head ($\pm 40°$ range around $x-axis$) and the transforms are computed relatively to the US volume's referential origin, as illustrated in Fig. 3. Finally, obtained slices are cropped into 250×250 pixels (with isotropic pixels spacing of 0.306 mm), and we guarantee a realistic acquisition frequency around 20 Hz (50 ms) between two successive slices. During this process, the associated reference prostate mesh is transformed in the same way, and saved at each new time step t, for further use during the evaluation process (see section II-D.2).

The ground-truth slice localization ($T_t = T_{S_t \to V_{ref}}$) is then computed relatively to our defined "reference volume" V_{ref}. This reference volume is a concept corresponding to the volume acquired just before starting the navigation, and to which we want to localize. For simplicity, slices are always referred to be extracted from V_i during data generation, while the choice of V_{ref} depends on the desired level of complexity we want to simulate: either $V_{ref} = V_i$ (as often performed in the literature) or $V_{ref} = V_{j \neq i}$. These two different data generation processes are referred respectively as "$S_i\, to\, V_i$" or "$S_i\, to\, V_j$" datasets (see Fig. 3). While the former dataset presents image dependencies and bias that do not represent the full complexity of real cases, the latter one includes input slices that are now independent and anatomically different from the reference volume to which we want to register them. The slice ground-truth localization is computed through the composition of several rigid transforms, as described in eq.4.

$$T_t = T_{S_t \to V_{ref}} = T_{Vi \to V_{ref}} * T^{-1}_{sweep,t} \text{ with:} \quad (4)$$

$$T_{Vi \to V_{ref}} = \begin{cases} T_{identity}, & \text{if } V_{ref} = V_i \\ T^{-1}_{Vj \to V_{pano}} * T_{Vi \to V_{pano}}, & \text{if } V_{ref} = V_j \end{cases}$$

In particular, for experiments using partial spatial input, as the given input volume for each time t is now a subpart of the reference volume (defined as $V_{sub,t}$), the ground-truth slice localization is adapted. Further description of $V_{sub,t}$ computation and ground-truth localizations are described in supplementary materials.

Temporal inputs must also be simulated. We use previous slice ground-truth localization T_{t-1} to simulate prior trajectory information (T_{prior}), during training only. For inference, however, previous predictions ($\widehat{T_{t-1}}$) of the network are directly used (see section II-D.1). Relative displacements between two successive images (T_{PT}) are computed using two successive global displacements (${T_t}^{-1} * T_{t-1}$) to which we add a maximum random noise of 1°. This simulates an inertial measurement unit with a realistic average sensor noise [34].

Finally, the given input reference volumes are cropped into different sizes according to the experimented spatial context condition: $250 \times 250 \times 128$ voxels for full-volumes, or $250 \times 250 \times 32$ for sub-volumes.

D. Evaluation protocol

1) Cumulative evaluation on trajectory: Because the effect of error accumulation might be detrimental to any navigation assistance, especially when using temporal inputs based on previous predictions (T_{prior}), series of registrations along trajectories must be evaluated. To provide a realistic drift evaluation, we simulated a complete navigation timeline, where predictions are made one by one, iteratively, using previous predictions results ($T_{prior} = \widehat{T_{t-1}}$).

Moreover, we mimicked a clinical navigation assistance by portion, between each biopsy site, using the clinically available 3D/3D registration (integrated in clinical workstations [10]) as drift re-alignments. We divided our simulated and simplified base-to-apex sweep trajectories into 12 subparts to reproduce *pseudo* 12-biopsy samplings scheme. The start of this series of previous predictions relies on simulated noisy ground-truth ($\widehat{T_0} \approx T_0 \pm 1°$). Then, at each new *pseudo* biopsy along the trajectory, T_{prior} is reset using the ground-truth transform towards the reference volume or sub-volume (eq.4).

2) Evaluation metrics: We evaluated the registration quality from different points of views: feature-based analysis using anatomic and geometric correspondence; image-based analysis using similarity of image content; and regression-based

analysis using statistical comparison of results to the ground-truth.

Regarding geometric error, Target Registration Error (TRE) is a frequently used metric in image registration to measure the impact of the registration quality directly on the organ of interest in which clinical targets are located [35]. The metric is also particularly useful because of its simplicity, fast computation, and ability to compare to the literature [36]. This measure is computed by comparing point pairs, most often fiducial landmarks, manually defined and localized by experts, and independently chosen from the registration process. In the presented study, we use prostate surface mesh points, provided by the clinical database for each patient (see section II-C). Such mesh-based evaluation allows computing a robust TRE, first due to the thousands of points available (far more numerous than internal fiducials), and secondly due to their 3D nature enabling to consider the registration impact on the whole prostate gland (rather than on few local fiducials). Being given prostate mesh points $p_{k=(1,\dots,N)}$, we compute the TRE (averaged for all time steps t) between the point set transformed by the estimated transform $\widehat{T}_t(p_k)$ and the point set transformed by the ground-truth transform $T_t(p_k)$ (eq.5). It is important noticing that TRE gives an estimate of the standard deviation of the normal distribution of predicted biopsy position around the real one. According to clinical requirements, let us remind that this error must not exceed 5 mm (see section I-A). To obtain about 95% of the predicted position conveniently located within the 5 mm limit (2 standard deviations), the desired TRE value to meet registration accuracy requirements is 2.5 mm.

$$TRE(mm) = \sqrt{\frac{\sum_{k=1}^{N} ||(T_t(p_k) - \widehat{T}_t(p_k))||^2}{N}} \quad (5)$$

Normalized cross-correlation (NCC) is a metric commonly used to estimate pixel similarity score between two images. It enables to have quality assessment of the registration results different from the SSIM, already used during the training optimization. This measure is computed between predicted resliced image and the corresponding input image.

