Medical Image Analysis (2019)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Siam-U-Net: encoder-decoder siamese network for knee cartilage tracking in ultrasound images

Matteo Dunnhofer[a,*], Maria Antico[b,c], Fumio Sasazawa[b,c,d], Yu Takeda[e], Saskia Camps[f,g], Niki Martinel[a], Christian Micheloni[a], Gustavo Carneiro[h], Davide Fontanarosa[c,i]

[a]*Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy*
[b]*School of Chemistry, Physics and Mechanical Engineering, Queensland University of Technology, Brisbane, Queensland, Australia*
[c]*Institute of Health Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia*
[d]*Department of Orthopaedic Surgery, Faculty of Medicine and Graduate School of Medicine, Hokkaido University, Sapporo, Japan*
[e]*Department of Orthopaedic Surgery, Hyogo College of Medicine, Nishinomiya, Hyogo, Japan*
[f]*Faculty of Electrical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands*
[g]*Oncology Solutions Department, Philips Research, Eindhoven, the Netherlands*
[h]*Australian Institute for Machine Learning, School of Computer Science, the University of Adelaide, Adelaide, Australia*
[i]*School of Clinical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia*

ARTICLE INFO

Article history:

Knee Arthroscopy, Knee cartilage, Ultrasound, Ultrasound Guidance, Visual Tracking, Fully Convolutional Siamese Networks, Deep Learning

ABSTRACT

The tracking of the knee femoral condyle cartilage during ultrasound-guided minimally invasive procedures is important to avoid damaging this structure during such interventions. In this study, we propose a new deep learning method to track, accurately and efficiently, the femoral condyle cartilage in ultrasound sequences, which were acquired under several clinical conditions, mimicking realistic surgical setups. Our solution, that we name Siam-U-Net, requires minimal user initialization and combines a deep learning segmentation method with a siamese framework for tracking the cartilage in temporal and spatio-temporal sequences of 2D ultrasound images. Through extensive performance validation given by the Dice Similarity Coefficient, we demonstrate that our algorithm is able to track the femoral condyle cartilage with an accuracy which is comparable to experienced surgeons. It is additionally shown that the proposed method outperforms state-of-the-art segmentation models and trackers in the localization of the cartilage. We claim that the proposed solution has the potential for ultrasound guidance in minimally invasive knee procedures.

1. Introduction

Ultrasound (US) imaging offers accurate and precise anatomical analysis, superior resolution and relative cost-effectiveness. Currently, it is the only real-time volumetric imaging modality

*Corresponding author:
Matteo Dunnhofer
Machine Learning and Perception Lab
Department of Mathematics, Computer Science and Physics (DMIF)
University of Udine
via delle Scienze 206
33100 Udine, Italy
e-mail: dunnhofer.matteo@spes.uniud.it (Matteo Dunnhofer), maria.antico@hdr.qut.edu.au (Maria Antico), sasazawa230@gmail.com (Fumio Sasazawa), yu.takeda@qut.edu.au (Yu Takeda), saskiacamps@gmail.com (Saskia Camps), niki.martinel@uniud.it (Niki Martinel), christian.micheloni@uniud.it (Christian Micheloni), gustavo.carneiro@adelaide.edu.au (Gustavo Carneiro), d3.fontanarosa@qut.edu.au (Davide Fontanarosa)

that is clinically available and compatible with surgical conditions. The knee is a particularly interesting region amenable to the use of US scanning in surgery-guided applications (Lueders et al., 2016), where most hard and soft tissue structures can be properly identified, segmented and tracked. Several publications have shown that tendons (Wong-On et al., 2015), ligaments (Oshima et al., 2016), menisci (Faisal et al., 2015), nerves (Faisal et al., 2015; Giraldo et al., 2015) and cartilages (Faisal et al., 2018b,a) can be clearly visualized using US imaging. Medical tools like arthroscopes (Tyryshkin et al., 2007) can also be visualized and tracked. US guided minimally invasive procedures (MIPs) that have been performed on the knee include needle guidance for injections (Morvan et al., 2012; Köroğlu et al., 2012; Hackel et al., 2016), tendon fenestration (Kanaan et al., 2013) and ligament reconstructions (Hirahara and Andersen, 2016).

Knee arthroscopy is a well-established MIP for diagnosis and treatment of disorders in knee joints. Its execution requires an initial small incision of the skin and soft tissues of the patient, and the successive insertion of the arthroscope, a flexible scope carrying a small camera, inside the joint. Through a video monitor, 2D images acquired by the camera are displayed to the surgeon, who is able to visualize the anatomical structures of the knee and to guide surgical instruments. Despite being a common procedure nowadays, this kind of intervention demands a great physical and mental effort from surgeons, with the consequent increased chance of damaging the knee structures (Jaiprakash et al., 2017). To overcome these problems, US guided knee arthroscopy is currently being studied (Wu et al., 2018). Automatic interpretation of 2D+time/3D+time US images of the knee could be a valuable tool able to offer accurate localization and visualization of the knee structures, ultimately reducing surgeon's operating stress. Furthermore, clinicians indicate that knee arthroscopy will be among the first types of MIPs that, in the near future, will be fully automated by robotic surgery (Wu et al., 2018). In these scenarios, the automatic interpretation of US images is required (Antico et al., 2019). A tracking tool can exploit the visual and temporal information acquired during the intervention, to interpret the variations in position and shape of the knee structures. Such a system would require a minimal user initialization, e.g. a contour or a segmentation and, in comparison with the surgeon, could produce a more accurate and repeatable localization.

Among the structures that are at risk during knee arthroscopy, cartilages are particularly vulnerable (Jaiprakash et al., 2017). Therefore they were chosen as the first target of the proof-of-concept work introduced in this paper. In US images, cartilages are typically clearly visible, but it is not straightforward to track them under surgical conditions, where their position, shape and appearance change due to the physics of the US beam, US probe shifts or knee joint flexion to different angles. In Figure 1, US images with the cartilages highlighted are shown.

In the past, several methodologies have been proposed to track anatomical structures in US images, such as tongue (Akgul et al., 1999; Roussos et al., 2009), heart's left ventricle (Carneiro and Nascimento, 2013; Huang et al., 2014), vessels (Guerrero et al., 2007) and liver landmarks (De Luca et al., 2015; Gomariz et al., 2019). These methodologies included, for example, active contour models and their variations (Akgul et al., 1999; Roussos et al., 2009), statistical approaches like Kalman filters (Guerrero et al., 2007), sparse representation and dictionary learning (Huang et al., 2014). One of the biggest limitations of the aforementioned methodologies is that these methods are model-centred and make many assumptions about the problem that may not be realistic. In addition, they also require the development of typically sub-optimal hand-designed representations. To address those issues, deep learning (DL) (Lecun et al., 2015) solutions have been introduced to the field of anatomical structure tracking. DL is a method that automatically learns optimal data representations. For example, Carneiro and Nascimento (2013) combined deep belief networks with a probabilistic non-Gaussian model to track the motion of the left ventricle. Nouri and Rothberg (2015) proposed convolutional neural networks (CNNs) with a learned distance metric, while Gomariz et al. (2019) developed a deep siamese neural network (SNN).

The latter solution is based on recently proposed SNNs for visual tracking (Held et al., 2016; Bertinetto et al., 2016b; Tao et al., 2016; Guo et al., 2017; Valmadre et al., 2017; Wang et al., 2017; Li et al., 2018b,a; Wang et al., 2018). The idea behind these methodologies is to treat the tracking as a similarity problem. Despite the outstanding results achieved on benchmark datasets of natural images, SNN-based visual trackers fail to be applied directly to medical domains due to their high architectural complexity and the unsuitable target object's state representation as bounding boxes. Here we try to reduce this gap by presenting a methodology that combines deep neural networks (DNNs) for segmentation of medical data and the recent SNN-based framework for visual tracking.