Regression analysis requires quantitatively and statistically comparing predictions to labels. For that, we use the coefficient of determination (R^2) which captures how well predictions match their expectations without having interpretability limitation or bias [37].

Finally, all these metrics are computed during the proposed cumulative evaluation, over a complete trajectory.

III. Experimental setup

A. Experiments

1) Ablation study: To determine independently whether a module improves the registration results or not, we performed an ablation study with several combinations of modules. The different compared forms of spatial context input are: local (slice S_{t-1}), partial (sub-volume), global (full-volume), respectively referred as *"local"*, *"part"*, *"glob"*. Finally, the SSIM-penalization module is added to each previous experiment to have a complete ablation study. We refer to them as *"local+SSIM"*, *"part+SSIM"* and *"glob+SSIM"*.

Besides, as our preliminary approach [19] already justified the benefit of T_{prior} in temporal context, such input is kept intact for the complete ablation study. However, as probe tracking requires an inertial sensor to measure the probe motions that may not be available in some clinical set-ups or in other state-of-the-art methods, we simulated two scenarii: with or without such input. They are respectively referred as *(Scenario Im+PT)* and *(Scenario Im)*.

Finally, as such temporal inputs (either previous localization results T_{prior} or probe tracking T_{PT}) can be strong additional information, a baseline assessment **"without network"** is also reported by computing the image position using only the geometric transformation T_{prior} or $T_{prior} * T_{PT}$. Such comparison allows demonstrating the network contribution based on all combined inputs and not only temporal inputs.

2) Comparison study : In the literature, two methods are close to ours regarding their (i) objectives: navigation assistance, (ii) clinical application: prostate biopsy, (iii) methodology: deep rigid mono-modal (US) 2D/3D registration. The following section described how we properly compare to them.

The **"FVR-Net"** refers to Guo's work [14]. A comparison with their work required a re-implementation of their code (available on GitHub), as well as a complete re-evaluation using the database and experimental conditions described in sections II-C and II-D.1. Indeed, contrary to the original work, we did not compute the inputed sub-volumes using the noisy ground-truth transform, but using the previous estimation ($\widehat{T_{t-1}}$) for a more realistic cumulative evaluation on complete trajectory. Moreover, experiments are tested over both $S_i\,to\,V_i$ and $S_i\,to\,V_j$ dataset levels, while the original paper presented only results for the first case.

Another study to compare with, is our preliminary work [19], referred as **"Pre-Net"**. In this former work, the database was constructed using a biopsy simulator [38] where no translation parameters could be simulated, and without any cumulative evaluation. However, the data generation involved slices in complex arbitrary orientations based of biopsy trajectories which is not easily comparable to the rest of the literature. The proposed comparison has required testing this former version of the network with the new data generation process and using the new cumulative evaluation on trajectory. This allows evaluating our previous method in a more realistic way and by having 3 more transformation parameters to predict. Moreover, among several proposed scenarii described [19], we selected the most successful (including prior registration results, probe tracking, and local spatial context) to compare with. From a methodological point of view, this scenario is mainly similar to our proposed local experiment, except this fomer work used 2D CNN architecture. Thus, the comparison allows evaluating the benefit of using our new proposed 3D convolutional structure.

B. Datasets

We simulated sweep trajectories (as described in section II-C.2) that we applied to a large database composed of

600 TRUS volumes, coming from biopsy sessions of 100 different patients. The datasets for the two different studies (III-A.1 and III-A.2) are structured in two different ways. For the ablation study, each tested experiment is trained, validated and tested on the same dataset splits, summarized in supplementary materials (Table 1). For the comparison study (with a more manageable number of experiences), we used a 5-fold validation strategy to increase the reliability of the results. The different folds are divided according to the number of patients (and not samples), and are described in supplementary materials (Table 2).

C. Settings

All the experiments were performed using the Nadam optimizer with an initial learning rate of 1e-4, allowing to obtain robust and rapid convergence. The weights of the network were all initialized with a Gaussian distribution (mean=0, std=0.01). The network was trained for 30 epochs (between 13000 and 42000 iterations) with batch size of K=50, 30, or 16 depending on the experiment input size (local, partial, global spatial context respectively). For higher performance training on the $S_i\,to\,V_j$ dataset, we applied transfer learning by reusing pre-trained model from $S_i\,to\,V_i$ experiments. The model was trained on A100 GPU from NVIDIA and using TensorFlow.

IV. Results

A. Ablation study

Table I provides the values of cumulative TRE, before and after registration over all the experiments, for the two scenarii considered, and for the two datasets ("$S_i\,to\,V_i$" and "$S_i\,to\,V_j$"). Non-parametric Wilcoxon test (alpha = 0.05) was performed on each paired experiments (column) and demonstrated significatively different distributions.

Both "$S_i\,to\,V_i$" or "$S_i\,to\,V_j$" datasets result in the same trends and conclusion. The best-case scenario is obtained through the local experiment, with probe tracking information (Sc. Im+PT). For "$S_i\,to\,V_i$", the TRE is corrected from an initial mean error of 9.52 ± 6.49 mm to a final error of 0.21 ± 0.28 mm. Whereas for "$S_i\,to\,V_j$", the final TRE is larger (about 2.68 ± 1.49), because of the higher difficulty level of the task including prostate deformations between images.

Both global or partial spatial information do not allow satisfactory improvements on the registration results, for any of the scenarii. Besides, the partial experiment results in even poorer registration quality compared to the global one.