Overall, in this paper we propose a DL methodology applied to US images to track the femoral condyle cartilage under several clinical conditions during MIP. In particular, our contribution is threefold:

1. The first real-time tracking algorithm for US images of the femoral condyle cartilage;
2. A novel combination of disparate DL architectures, named Siam-U-Net, which merges U-Net (Ronneberger et al., 2015) and the siamese framework (Bertinetto et al., 2016b,a);
3. The first use, in the context of visual tracking, of an end-to-end learning strategy that leverages a training loss generally used for segmentation tasks.

To train and evaluate our model, multiple US scans were taken from knees of six volunteers. Volumetric US images were acquired during leg flexion to mimic possible positions of the leg during the intervention, and while the US probe shifted on the surface of the knee. From the US images obtained, given an initial cartilage segmentation, the structure was tracked either in the consecutive US frames, referred as to temporal tracking or both within neighbouring US slices of the same volume and consecutive frames, defined as to spatio-temporal tracking. We show that using segmentation architectures inside the siamese tracking framework is an effective way to localize the femoral cartilage in 2D US sequences with a minimal user intervention.

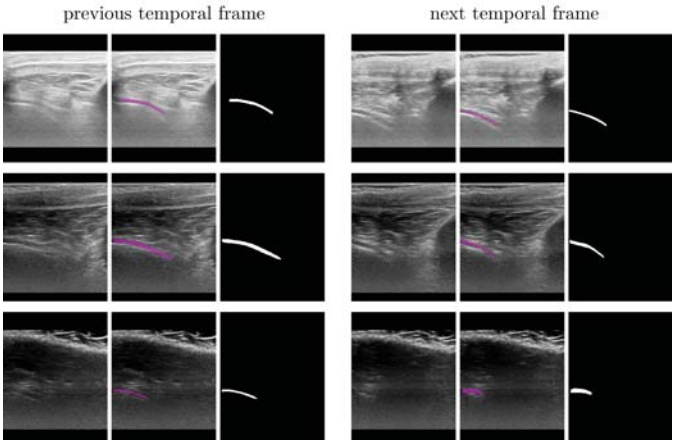

Fig. 1. Visual examples of US images of the knee, with the highlight of the femoral condyle cartilage. Each of three-image blocks shows a 2D US image, the same US image with the cartilage's ground-truth segmentation (in pink) drawn by a surgeon, and the corresponding binary cartilage mask, respectively. Each row of images shows the transformation of the cartilage from a previous temporal frame of a US sequence to the successive temporal frame. First two rows depict examples of translation of the US probe. The third row presents an example of transformation while the knee is flexing.

Despite the fact that we propose a 2D+time approach, our solution is fully volumetric, in the sense that it is capable of tracking, both temporally and spatially, the condyle cartilage in any section of 3D+time US sequences.

The proposed solution exhibits a segmentation accuracy, in terms of Dice Similarity Coefficient (DSC) (Dice, 1945; Sørensen, 1948), that is comparable to the one produced by two expert operators and that is higher than the segmentation models proposed by Ronneberger et al. (2015) and by Léger et al. (2018). Our solution also offers better performance than the state-of-the-art trackers OSVOS (Caelles et al., 2017) and RGMP (Oh et al., 2018) which were developed for video object segmentation.

2. Related Work

Our solution can be placed at the intersection of three research areas: visual tracking, US tracking and medical image segmentation. In this section, we review the most relevant works to our methodology.

2.1. Visual Tracking

In its simplest form, the visual tracking problem consists of the consistent recognition of a target in consecutive video

frames. The most used target representation is a bounding box that encloses the object of interest. If a more precise localization is needed, a segmentation that identifies the object pixel-by-pixel should be used. In the computer vision literature, the first approach is known as visual object tracking (VOT), while the second is referred as to video object segmentation (VOS).

2.1.1. Visual Object Tracking

In VOT problems, a moving target object must be identified within a searching area (usually bigger than the target) in each video frame. The target is localized in the searching area's sub-region that has the highest visual similarity with the target in the previous frames. In the past years, SNNs have been used for VOT mostly because of their computational efficiency and good accuracy on existing benchmark datasets (Held et al., 2016; Bertinetto et al., 2016b; Tao et al., 2016; Guo et al., 2017; Valmadre et al., 2017; Wang et al., 2017; Li et al., 2018b,a; Wang et al., 2018). An SNN (Bromley et al., 1993) is a particular neural network architecture commonly used to learn representations of two input objects by optimizing a training loss that compares their similarity in higher-level feature spaces. In VOT, this idea is exploited to define a similarity map obtained by comparing the target representation and every sub-matrix of the searching area representation, using as comparison metric the cross-correlation. This solution, known in literature as SiamFC, was firstly proposed by Bertinetto et al. (2016b). Subsequently, SiameseRPN (Li et al., 2018b) increased the detection accuracy by fusing a Region Proposal Network (Ren et al., 2015) and the cross-correlation operation. Li et al. (2018a) proposed to aggregate the CNN features through layer-wise and depth-wise convolutions to enhance the cross-correlation. Wang et al. (2018) suggested a siamese architecture to unify the VOT and VOS tasks. Their proposed network is initialized with a ground-truth bounding box and is able to propagate both the box and the segmentation mask that identify and localize the target object through the video.

All these methods have high performance in terms of speed, as they are able to produce the target representations in real-time, i.e. they are able to process more than 30 images per second. This is clearly an advantage which we want to include in our solution. However, these methods are not directly suited for our problem, because a bounding box representation of the target is not sufficient to produce precise information about the location and shape of the cartilage. Additionally, the CNN employed by Wang et al. (2018) has many learnable parameters that are not needed for the problem of tracking a single object like the cartilage and that would lead to overfitting, given the limited number of training examples available for our task. Very deep neural networks can achieve outstanding results, but the main drawback is the necessity of large sets of information rich data. Compared to natural images (on which the presented methods perform well), US images are less informative and thus, networks with less parameters can be used. Lowering the number of parameters reduces the chances of overfitting and increases the processing speed of the network.

2.1.2. Video Object Segmentation

To tackle the VOS problem, different methodologies have been proposed. MaskTrack (Perazzi et al., 2017) introduced a pixel-labeling CNN that frame-by-frame refines, through a combination of offline and online learning strategies, the previously detected segmentations. Several other papers (Grundmann et al., 2010; Tsai et al., 2012; Marki et al., 2016) used spatio-temporal graph representations to distribute the labels estimates to the pixels of consecutive frames. Alternative approaches independently segmented every single frame (Caelles et al., 2017; Voigtlaender and Leibe, 2017; Maninis et al., 2018) using an online training scheme. One of the most relevant works in this direction (Caelles et al., 2017) proposed to use one-shot learning to fine-tune online a Fully Convolutional Network (FCN) (Long et al., 2014) which was pre-trained to distinguish target object pixels from the ones of the searching area. This solution allowed to reach superior results, but with the drawback of an online pre-processing time of up to 10 minutes. The employment of SNNs in VOS was firstly introduced by Oh et al. (2018), who proposed an encoder-decoder fully convolutional siamese architecture with a global convolution operator that was trained to produce a segmentation mask for every

frame, given as input: the current frame, the mask produced at the previous time step and the initial ground-truth mask. Our proposed Siam-U-Net follows a similar approach, but it substitutes the global convolution operation with the depth-wise cross-correlation. This allows to produce a high level activation map, which is then refined by the decoder into a fine-grained segmentation.

Despite the promising segmentation accuracies achieved by the methods described above, their high complexity will not allow the production of segmentations in a very short time. In fact, these solutions can process from less than an image to a maximum of 10 images per second. Thus, they are not suited for real-time applications like our problem of interest. Moreover, no methodology took advantage of the DSC as a training loss, which was shown to lead to better segmenting performance (Milletari et al., 2016).

2.2. *Tracking in US Images*

Visual tracking in US images has received increased interest in the past. Akgul et al. (1999) and Roussos et al. (2009) used variations of active contours to track the motion of the tongue. These methods rely on image gradient and energy-based functions to draw a contour around the edges of the target object. Even though it is a common technique in computer vision, this kind of methods suffer from initialization robustness, which can lead to drifting over time. Guerrero et al. (2007) proposed a real-time algorithm for vessel segmentation and tracking. Their solution used an elliptical model to segment vessels and Kalman filters to track their shape through temporal sequences. A main drawback of this solution is the assumption that anatomical structures can always be represented through elliptical models, thus reducing the generalization capabilities to structures with other shapes. Huang et al. (2014) presented a method that employs multiscale sparse representation and dictionary learning to track the endocardial and epicardial contours of the left ventricle. Despite achieving great results, the biggest limitation of dictionary learning is the assumption that samples can be represented by a linear combination of dictionary items. In contrast, our methodology uses convolutional neural networks (CNNs) to build powerful image representations through non linear operations.

Overall, the biggest limitations of the methods above are that they are model-centred or use linear data-driven methodologies. Furthermore, they make assumptions about the problem that may not hold in practice and they sometimes require the development of sub-optimal hand-designed representations.

More recently, DL based methodologies have been applied to US data. Carneiro and Nascimento (2013) fused deep belief networks and multiple dynamic models by means of a probabilistic non-Gaussian state-space distribution to track the left ventricle. Despite the good results, this method is difficult to be extended to other medical context since the transition model involved takes into account information that is too specific for the cardiac cycle (e.g., it only considers the two cardiac phases of the cycle: diastole and systole). Additionally, the observation model is based on shallow artificial neural networks. In contrast, we employ a CNN based architecture which is proven to work better for spatial data, such as images (Lecun et al., 1998; Krizhevsky et al., 2012). Nouri and Rothberg (2015) proposed a CNN to track liver landmarks in 2D+time US sequences. Their proposed model was trained by optimizing a distance metric between two US image patches. At test time, different image patches were sampled in the current frame around the previous known target location, and the coordinates of the patch with the predicted lower metric value were chosen as new position for the target. We propose a method with a single forward pass, different from the candidate generation procedure proposed by the authors that can harm the processing speed of the tracker, since many image comparisons are to be executed. Gomariz et al. (2019) tackled the liver landmark tracking problem with a SNN and a location prior. This was the first attempt to apply SNNs to US images, but its tracking capabilities are limited to the prediction of the position of the target object, which is represented by the coordinates of a single point. This is not sufficient for our problem of interest that requires precise localization and shape definition of a structure that is characterized by a highly variable appearance.

In general, despite the good reported results, all the approaches mentioned above are not directly applicable to our task because they propose ad-hoc implementations that are optimized for their problem of interest, thus reducing their capability of generalization to other use cases.

2.3. Medical Image Segmentation

FCNs for semantic segmentation were firstly introduced by Long et al. (2014). Their idea was to exploit the knowledge of a CNN pre-trained for natural image classification to perform image segmentation. To this end, the authors added an expanding block to the pre-trained CNN. The block was used to generate the output segmentation by enlarging the CNN intermediate features through convolutional and up-sampling layers. The weights of the newly added module were then learned by means of a supervised segmentation task. This solution showed very good results with respect to previous methodologies (Ce Liu et al., 2011; Farabet et al., 2013; Tighe and Lazebnik, 2013; Pinheiro and Collobert, 2014). However, the required classification pre-training on the ImageNet dataset (Deng et al., 2009) is (still today) very computationally expensive and only suited for natural image processing applications. To overcome these problems, Ronneberger et al. (2015) proposed a novel fully convolutional architecture, named U-Net, that could be trained end-to-end and with few training samples. The structure of the U-Net extended the one from FCN by Long et al. (2014) and it was the combination of a contracting part (the encoder), composed of convolutional and max-pooling layers, and an expanding part (the decoder), consisting of the aggregation of the encoder intermediate features, up-sampling and convolutional layers. Thanks to its outstanding results in many clinical domains (Milletari et al., 2016; Ben-Cohen et al., 2016; Oktay et al., 2016; Çiçek et al., 2016; Yu et al., 2017), today this methodology is considered the standard architecture for medical image segmentation. Despite this, U-Net has not been effectively adapted to include temporal data. Therefore, U-Net was chosen to form just the base CNN architecture of the cartilage tracker proposed in this paper. Léger et al. (2018) tried to include previously computed segmentation masks into U-Net's architecture as an additional input channel. The idea was to use prior information for aiding the task of 3D segmentation by means of a 2D model. Experimental validation showed the proposed model to be stronger than U-Net in segmenting 3D CT scans of the bladder. In principle, the presented methodology could be applied to track anatomical structures in temporal sequences of 2D images. However, tracking requires fast elaboration times and processing searching areas as large as the image size is usually very time consuming. Moreover, the target object has usual motion patterns that can be exploited to reduce computational time and effort in its search. The solution proposed by Léger et al. (2018) does not take into account these considerations.

3. Materials and Problem Formulation

For this study, a dataset of 3D+time images was built by mimicking possible MIP scenarios. In this section we describe how the US data was acquired, labeled and organized. We also give a precise formulation of the problem of tracking the femoral condyle cartilage.

3.1. US Data Acquisition and Labels Generation

To build the US dataset, knees of six healthy volunteers (male and female) have been scanned at the Queensland University of Technology using a Philips EPIQ7 US workstation with a VL13-5 mechanically swept probe (Philips Healthcare, Eindhoven, Netherlands). The ethics approval for data acquisition was granted by Queensland University of Technology Ethics Committee (No. 1700001110). All the volunteers signed an informed consent before the data collection.

The US probe was positioned anteriorly to the knee, and the scans were performed through the volunteer's patellar tendons as shown in Figure 2. The rationale for this choice was to allow enough space for the insertion and manipulation of the surgical instruments through the medial and lateral parapatellar portals (the soft spots at both sides of the patella), as in realistic intra-operative knee arthroscopy scenarios. The US probe was hand-held by an experienced orthopedic surgeon. The US scans were performed with the knees fully submerged in water to minimize

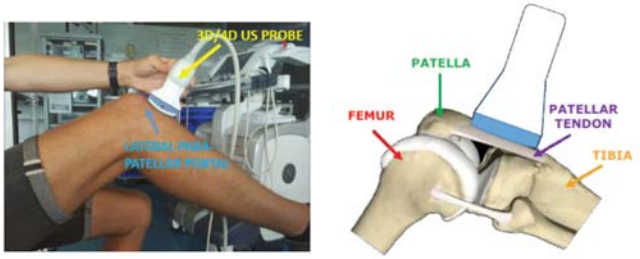

Fig. 2. US probe positioning. On the left: lateral view of the knee joint with the probe placed on the patellar tendon. On the right: schematic US probe positioning representation, showing the positions of reference structures relative to the probe.