Adding a spatial penalty does not seem to improve registration quality (similar range of results than the configuration without it), regardless of the spatial input being tested. Those results are still less accurate than the local experiment.

Finally, all experiments demonstrate better results compared to the baseline assessment ("without network"), suggesting a good network contribution and a good processing between all kinds of inputs.

Both Fig. 4-A and 4-C illustrate parameters evolution over a trajectory. Base-to-apex motions (around $x-axis$) are well illustrated in the graph of "$S_i\,to\,V_i$" dataset (Fig.4-A). Let us note that the variations of the translation parameters come from the offset between the probe head rotation center and the volume's referential origin (see section II-C). For "$S_i\,to\,V_j$" graph (Fig.4-C), parameter motions come from composition of base-to-apex motions and transform toward another volume (see eq. 4). Fig. 4-B and 4-D display the associated TRE evolution over the same trajectory, before and after registration, demonstrating an efficient error correction.

The results on "$S_i\,to\,V_i$" dataset show a good trajectory reconstruction without any drift and a good TRE decrease thanks to registration. The range of each parameter value is well respected, even for out-of-plane parameters ($\theta x, \theta y, tz$) and for parameters without expected variation. On "$S_i\,to\,V_j$" dataset, however, a drift is more visible between predicted and ground-truth evolution. The 12 T_{prior} resettings are clearly observed over the trajectory and illustrate a drift control between two successive biopsy sites.

B. Comparison study

Table II summarizes the comparison against other methods: FVR-Net and Pre-Net. "SpT-Net" refers to our best proposed spatiotemporal network (local, Scenario Im+PT). The presented results come from a complete re-evaluation on our two datasets, using a 5-fold validation strategy, and with the proposed cumulative evaluation (see II-D.1). We also reported the average running time per input slice, as well as the networks' number of parameters.

Both studies benefitting from T_{prior} and relative probe tracking, Pre-Net (preliminary work) and SpT-Net (current work), seem more accurate than FVR-Net which relies only on sub-volume input and SSIM-penalization. Moreover, even our scenario without probe tracking (Sc.Im, local, Table I) performs similarly to FVR-Net (Table II), which demonstrates competitive results of our method. Compared to Pre-Net, our new model benefitted from a new 3D convolution and achieved better results in both datasets, suggesting a more appropriate network architecture.

To conclude, our best selected scenario of SpT-Net outperforms by far the rest of the literature methods, on the two datasets evaluated. For both datasets, FVR-Net results are far from the clinical requirements expected in terms of accuracy and computational time. This may be due to their sub-volume generation (with an orientated initialization) and due to costly operations involved in their complex network architecture. Finally, our SpT-Net meets clinical conditions for the simple dataset and is close to them in the more complex one.

C. Qualitative results

Fig. 5 illustrates the registration quality directly on US images, by comparing the obtained images (after predicted registration) to the ground-truth image, through pixels difference. The ground-truth image $S(V_{ref}, T_t)$ is obtained after reslicing V_{ref} using ground-truth registration T_t. This computation is needed for the simulated "$S_i\,to\,V_j$" dataset, as the input image S_t (simulated from V_i) may present different anatomical prostate shapes compared to ground-truth image. The following rows show the resliced image from

TABLE I: Performances of the proposed SpT-Net with ablation studies: over several forms of spatial context, across the different scenarii, and for the two tested datasets. The different spatial context forms are: local slice, partial sub-volume, or global full-volume (respectively referred as *"local/part/glob"*). *"+SSIM"* referred to the addition of SSIM-penalization module.*"without network"* referred to the baseline assessment computing registration with only temporal input. The two tested scenarii are *Scenario Im+PT* and *Scenario Im*, respectively with/without probe tracking.

Cumulative evaluation on S_i to V_i	TRE before	TRE after local	glob	part	local+SSIM	glob+SSIM	part+SSIM
Scenario Im							
SpT-Net (with $T_{prior} = \widehat{T_{t-1}}$)	9.52 ± 6.49	2.74 ± 6.04	3.83 ± 7.11	4.57 ± 6.42	2.91 ± 6.00	3.43 ± 6.71	4.69 ± 6.43
Without network (only T_{prior})		5.30 ± 6.32	6.17 ± 7.23	5.43 ± 5.89	5.37 ± 6.28	5.88 ± 6.88	5.55 ± 5.75
Scenario Im+PT							
SpT-Net (with $T_{prior} = \widehat{T_{t-1}}$)	9.52 ± 6.49	**0.21 ± 0.28**	**0.21 ± 0.19**	6.43 ± 6.30	0.62 ± 0.95	0.52 ± 0.55	7.13 ± 7.19
Without network (only $T_{prior} * T_{PT}$)		1.01 ± 2.23	1.00 ± 2.24	6.42 ± 6.32	1.33 ± 2.30	1.24 ± 2.23	7.09 ± 7.22

Cumulative evaluation on S_i to V_j	TRE before	TRE after local	glob	part	local+SSIM	glob+SSIM	part+SSIM
Scenario Im							
SpT-Net (with $T_{prior} = \widehat{T_{t-1}}$)	11.72 ± 6.28	5.06 ± 5.70	8.87 ± 5.82	5.01 ± 7.82	5.80 ± 7.00	7.30 ± 7.25	7.28 ± 8.50
Without network (only T_{prior})		7.70 ± 6.24	9.65 ± 6.11	6.35 ± 6.88	5.91 ± 6.77	9.39 ± 7.65	7.63 ± 7.67
Scenario Im+PT							
SpT-Net (with $T_{prior} = \widehat{T_{t-1}}$)	11.72 ± 6.28	**2.68 ± 1.49**	2.81 ± 2.00	5.28 ± 6.30	3.97 ± 2.03	3.15 ± 2.17	5.91 ± 7.19
Without network (only $T_{prior} * T_{PT}$)		3.11 ± 2.67	3.15 ± 2.94	5.23 ± 6.32	4.28 ± 2.93	3.53 ± 3.10	5.84 ± 7.22