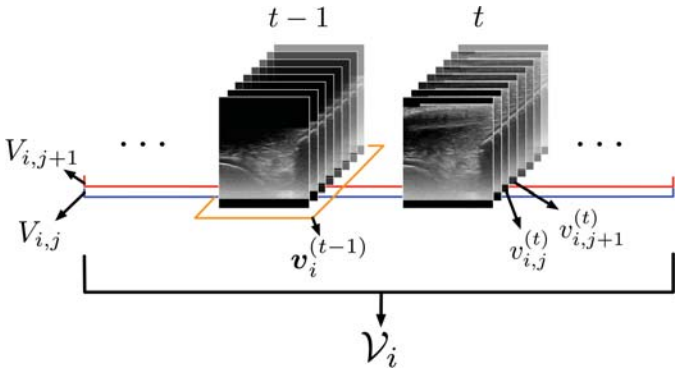

Fig. 3. Visual representation of the notation used throughout the paper. Each 3D+time US sequence is denoted as $\mathcal{V}_i$. The volumes belonging to $\mathcal{V}_i$ are referred as $\boldsymbol{v}_i^{(t)}$ (highlighted by the orange line) for the temporal step t. Each 2D+time sequence $V_{i,j}$ (highlighted by the red and blue lines) comprises the slices $v_{i,j}^{(t)}$ which in turn belong to the volumes $\boldsymbol{v}_i^{(t)}$ respectively.

possible acoustic coupling issues. To mimic normal conditions during surgical procedures, we acquired 35 3D+time sequences (3D volumes in time), for a total of 151 full 3D volumes, flexing the knee from 0 to 30 degrees (F30), and translating the probe along the patellar tendon with the knee flexed at 0 degrees (T0) or at 30 degrees (T30). Table 1 reports a summary of the dataset collected. MRI scans of the knees of the same volunteers have also been acquired in identical geometric conditions and manually fused with the US volumes by an experienced surgeon to accurately identify all the anatomic structures. During knee flexion, the expert operator always tried to capture the US volume from the lower end of the patella to the upper end of the tibia longitudinally, and containing the articular cartilage on both sides of femoral condyles transversely. The US volumes collected had a size of approximately $(4 \times 4 \times 3)$ cm^3 and were acquired at 1 Hz refresh rate.

In the images, typically the femoral cartilages appearance is an hypoechoic band on top of a clear hyperechoic line outlining the bone contour of the femoral condyles. The border between the cartilage layer and Hoffa's fat pad is also typically clearly visible as a thin hyperechoic line parallel to the bone contour. The pixel dimensions are ~0.19mm. The reference segmentations of the femoral cartilages have been manually created by an expert orthopaedic surgeon (Operator 1), along the sagittal slices within the US volumes acquired using MeVisLab (MeVis Medical Solutions AG, Germany). The total number of annotated slice was 18278.

3.2. Problem Formulation

The resulting dataset used for this work is composed of a set of temporal sequences of 3D+time US images and respective labels. We denote it as $\mathcal{D}_{3D+time} = \{(\mathcal{V}_i, \mathcal{G}_i)\}_{i=1}^{35}$, where each pair $(\mathcal{V}_i, \mathcal{G}_i)$ is obtained from ordered sequences of volumes $\mathcal{V}_i = \{\boldsymbol{v}_i^{(t)}\}$ and $\mathcal{G}_i = \{\boldsymbol{g}_i^{(t)}\}$, $t \in \{0, \dots, T-1\}$, $T \in \mathbb{N}$. Each $\boldsymbol{v}_i^{(t)} \in \{0, \dots, 255\}^{r \times c \times d}$ is a US volume of $r \times c \times d$ voxels (in our case $r = 313, c = 255, d = 256$) and $\boldsymbol{g}_i^{(t)} \in \{0, 1\}^{r \times c \times d}$ is the respective reference segmentation volume. Each 2D+time sequence $V_{i,j}$ is composed by considering each $v_{i,j}^{(t)} \in \{0, \dots, 255\}^{r \times c \times 1} \subset \boldsymbol{v}_i^{(t)}$, $j \in \{0, \dots, d-1\}$, i.e. the 2D matrix component (belonging to the volume $\boldsymbol{v}_i^{(t)}$) which we refer as slice, for which the 2D mask $g_{i,j}^{(t)} \in \{0, 1\}^{r \times c \times 1} \subset \boldsymbol{g}_i^{(t)}$ presents a localization of the cartilage. In formal terms $V_{i,j} = \{v_{i,j}^{(t)} \,|\, \forall t \; \exists g_{i,j}^{(t)} \neq 0^{r \times c \times 1}\}$. In Figure 3, we show a visual representation of the notation employed in this paper.

The entire dataset is divided into training and testing sets subject-wise, i.e. with no overlap in terms of volunteers in the training and testing sets. In Table 1, details about the acquired data are reported, while in Figure 4, the distribution of the contoured slices is shown for each subject.

The use of sequences of 2D data, and so following a 2D+time tracking approach (instead of a 3D+time approach), was motivated by the fact that this setting allowed significantly less computational effort for data processing. In fact, dealing with sequences of 3D volumes would have required the reduction of

Table 1. Summary of the dataset collected for the study. For each volunteer, we report the scanned legs (L: left, or R: right), the scan type (probe translation with the knee at 0 (T0) or 30 degrees (T30) flexion, or knee flexion from 0 to 30 degrees (F30), the number of volumes acquired and the number of 2D US slices contoured by the expert Operator 1.

Subject id	Leg scanned	Scan modalities	# volumes	# annotated slices
1	L, R	T0, T30, F30	28	3402
2	L, R	T0, T30, F30	24	3245
3	L, R	T0, T30, F30	29	2657
4	L, R	T0, T30, F30	28	3119
5	L, R	T0, T30, F30	23	3872
6	L, R	T0, T30, F30	19	1983

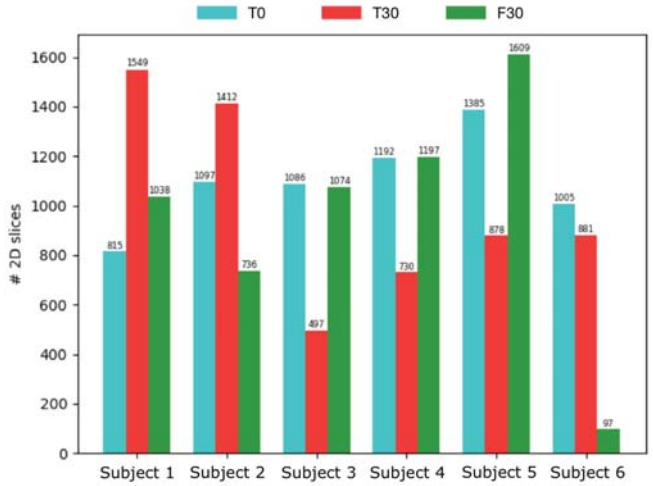

Fig. 4. The distribution of the 2D contoured slices shown for each subject included in the dataset.

the volumetric dimensions of the data to fit in the memory of currently available machines, with the consequent loss of valuable information. A 3D+time approach would also need a much larger amount of labeling effort to produce sufficient samples for making DL methods work well, given that each volume can have up to 203 2D annotated slices. Moreover, some of 3D volumes acquired in this work were just partially contoured (i.e. not all the 2D slices composing the volumes and effectively containing a cartilage were segmented by the expert) making them unusable for 3D+time processing.

Our problem of interest is the precise localization of the femoral condyle cartilage in each of the 2D slices that compose a 2D+time US sequence, given an initial 2D reference segmentation for the first slice of the sequence. In formal terms, given a temporal sequence $V_{i,j}$, containing T slices and an initial reference segmentation of the cartilage $g^{(0)}_{i,j}$, drawn by an expert, our method will produce the masks $s^{(t)}_{i,j} \in \{0,1\}^{r \times c \times 1}, t \in \{1,\dots,T-1\}$ that successfully locate the femoral cartilage. With this setting, the cartilage location and shape representations, $s^{(t)}_{i,j}$, are expressed as binary segmentations.

4. Method

The key idea of this paper is to combine an encoder-decoder neural network architecture such as U-Net (Ronneberger et al., 2015) with the siamese tracking framework (Bertinetto et al., 2016b,a). We begin this section by describing the novel DL architecture, Siam-U-Net, that is used to produce a cartilage segmentation within a 2D US image, given the information about the structure's visual appearance in the previous time frame and the searching area where the cartilage is supposed to be present. After discussing training procedure of the network, we introduce how the architecture is used to effectively track the cartilage in a 3D sequence.

4.1. Siam-U-Net Architecture

The neural network architecture we propose takes inspiration from the encoder-decoder architecture of U-Net (Ronneberger et al., 2015), and the cross-correlation operation used in the traditional siamese framework for visual tracking. A graphical representation of the proposed network is depicted in Figure 5.

The network receives as input two cropped images, a smaller one for the target cartilage and a bigger one for the searching area. These image crops are passed through the encoder branch denoted as $E_{\theta_E}(\cdot)$, whose weights θ_E remain the same for the two inputs. The encoder is composed of a sequence of five computational blocks each including a set of 3×3 convolutional layers and 2×2 max pooling operators applied with a stride of 2 to reduce the size of the feature maps. Each convolutional layer is followed by batch normalization (Ioffe and Szegedy, 2015), ReLU activation and a dropout (Srivastava et al., 2014) layer.

After the target and searching area are processed by the encoder, the cross-correlation operation is performed. The target representation is depth-wise, i.e. feature map by feature map, cross correlated to the searching area representation, as proposed by Bertinetto et al. (2016a); Li et al. (2018a). This

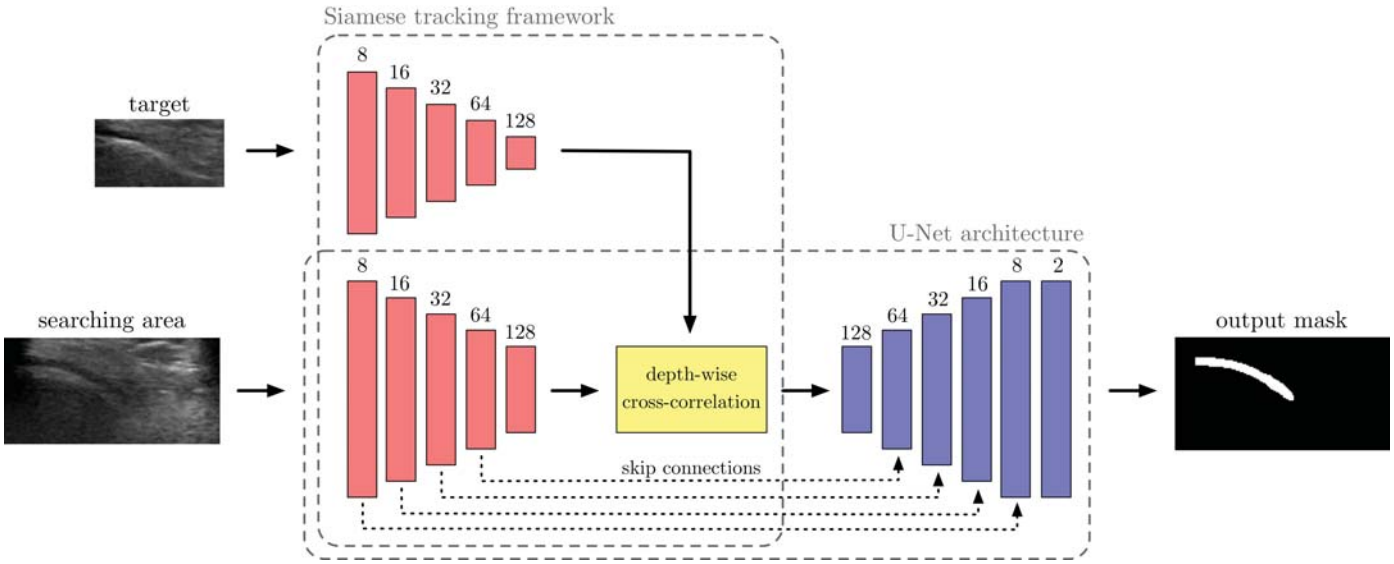

Fig. 5. Graphical visualization of the novel DL architecture, Siam-U-Net, proposed to track the femoral condyle cartilage. The network takes as input the target and the searching area (showed on the left) which are passed through the encoder $E_{\theta_E}(\cdot)$ represented by the red blocks. Then the target representation is depth-wise cross-correlated to the searching area representation. This operation encodes the information regarding the relative position of the cartilage inside the searching area. This embedding is combined with the intermediate feature maps produced by the encoder on the searching area (skip connections), and it is used by the decoder $D_{\theta_D}(\cdot)$ (blue blocks) to build the segmentation of the cartilage inside the searching area. The values above each block indicate the depth of the feature maps. The rectangles with dashed borders enclose the siamese tracking framework and the U-Net architecture that were used to create this novel network.

procedure is implemented as a convolutional layer applied to the searching area feature maps, using the target embedding as convolutional kernel. Zero-padding is applied to the cross-correlated feature maps to match the dimensions of the searching area embedding. The depth-wise cross correlation allows the comparison of the target cartilage image with the slice area where it is supposed to be present. The output of this operation encodes implicit information about the position of the cartilage inside the searching area into a three-dimensional representation, and is indeed a similarity map that is richer than the bi-dimensional one produced by the standard cross-correlation operation. Moreover, to make the correlation meaningful, the weights θ_E of the encoder are shared for the two input images. Since these belong to the same image domain, it makes sense to learn the same hierarchy of features and so to apply the same transformation to the two patches.

After the cross-correlation, the output binary mask is built by the decoder branch $D_{\theta_D}(\cdot)$ that uses four blocks composed sequentially of: the bilinear up-sampling of the feature maps of the previous layer, followed by a 2×2 convolution; a concatenation with the feature representations produced by each encoder block on the searching area (in the literature referred as to skip-connections); and two 3×3 convolutional layers. The latter are followed by batch normalization, ReLU and dropout. To generate the output segmentation, a 1×1 convolutional layer with two output channels is employed after the last block. The first output channel is for the prediction of the foreground object. i.e. the cartilage, while the second one is for the prediction of the pixels belonging to the background of the slice. This last layer is followed by a softmax activation function. The idea here is to refine the high level similarity map produced by the depth-wise cross-correlation operation through the layers of the decoder. Skip connections coming from the searching area branch are used to provide lower level (hence, more detailed) feature context and consequently compute a more fine-grained segmentation of the cartilage in the searching area.

In contrast to U-Net, which uses blocks with 64, 128, 256, 512, 1024 convolutional feature maps respectively, we implemented lighter blocks (i.e. they are composed of a smaller number of parameters) with 8, 16, 32, 64, 128 convolutional feature maps respectively. This modification was done to reduce the computational effort and improve the processing speed of the

network. In addition, we took advantage of the dropout layer to improve generalization.

4.2. Training Procedure

We trained Siam-U-Net end-to-end using the US data acquired as described in Section 3.1. To compose the training mini-batches, two slices belonging to the same subject, to the same leg, to the same US scanning modality and to the same 3D+time sequence were sampled. The first sampled slice was chosen inside the volume of temporal index $t-1$ at slice index j, i.e. $v_{i,j}^{(t-1)}$, while the second sampled slice was randomly chosen among

$$\left\{ v_{i,k}^{(t-1)}, v_{i,k}^{(t)} \;\middle|\; \begin{array}{l} v_{i,k}^{(t-1)}, v_{i,k}^{(t)} \in V_{i,j}, \\ k \in \{j - S_{max}, \ldots, j-1, j, j+1, \ldots, j+S_{max}\}, \\ S_{max} \in \mathbb{N} \end{array} \right\} \tag{1}$$

that is the set of spatially near slices that either belong to the $(t-1)$-th or to the t-th volume. Each mini-batch is composed of B pairs, sampled uniformly from intra-volume and inter-volume slices. We believe that useful information for the temporal tracking can be acquired also intra-volume (e.g. from the cartilage anatomical variations between spatially near slices), as this setting could provide changes of the cartilage appearance that are similar to the ones that could be found in inter-volume tracking. In addition, this process allows to augment the number of training samples, with the potential of improving generalization.