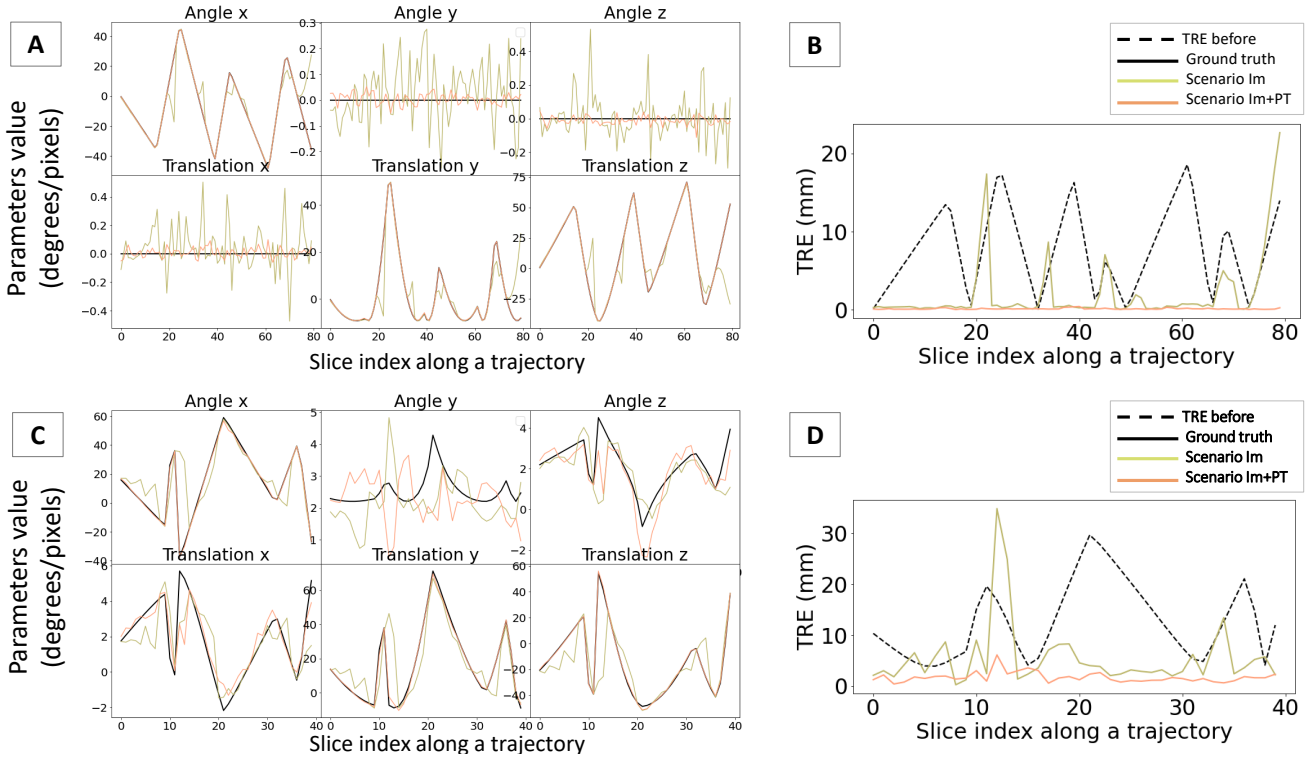

Fig. 4: Temporal analysis over a complete trajectory example: for local experiment, on both datasets. Results on "S_i *to* V_i" dataset are illustrated on A-B, whereas "S_i *to* V_j" results are illustrated on C-D. (A, C) Angle and translation parameters evolution over trajectory: ground-truth (black) and predicted (color) evolution according to the different scenarii. (B, D) TRE evolution over trajectory: before any registration (black), and after registration with different scenarii (color).
N.B: (x,y) is the US probe plane and z the out-of-plane axis.

V_{ref} before any registration $S(V_{ref}, Id)$, and after predicted registration $S(V_{ref}, \widehat{T_t})$ depending on the selected method: respectively SpT-Net (best), FVR-Net, or Pre-Net. The last column illustrates the overlapping between the two compared prostate mesh: the ground-truth prostate (green), and the predicted prostate after registration (in red dotted line). This allows to better visualize the registration error directly on prostate anatomy, and better interpret its impact for the clinical application. Finally, the associated TRE and NCC measures, for each experiment, are also reported.

The chosen example illustrates a typical case with TRE value close to our obtained mean results (Table II). We observe

TABLE II: Performance comparison with state-of-the art methods after complete re-evaluation of the method with the proposed cumulative evaluation and on our datasets *(re-evaluated)*. A 5-fold cross validation strategy was used. The average initial error before any registration are similar for each experiment and are reported at the top. p-values between state-of-the-art methods and SpT-Net (on both NCC and TRE values) are also reported.
N.B: R^2 score can not be computed for non varying parameters (namely $tx, \theta y, \theta z$ in the "$S_i\,to\,V_i$" dataset).

Methods ($S_i\,to\,V_i$)	**TRE (mm)** before at 9.07 ± 6.55	**NCC** before at 0.84 ± 0.12	**Coef. of Determination R^2** ($tx, ty, tz, \theta x, \theta y, \theta z$)	**RunTime** (s)	**Hyperparameters** (M)	**p-value** (vs SpT)
Our SpT-Net (best)	**0.19 ± 0.77**	**0.99 ± 0.04**	**(∅, 0.99, 0.99, 0.99, ∅, ∅)**	0.06 ± 0.02	71	
FVR-Net (re-evaluated)	2.28 ± 8.52	0.95 ± 0.11	(∅, 0.84, 0.80, 0.81, ∅, ∅)	0.17 ± 0.06	68	$p \leqslant 0.001$
Pre-Net (re-evaluated)	0.26 ± 0.76	0.98 ± 0.03	(∅, 0.99, 0.99, 0.99, ∅, ∅)	0.03 ± 0.01	141	$p \leqslant 0.001$
Methods ($S_i\,to\,V_j$)	**TRE (mm)** before at 11.14 ± 6.13	**NCC** before at 0.81 ± 0.11	**Coef. of Determination R^2** ($tx, ty, tz, \theta x, \theta y, \theta z$)	**RunTime** (s)	**Hyperparameters** (M)	**p-value** (vs SpT)
Our SpT-Net (best)	**2.21 ± 2.42**	**0.93 ± 0.09**	**(0.98, 0.98, 0.99, 0.99, 0.98, 0.94)**	0.06 ± 0.02	71	
FVR-Net (re-evaluated)	4.36 ± 12.18	0.85 ± 0.16	(0.83, -0.78, 0.59, 0.50, -2.99, -10)	0.17 ± 0.06	68	$p \leqslant 0.001$
Pre-Net (re-evaluated)	2.26 ± 3.16	0.81 ± 0.12	(0.94, 0.98, 0.99, 0.99, 0.97, 0.89)	0.03 ± 0.01	141	$p \leqslant 0.001$

very similar image results for SpT-Net compared to other methods, with a global mesh overlapping. The errors on the "$S_i\,to\,V_j$" dataset are larger due to the difficulty of including prostate deformations between images.

V. Discussion

A. Methodological contribution

1) Adding spatial context through inputs and/or penalization: It seems that adding spatial context information through input volumes does not benefit the registration task, indicating that the network learns little information from these additional inputs. Although the features balance is guaranteed by the network architecture, other challenges may hinder the performances, such as: a bigger searched space, more parameters to optimize (compared to a single branch network), as well as a different nature of features dimension (2D or 3D). Indeed, despite the extension parameter used to transform 2D-features maps into 3D space, they are still 2D-related while the other branch directly deals with 3D-features.

The addition of partial spatial context (sub-volume) seems even more penalizing to the network, suggesting that T_{prior} information is less contributing when used for sub-volume initialization than when directly fed as vector input to the network, as in other experiments.

Adding the SSIM-penalization loss term resulted in a poor improvement of registration quality compared to MSE loss alone. This might be explained by the unstable convergence involved by the SSIM, also reported by [14]. Besides, SSIM penalization is computationally greedy.

Finally, the best tested form of spatial context is the local context input, through the addition of the previous slice S_{t-1}. This local input is better processed using the new proposed 3D CNN, which demonstrates a better capture of the concatenated 2D inputs channel dependencies.

2) Adding temporal information through prior information and probe tracking: Compared to the study without prior navigation information (FVR-Net), we observed a significant improvement of registration quality with the addition of T_{prior}. Indeed, giving directly information about the previous localization in the reference 3D space seems to facilitate learning. It can be compared, more or less, to the idea of an initialization during traditional registration. Moreover, such information directly added in a form similar to the output vector T_t, can be easily used and interpreted by the network.

The scenario with probe tracking information (Im + PT) outperforms the other (only Im), both in terms of accuracy and in terms of reliability, over all experiments. Indeed, this input gives information about the relative motion that has been applied since the previous registration (T_{prior} for S_{t-1}) and can thus be easily combined with the relative in-plane motion between the two successive slices context, in the local experiment. Furthermore, additional tests (not described in this study) suggested that relative probe tracking information is better processed when inputted separately from T_{prior}. Indeed, relative and global temporal information have their own independent benefit and must be kept dissociated.

However, T_{prior} input can be really close to the current T_t to estimate, especially when combined with probe tracking input. To confirm that such strong temporal context contribution neither biases the network capacity nor mislead conclusion about spatial contribution, additional tests were done.

First, we compared each experiment results to a baseline assessment "without network" (using only the geometric transformation, see sec.IV-A) and we observed improved results using the network indicating that other inputs are well processed.

Secondly, we conducted the same comparison between spatial forms, for all experiments but without T_{prior} (not presented in this study). Indeed, as spatial context and previous transform input can share a lot of common and redundant information about global localization within the 3D environment, such experiment evaluated the spatial contribution independently and without any interference. The results of this experiment (not presented in this study) illustrate the same trends as in Table I, which demonstrate that additional complex spatial forms do not seem well adapted for help. However, we obtained lower registration quality, suggesting temporal context seems critical.