Before being fed to the SNN, both target and searching area were resized to *height*×*width*×*channels* (in practice [48×80×1] pixels for the target and $[64 \times 160 \times 1]$ pixels for the searching area) by respecting the aspect ratio of the cartilage. The fixed size for the searching area was obtained by assuring that: 1) the resizing process of the cropped slice would not alter the visual aspect of the cartilage; and 2) the feature maps produced by the encoder $E_{\theta_E}(\cdot)$ would be large enough to contain meaningful information. In a similar way, in order to guarantee that the target representation was informative enough, we used resizing dimensions that satisfied the architectural constraints (imposed by the max-pooling operations that halve the feature maps' dimensions) of the encoder and that allowed the feature maps to keep enough spatial information.

The training objective was set to reduce the DSC dissimilarity (Milletari et al., 2016), referred as to DSC loss, between the masks outputted by the network and the reference segmentation of the second slice of the input pair. This is novel in the panorama of VOS, where the Cross Entropy (CE) loss is often utilized. The use of the DSC loss as training cost is motivated by its robustness against class imbalance.

4.3. Tracking Procedure

In this section we describe how the presented network is employed to continuously track the knee cartilage in a 2D+time sequence.

Given a US sequence, two temporal consecutive slices at each time step are considered. For the first one a segmentation estimate is known, while for the second one it must be produced by Siam-U-Net. Given $v_{i,j}^{(t-1)}, v_{i,j}^{(t)} \in V_{i,j}$ as consecutive slices and

$$b^{(t-1)} = [x_{tl}^{(t-1)}, y_{tl}^{(t-1)}, x_{br}^{(t-1)}, y_{br}^{(t-1)}] \tag{2}$$

the smallest bounding box (defined by the top left and the bottom right vertices) enclosing the non-zero elements of the segmentation at time step $t-1$, $s_{i,j}^{(t-1)}$, the target crop is defined in $v_{i,j}^{(t-1)}$ as follows

$$b_{target}^{(t)} = [x_{tl}^{(t-1)} - P_1, y_{tl}^{(t-1)} - P_1, x_{br}^{(t-1)} + P_1, y_{br}^{(t-1)} + P_1], \tag{3}$$

where P_1 is a scalar that allows to enlarge the bounding box in order to include some context area around the cartilage segmentation. The searching area crop is obtained in $v_{i,j}^{(t)}$ as follows

$$b_{search}^{(t)} = [0, y_{tl}^{(t-1)} - P_2, c, y_{br}^{(t-1)} + P_2], \tag{4}$$

where P_2 is a scalar used to vertically increase the image context for this slice region. The definition of this crop area is based on two assumptions: 1) the physical layout of the data acquisition strongly limits vertical shifts of the cartilage and 2) the motion of the probe during US acquisition prevents the definition of horizontal shifts limits. Therefore, we selected the whole width of the slice and a limited vertical zone expressed by P_2 as crop area. The two cropped images are fed to the Siam-U-Net which outputs the binary segmentation that locates

the cartilage inside the searching area. The output mask $s_{i,j}^{(t)}$ is constructed by placing Siam-U-Net's output mask inside a matrix filled with zeros at the coordinates of $b_{search}^{(t)}$.

At the beginning of the tracking process, the known estimate of the cartilage, $s_{i,j}^{(0)}$, is set to be the reference contour $g_{i,j}^{(0)}$, i.e. $s_{i,j}^{(0)} := g_{i,j}^{(0)}$. In the next step, the segmentation produced by the network, $s_{i,j}^{(1)}$, is used to crop the target and the search area inside the slices $v_{i,j}^{(1)}, v_{i,j}^{(2)}$ respectively. This process is then repeated for all the slices that compose the sequence. The described procedure is depicted in Figure 6.

5. Experimental Setup

In this section we first report how the experimental datasets and procedures have been set up. Then we discuss the error measures employed to validate our methodology. Finally, we present the details of the implementation of the training and tracking procedures.

5.1. Dataset Splits

To validate the performance of our solution, we performed a cross validation across the different subjects that compose our US dataset. To this end, we ran six different experiments, where in each one we considered five subjects (80%) for training and one for testing (20%). To optimize the architecture and training hyper-parameters, we ran a first experiment using four subjects for training, one for validation and one for testing. This training, validation and test split was optimized in order to obtain sets with the most similar distributions of samples with respect to the different types of US scans. After their optimization, the hyper-parameters were kept fixed across the six experiments. In Figure 7 the distributions of the 2D slice samples considered in the six experiments are shown. Each subject $X \in \{1, .., 6\}$ is used as test subject in the Split X experiment.

5.2. Testing Sequences

To evaluate the performance of our methodology we ran Siam-U-Net on all the 2D+time sequences of the subjects who were chosen for testing. In particular, given the sequence $V_{i,j}$ and the initial segmentation $g_{i,j}^{(0)}$ for the slice $v_{i,j}^{(0)}$, we let the tracker run until the end of the sequence, i.e. $\forall v_{i,j}^{(t)} \in V_{i,j}, t > 0$.

Table 2. Summary of the test sequences for the temporal tracking setting. Each column reports respectively: the number of test sequences; the total number of slices that have been processed; the average number (± standard deviation) of slices that composed the sequences (i.e. circa 4 slices); the minimum and maximum number of slices in the sequences.

Split	# sequences	# slices	Average sequence length	Min-max sequence lengths
1	849	2224	3.62 ± 1.4	2-6
2	746	1759	3.36 ± 1.1	2-6
3	620	1533	3.47 ± 1.4	2-6
4	720	1701	3.36 ± 1.0	2-5
5	957	2626	3.74 ± 0.8	2-5
6	414	1127	3.72 ± 0.9	2-5

Table 3. Summary of the test sequences for the spatio-temporal tracking setting.

Split	# sequences	# slices	Average sequence length	Min-max sequence lengths
1	849	13633	17.06 ± 11.9	2-69
2	746	9356	13.54 ± 10.3	2-54
3	620	8535	14.77 ± 11.2	2-66
4	720	9808	14.62 ± 10.7	2-54
5	957	14070	15.70 ± 10.5	2-61
6	414	5892	15.23 ± 10.2	2-54

We then compared each produced prediction mask $s_{i,j}^{(t)}$ with the corresponding reference $g_{i,j}^{(t)}$. In VOT literature this evaluation procedure is referred as to one-pass evaluation (OPE) (Wu et al., 2013).

To assess the tracking capabilities of our solution, we set up two testing settings. For the first, we considered all the 2D+time sequences in which each slice belongs to the same volunteer, the same volunteer's leg, the same angle of scanning and the same 3D+time sequence, but to temporally consecutive US volumes. In this way we can assess the *temporal* tracking capabilities of our solution.

With the second procedure, each pair can include slices belonging either to a consecutive or to the same volume. In the latter case, if $v_{i,j}^{(t)}$ is the first slice of the pair, the second slice is chosen as the nearest slice $v_{i,j\pm1}^{(t)} \in V_{i,j\pm1}$ inside the volume at temporal step t. Given $v_{i,j}^{(t)}$, the pairing slice is randomly selected between $v_{i,j}^{(t+1)}$ and $v_{i,j\pm1}^{(t)}$ using a uniform distribution. We refer this setting as to *spatio-temporal* tracking.

Tables 2 and 3 summarize the test sequences used for each split.