In conclusion, it seems that the superiority of the proposed method lies in the simplest possible combination between temporal and spatial context: prior position (vector input), probe

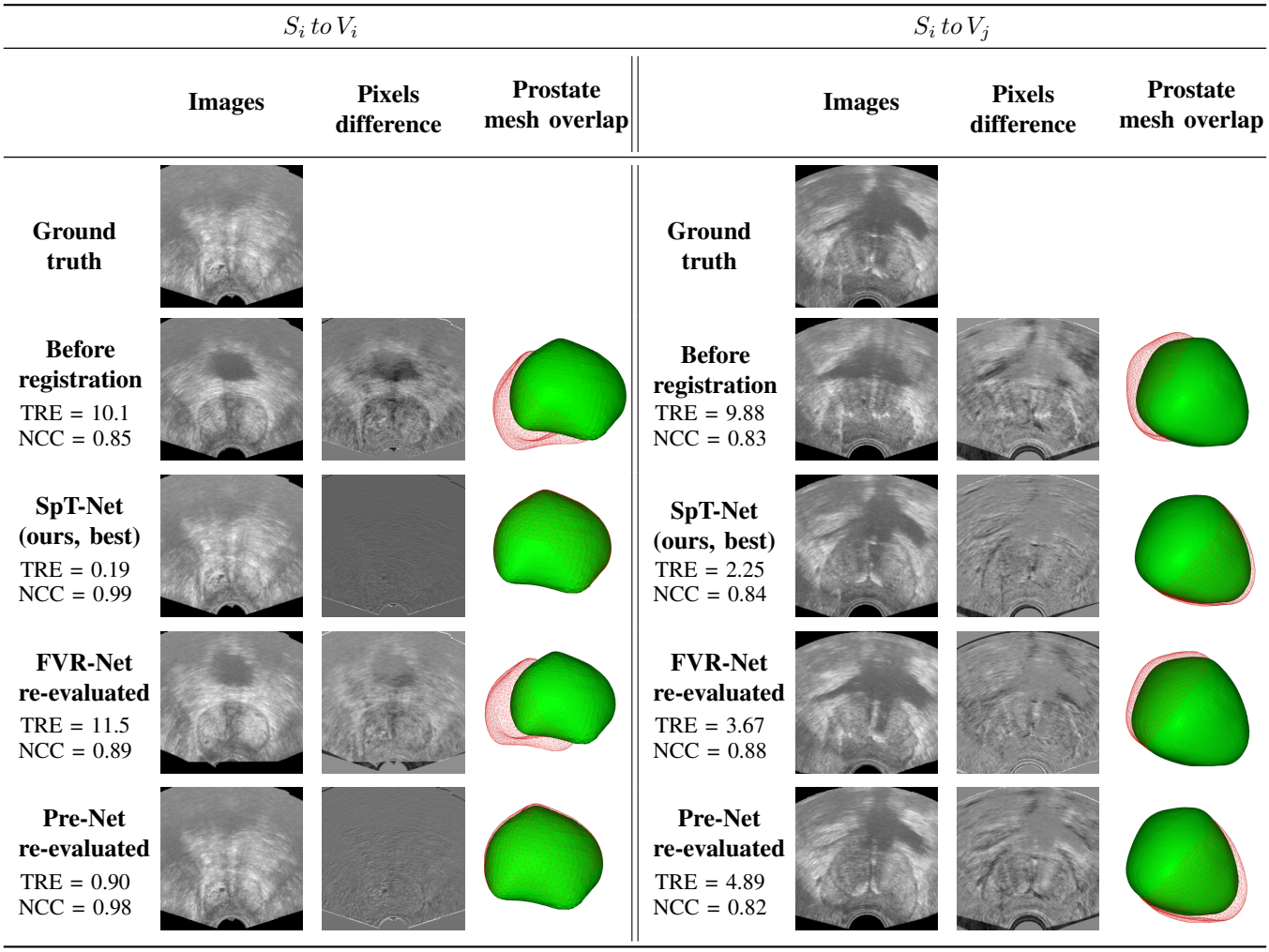

Fig. 5: Illustrative examples of registration quality on US images on S_i to V_i (left) and S_i to V_j (right) datasets. Obtained resampled images are compared to the ground-truth images through pixels difference. Prostate mesh superposition (ground-truth in green, predicted in red dotted-line) is illustrated in the last column as well as the related TRE and NCC measures.

tracking (vector input), and local spatial context (previous image input).

B. Validation contribution

1) Cumulative evaluation on trajectories: We experimented temporal-based predictions that depend on the predictions at previous steps (T_{prior} or initialized sub-volume). As temporal context can be really impacted by drifted results (error accumulation phenomenon), we evaluated its impact, unlike other works in the literature [14], [19]. Indeed, our previous work [19] was limited to the "ideal" case where the previous predictions (T_{prior}) were consistently perfect, by using directly ground-truth registration of the previous step. Similarly, Guo's study [14] oriented their sub-volumes based on a range around the current ground-truth localization. Such method is even more biased as it relies on the current localization and not the previous one (as in [19]).

We demonstrated that drift can have a significant impact on the overall accuracy. Such drift can be corrected when reaching each biopsy location, using 3D/3D registration already integrated in clinical workstations. These registrations require a few seconds but can be performed at each biopsy targetting, as the urologist needs to stop in order to collect the prostate core and change the needle. The access and use of the resetting information seem reasonable and feasible during current clinical practice timeline.

Thus, the proposed cumulative evaluation, based on successive 2D/3D registrations over a trajectory with drift resettings, respects both clinical procedure realism and timeline.

2) Generalization capabilities on two clinical difficulty levels: As currently available clinical data does not provide 2D US image flow and corresponding prostate localization, we generated two datasets with two levels of difficulty. The first dataset ("S_i to V_i") is the most common evaluation process proposed in the registration literature and allows for comparison. But it does not take into account the complexity of real cases. Indeed, the 2D images are extracted from the same volume they are registered to, and some dependencies and correlation between the 2D slice and volume can facilitate or even bias the learning. The developed new data generation process ("S_i to V_j" dataset) allows including additional complexity: input slices are now independent from the reference volume in which we

want to register them, and may present different anatomical deformations and noise patterns. our evaluation using this data generation process is, to the best of our knowledge, more robust than other literature validation.

Our model outperforms by far the rest of the literature methods, on the two datasets evaluated. It demonstrates robust generalization capabilities even on more complex and more realistic tasks.

3) Clinical requirements and application feasibility: Our best results meet clinical requirements of prostate biopsy navigation (see section I-A), with a response time about 70 ms per registration (averaged on all test samples), and with an average error close to (respectively below) the expected 2.5 mm for complex (respectively simple) tasks. This allows envisioning real-time 2D/3D US registration, and thus navigation assistance, from an initial position to a biopsy site.

Our method is easily reproducible thanks to commonly used network layers and simple inputs. However, some inputs can be more specific such as probe tracking and registration resets. While most studies tend to perform robust registration without requiring hardware tracking, we conclude that the addition of an inertial probe sensor can be important. These tracking devices are small and cheap and could easily be integrated for a clinical application. Moreover, our simulated prior resets rely on 3D/3D registration, which are available on several clinical commercial platforms or can be performed with conventional or deep learning methods.

Finally, even if our model demonstrates robust generalization capabilities on more complex tasks, the simulated database still presents some limitations. For now, only simple base-to-apex motions are used to simulate the 2D US flow and pseudo biopsy schemes are simulated over these sweeps. Such sweeps allow avoiding resampling bias in US-cone orientation but does not mimic a complete realistic biopsy procedure. Besides, even if "$S_i\,to\,V_j$" dataset tends to add different prostate volume configurations, real-time clinical deformation of the organ in the complete 2D US flow can be more complex. These limitations must be solved for further development compatible with real cases scenario.

VI. Conclusion

This paper introduces a spatiotemporal registration network (SpT-Net) to localize continuously a 2D US image relatively to a previously available reference US volume, acquired at the beginning of the procedure. Our best model is obtained using: prior navigation trajectory information, based on previous registration results and probe tracking, in addition to local spatial context, through a 3D CNN architecture. The conducted experiments suggest that adding new kinds of spatial context (input volume/sub-volume, ssim-penalization) does not always serve the purpose in the most effective way.

We developed an accurate clinical validation of the method including realistic cumulative evaluation on trajectory and new database generation process with two levels of registration difficulty. Such evaluation is, to the best of our knowledge, more robust than any other validation approach proposed in similar work. We obtained promising results, which respect clinical requirements and application feasibility, and which outperform similar state-of-the-art methods. This makes our approach a promising tool for prostate biopsy navigation assistance and more generally for any US image-guided procedures.

Further improvements will include the generation of new databases from data collected during real clinical biopsy procedures. Finally, as we demonstrated the profit of using temporal context, new questions emerged about using sequence of input images through dynamical structure to retain the global and complete biopsy trajectory.

Acknowledgment

Sincere thanks to clinicians from the Urology department of the Grenoble University Hospital, for their collaboration in data acquisition. This work was partly supported by the French Agence Nationale de la Recherche, "Investissement d'Avenir" program (grants MIAI@Grenoble Alpes under reference ANR-19-P3IA-0003 and CAMI Labex under reference ANR-11-LABX-0004), by Région Rhône-Alpes (project ProNavIA) and by the PSPC project DIANA.

References

[1] D. J. Gillies *et al.*, "Real-time registration of 3d to 2d ultrasound images for image-guided prostate biopsy," *Med. Phys.*, vol. 44, no. 9, pp. 4708–4723, 2017.

[2] V. Karnik *et al.*, "Assessment of image registration accuracy in three-dimensional transrectal ultrasound guided prostate biopsy," *Med. Phys.*, vol. 37, pp. 802–13, feb 2010.

[3] F. Cornud *et al.*, "Trus–mri image registration: A paradigm shift in the diagnosis of significant prostate cancer," *Abdom. Imaging*, vol. 38, no. 6, pp. 1447–1463, dec 2013.

[4] A. M. Brown *et al.*, "Recent advances in image-guided targeted prostate biopsy," *Abdom. Imaging*, vol. 40, no. 6, pp. 1788–1799, aug 2015.

[5] Q. Zeng *et al.*, "Weakly non-rigid mr-trus prostate registration using fully convolutional and recurrent neural networks," in *Med. Imaging 2020: Image Processing*, I. Išgum and B. A. Landman, Eds., vol. 11313. SPIE, 2020, pp. 754–760.

[6] Q. Zeng *et al.*, "Label-driven magnetic resonance imaging (mri)-transrectal ultrasound (trus) registration using weakly supervised learning for mri-guided prostate radiotherapy," *Phys. Med. Ampmathsemicolon Biol.*, vol. 65, no. 13, p. 135002, jun 2020.

[7] Y. Hu *et al.*, "Label-driven weakly-supervised learning for multimodal deformable image registration," *CoRR*, vol. abs/1711.01666, 2017.

[8] H. Guo *et al.*, "Deep adaptive registration of multi-modal prostate images," *Comput. Med. Imaging Graph.*, vol. 84, p. 101769, 2020.

[9] G. Haskins *et al.*, "Learning deep similarity metric for 3d mr–trus image registration," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 3, pp. 417–425, oct 2018.

[10] M. Baumann *et al.*, "Prostate biopsy tracking with deformation estimation," *Med. Image Anal.*, vol. 16, no. 3, pp. 562–576, apr 2012.

[11] S. Xu *et al.*, "Real-time mri-trus fusion for guidance of targeted prostate biopsies," *Comput. Aided Surg.*, vol. 13, no. 5, pp. 255–264, 2008.

[12] A. Bhardwaj *et al.*, "Rigid and deformable corrections in real-time using deep learning for prostate fusion biopsy," in *Med. Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*, B. Fei and C. A. Linte, Eds., vol. 11315. SPIE, 2020, pp. 486–499.