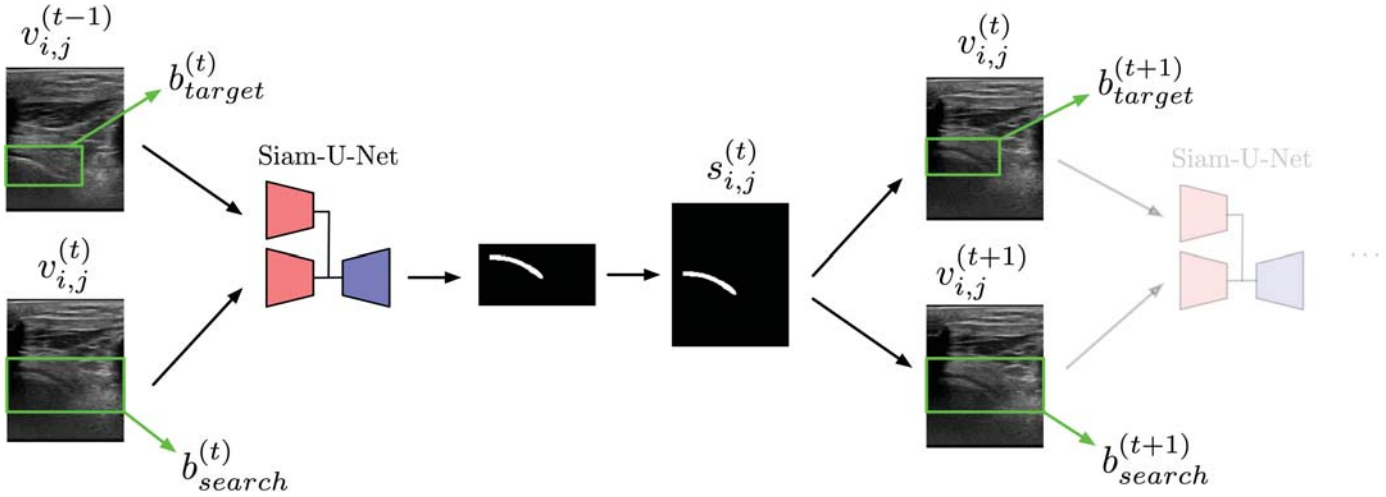

Fig. 6. Schematic view of the proposed cartilage tracking procedure. On the left, the two consecutive slices $v_{i,j}^{(t-1)}, v_{i,j}^{(t)}$ are cropped by the bounding boxes $b_{target}^{(t)}$ and $b_{search}^{(t)}$ (represented in green), respectively. The two cropped images are fed to Siam-U-Net, which produces the segmentation of the target cartilage inside the searching area. The prediction mask $s_{i,j}^{(t)}$ is then assembled by placing the output mask at the coordinates of $b_{search}^{(t)}$. $s_{i,j}^{(t)}$ is later used to compute $b_{target}^{(t+1)}$ and $b_{search}^{(t+1)}$ in order to crop the slices $v_{i,j}^{(t)}$ and $v_{i,j}^{(t+1)}$

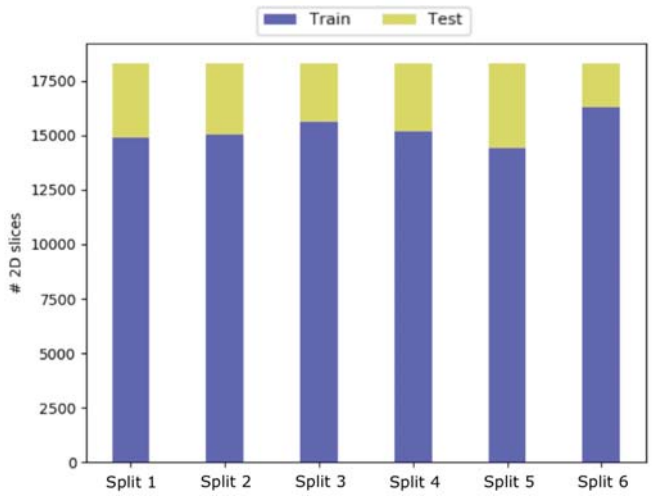

Fig. 7. Summary of the ratios of training and testing samples in the different experiments done.

5.3. Error Measures

For both the temporal and spatio-temporal tracking settings, we measured the DSC (Dice, 1945; Sørensen, 1948) between the predictions of Siam-U-Net and their respective reference segmentations. The DSC is a set similarity score that ranges in $[0, 1]$, which is measured as two times the number of overlapping pixels between two binary segmentations, normalized by the sum of the total number of pixels contained in the two. A DSC equal to 0 means that the two segmentations do not overlap, while a DSC of 1 defines a perfect overlap situation. The use of this index was motivated by the fact that it is agnostic to the size of the segmentation. Comparing to a distance-based measure (e.g., Hausdorff distance), DSC enables the computation of results in situations where objects have varying dimensions, which is the case for our problem. Across different slices, the cartilage can be very small (composed of around 4 pixels) or occupy a much larger part of the field of view (up to 1403 pixels). Computing the mean and standard deviation of the Hausdorff distance in this scenario would result in a widespread distribution, hiding the real amount of error made by the model.

As an aggregate metric, we computed the average value (along with standard deviation) of the DSC across all the slices for which a prediction is given by Siam-U-Net. Additionally, the boxplots containing the information regarding the median, the upper and lower quartiles, and range of the DSC values are reported.

Furthermore, we build the success plots for the two testing settings. The success plot (Wu et al., 2013) is used in VOT to evaluate the accuracy of a tracker and it is built by counting the number of frames that obtained a positive prediction. A prediction is considered positive if the intersection-over-union (IOU) between the predicted and the ground-truth bounding-boxes is above some threshold defined in the range $[0, 1]$, otherwise the

prediction is negative. Varying the thresholds for the IOU, different values of accuracy are obtained. With enough samples, Wu et al. (2013) showed that the area under the curve (AUC) of the success plot tends to be the average IOU. For our purposes we followed a similar approach, presenting a setup substituting the IOU with the DSC.

5.4. Evaluation Procedures

To extensively assess the performance of our methodology we employed six evaluation setups.

Evaluation 1. In this first setup, we evaluated the general performance of our methodology by running Siam-U-Net on all the temporal and spatio-temporal 2D+time sequences obtained from the testing subject's 3D+time sequences. The predicted segmentations were compared with the respective references using the DSC. The distribution of the predictions was assessed by mean, standard deviation, boxplots and success plots. The processing speed of the network was also determined, by measuring the processing time (in milliseconds) to obtain a prediction. The reciprocal of the average measured time was used to express the number of slices-per-second. Finally, qualitative examples of the predictions were obtained. In this setup, the general capabilities of tracking the cartilage, in real-time, were evaluated.

Evaluation 2. To make sure that Siam-U-Net developed a tracking performance which is consistent and robust through time, we evaluated the performance of our solution at different temporal steps. For the temporal tracking setting, we evaluated the distribution of the DSC after every prediction (i.e. when $t = 1, 2, 3, 4, 5$) by measuring mean and standard deviation. In the spatio-temporal setting instead, the same distribution was evaluated after each temporal step (i.e. when two consecutive slices belonged to different volumes), and after $J = 1, 3, 5, 10, 15$ slices processed inside each volume $\boldsymbol{v}_i^{(t)}$. Both results were obtained considering the DSC distributions across all the six experiments.

Evaluation 3. To further establish that Siam-U-Net learned an effective tracking ability through its architectural modules, a quantitative and a qualitative examinations were performed on the siamese encoder $E_{\theta_E}(\cdot)$ and the decoder $D_{\theta_D}(\cdot)$. In the first setting, we measured the mean DSC and standard deviation considering the scenario where the $E_{\theta_E}(\cdot)$'s branch processing the target cartilage is not active. This was done by replacing $E_{\theta_E}(\cdot)$'s branch intermediate features with a zero filled tensor, before being inputted to the depth-wise cross correlation layer. In this way, we can assess the importance of the information encoded by the target patch branch, and the robustness of the searching area branch in providing meaningful features for producing segmentations without the target cartilage. For the second setup instead, given a target cartilage image, different runs of Siam-U-Net with vertically shifted searching areas were performed. The activations of the decoder's feature maps after block 2, 4 and the output respectively, were visualized as heatmaps by reducing the range of the computed values in [0, 1] (by subtracting the minimum of the values and then dividing by the width of the range). The intention of this test was to examine the decoder's learned features in reflecting effectively the position variations of the target cartilage inside the searching areas.