[13] S. Zhang *et al.*, "2d ultrasound and 3d mr image registration of the prostate for brachytherapy surgical navigation," *Medicine (Baltimore)*, vol. 94, no. 40, 2015.

[14] H. Guo *et al.*, "End-to-end ultrasound frame to volume registration," in *Med. Image Comput. Comput. Assist. Interv. – MICCAI 2021*. Cham: Springer International Publishing, 2021, pp. 56–65.

[15] S.-Y. Selmi *et al.*, "Hybrid 2d-3d ultrasound registration for navigated prostate biopsy." *Int. J Comput. Assist. Radiol. Surg.*, vol. 13, no. 7, pp. 987–995, jul 2018.

[16] C. Beitone *et al.*, "Towards real-time free-hand biopsy navigation." *Med. Phys.*, vol. 48, no. 7, pp. 3904–3915, jul 2021.

[17] D. J. Gillies *et al.*, "Ring navigation: An ultrasound-guided technique using real-time motion compensation for prostate biopsies," in *Med. Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, B. Fei and R. J. W. III, Eds., vol. 10576. SPIE, 2018, pp. 404–409.

[18] T. De Silva *et al.*, "2d-3d rigid registration to compensate for prostate motion during 3d trus-guided biopsy." *Med. Phys.*, vol. 40, no. 2, p. 022904, feb 2013.

[19] T. Dupuy *et al.*, "2d/3d deep registration for real-time prostate biopsy navigation," in *SPIE Med. Imaging 2021*, ser. Med. Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling, vol. 11598. International Society for Optics and Photonics., feb 2021, p. 115981P.

[20] B. Hou *et al.*, "Predicting slice-to-volume transformation in presence of arbitrary subject motion," in *Med. Image Comput. Comput. Assist. Interv. – MICCAI 2017*. Cham: Springer International Publishing, 2017, pp. 296–304.

[21] T.-F. Yu *et al.*, "Slice localization for three-dimensional breast ultrasound volume using deep learning," *DEStech Trans. Eng. Technol. Res.*, 2019.

[22] S. S. Mohseni Salehi *et al.*, "Real-time deep pose estimation with geodesic loss for image-to-template rigid registration," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 470–481, 2019.

[23] B. Almogadwy *et al.*, "A deep learning approach for slice to volume biomedical image integration," in *Proc. of the 2019 11th Int. Conf. Bioinf. Biomed. Technol.*, ser. ICBBT'19. Stockholm, Sweden: Association for Computing Machinery, 2019, pp. 62–68.

[24] Y. Li *et al.*, "Standard plane detection in 3d fetal ultrasound using an iterative transformation network," in *Med. Image Comput. Comput. Assist. Interv. – MICCAI 2018*. Springer International Publishing, 2018, pp. 392–400.

[25] Z. Chen *et al.*, "Real-time and multimodal brain slice-to-volume registration using cnn," *Expert Syst. Appl.*, vol. 133, pp. 86–96, 2019.

[26] W. Wei *et al.*, "Towards fully automatic 2d us to 3d ct/mr registration: A novel segmentation-based strategy," in *2020 IEEE 17th Int. S. Biomed. Imaging (ISBI)*, 2020, pp. 433–437.

[27] R. Prevost *et al.*, "3d freehand ultrasound without external tracking using deep learning," *Med. Image Anal.*, vol. 48, pp. 187–202, 2018.

[28] K. Miura *et al.*, "Localizing 2d ultrasound probe from ultrasound image sequences using deep learning for volume reconstruction," in *Med. Ultrasound, Preterm, Perinatal Paediatric Image Anal.* Springer International Publishing, 2020, pp. 97–105.

[29] M. Luo *et al.*, "Self context and shape prior for sensorless freehand 3d ultrasound reconstruction," in *Med. Image Comput. Comput. Assist. Interv. – MICCAI 2021*. Cham: Springer International Publishing, 2021, pp. 201–210.

[30] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. of the 28th Int. Conf. Neural Information Processing Systems - Volume 2*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, p. 2017–2025.

[31] H. Guo *et al.*, "Sensorless freehand 3d ultrasound reconstruction via deep contextual learning," in *Med. Image Comput. Comput. Assist. Interv. – MICCAI 2020*. Cham: Springer International Publishing, 2020, pp. 463–472.

[32] A. Singh *et al.*, "Deep predictive motion tracking in magnetic resonance imaging: Application to fetal imaging," *IEEE Trans. Med. Imaging*, vol. 39, no. 11, pp. 3523–3534, 2020.

[33] P.-H. Yeung *et al.*, "Learning to map 2d ultrasound images into 3d space with minimal human annotation," *Med. Image Anal.*, vol. 70, p. 101998, 2021.

[34] M. Kok *et al.*, "Using inertial sensors for position and orientation estimation," *Found. Trends® Signal Process.*, vol. 11, no. 1-2, pp. 1–153, 2017.

[35] J. B. West *et al.*, "Fiducial point placement and the accuracy of point-based, rigid body registration," *Neurosurg.*, vol. 48, no. 4, pp. 810–817, apr 2001.

[36] H. R. Boveiri *et al.*, "Medical image registration using deep neural networks: a comprehensive review," *Comput. Elec. Eng.*, vol. 87, p. 106767, 2020.

[37] D. Chicco *et al.*, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, p. e623, jul 2021.

[38] G. Fiard *et al.*, "Simulation-based training for prostate biopsies: Towards the validation of the biopsym simulator." *Minim Invasive Ther Allied Technol*, vol. 29, no. 6, pp. 359–365, dec 2020.