Evaluation 4. To support the use of the DSC as training loss, a comparison between Siam-U-Net trained with the DSC loss and the same network trained using the CE loss was done. For the CE loss setting, the same architectural and training hyper-parameters used for the DSC loss were maintained. The two different networks were then tested by measuring the average DSC and standard deviation using the temporal test sequences presented in Table 2. The predictions of the two obtained models were also evaluated qualitatively.

Evaluation 5. The assessment of Siam-U-Net against the expert performance was based on a comparison with the intra-operator error. Six US volumes (two for every scanning modality) were re-annotated by Operator 1, and a second expert (Operator 2) was asked to contour them. The volumes were randomly chosen by making sure that they would vary among different volunteers, legs and scanning angles. In two separate sessions, each expert was provided with one volume at a time

and asked to contour the cartilage on each of the sagittal US slices comprised in that volume. This was done to measure the annotator consistency in outlining the femoral cartilage, avoiding the introduction of other possible sources of variability in the intra-observer study. After that, the DSC between the new and the reference annotations was computed in order to estimate the experts' consistency. The distribution was again evaluated through mean, standard deviation and a boxplot. We also assessed the p-values of a two-sample test (Welch, 1947) to evaluate the correlation between the DSC distributions of: Operator 1 and Operator 2; Siam-U-Net and Operator 1; Siam-U-Net and Operator 2.

Evaluation 6. To further validate our proposed methodology, a comparison with state-of-the-art segmentation models was performed. In particular, we implemented U-Net following the architectural details provided by Ronneberger et al. (2015). U-Net was trained by optimizing the DSC loss with the Adam optimizer (Kingma and Ba, 2014) for 30 epochs with an initial learning rate of 10^{-4} that was successively halved at epochs 10 and 20. Batches of 24 slices were used. A weight decay of $5 \cdot 10^{-4}$ was also added as regularization term. A comparison with the solution of Léger et al. (2018) was also performed. As suggested by the authors, an extra input channel containing a binary mask of the cartilage was added to U-Net's architecture. The proposed model was trained to perform cartilage's contour propagation. Given as inputs a previously known segmentation of the cartilage and a US slice, the network shall predict the segmentation that localizes the cartilage inside the US image. The model was trained with the same hyperparameters used for U-Net except for the number of epochs, that was set to three. During training, for each sample, the input binary mask was selected among the 10 reference segmentations $\{g^{(t)}_{i,j+k}, k \in \{-10, \dots, 0\}\}$ adjacent to slice $v^{(t)}_{i,j}$, as detailed by the authors. At test time, the mask outputted by the network at each step is later used as input segmentation at the successive prediction. In addition to the tests above, we performed a comparison with two VOS state-of-the-art methods. In particular, we implemented the solutions of Caelles et al. (2017) and of Oh et al. (2018), which are referred as to OSVOS and RGMP respectively. The former is currently the best performing solution in the single-object VOS panorama, while the latter is the best in terms of processing speed and it is also the solution most similar to Siam-U-Net, as both use SNNs. Both methodologies publicly provided their source code and we adapted them to the acquired US data. Six experiments were run using 5 subjects for training and one for testing, as done for Siam-U-Net. In each experiment, RGMP was trained for 10 epochs using all the 2D+time US sequences, obtained from the training subjects. The only modifications to OSVOS were the use of the Adam optimizer (Kingma and Ba, 2014) (instead of the Stochastic Gradient Descent algorithm), the learning rate of 10^{-4} and the number of epochs (500). These were done in order to reduce the online training time (from 10 minutes to circa 3).

For all the experimental setups, after training, the models were then tested with the 2D+time sequences obtained from the testing subject in the temporal tracking setting (which were presented in Table 2). As done for Siam-U-Net in Evaluation 1, the average DSC, standard deviation, boxplots and the number of slices-per-second were measured.

5.5. Implementation Details

In this section we report the results of the hyperparameters search which led to the best performance on the validation set.

Before being fed to the neural network, the target and searching area were resized to $[48 \times 80 \times 1]$ pixels and $[64 \times 160 \times 1]$ pixels, respectively. In our dataset, the average dimensions of the bounding boxes enclosing the target were 36 pixels in height and 72 pixels in width. The average dimensions for the searching areas were 40 pixels and 160 pixels. The padding values were set to $P_1 = 8$ pixels and $P_2 = 20$ pixels. Successively, the cropped and resized images were normalized by dividing each pixel value by 255. Before the cropping and resizing of the target and the searching area, each slice and its respective reference mask were resized to $[196 \times 160 \times 1]$ pixels to improve the speed of the network while processing smaller images. The dimensions were chosen making sure that the resized slices had an aspect ratio similar to the original slices. Using the valida-

tion set, we evaluated that this resizing process caused a performance loss (in terms of DSC) of around 1%, but it allowed an improvement of ×1.6 in the processing speed of our solution.

The model was trained for 75000 iterations using the Adam optimizer (Kingma and Ba, 2014). The initial learning rate was set to 10^{-4}, and then halved two times, at iterations 45000 and 60000, respectively. A weight decay of 0.0005 was also added to the DSC loss as regularization term. Each mini-batch was composed of $B = 64$ pairs. In the composition of training pairs, the number of possible nearest slices S_{max}, was set to 10. We experimented removing the constraint of choosing just the S_{max} nearest slices and instead we composed training pairs of random inter and intra volume slices. The motivation for this was to learn the most generic transformations of the cartilage, however this setup did not achieve good performance. The rate of the Dropout layer was set to 0.4.

At test time, no online update of the network's parameters was performed. Additionally, the foreground output masks $s_{i,j}^{(t)}$, that had a size of $[196 \times 160 \times 1]$ pixels were resized to match the size of the reference segmentations, which is $[313 \times 255 \times 1]$ pixels.

Experiments have been conducted running our Python code with the PyTorch (Paszke et al., 2019) machine learning framework on an Intel Xeon E5-2690 v4 @ 2.60GHz CPU with 320 GB of RAM, four NVIDIA TITAN V GPUs and an NVIDIA TITAN Xp GPU each with 12 GB of memory. The training took around 7 hours.

6. Results and Discussion

Evaluation 1. In Table 4 and in Figure 8, we show the results achieved for Evaluation 1.

The average DSC across all experiments is 0.70 ± 0.16 for the temporal tracking setting while it is 0.71 ± 0.16 for the spatio-temporal setting. The median averaged between the six experiments resulted in 0.75 for both settings. The boxplots show compact distributions of the predictions. The low difference between the results of the two settings suggests that the proposed model is robust to the increased length of the sequences and it is able to overcome the variations of the cartilage appearance both in inter and in intra volume scenarios.

Table 4. Results of Siam-U-Net obtained, on Evaluation 1, for the temporal (left column of results) and for the spatio-temporal (right column) tracking settings.

Split	Temporal tracking average DSC	Spatio-temporal tracking average DSC
1	0.74 ± 0.15	0.73 ± 0.16
2	0.69 ± 0.20	0.71 ± 0.16
3	0.69 ± 0.16	0.70 ± 0.15
4	0.69 ± 0.17	0.68 ± 0.18
5	0.73 ± 0.14	0.73 ± 0.14
6	0.69 ± 0.15	0.68 ± 0.16
Total	0.70 ± 0.16	0.71 ± 0.16

The results here obtained do not depend on the dataset split, thus on the subject, the knee and the scan type. This indicates that our solution captures the variability that occurs among different subjects and is able to generalize well to new cases.

The success plots for the temporal and spatio-temporal experimental scenarios are presented in Figure 9. It can be seen that Siam-U-Net presents a high percentage (> 80%, on the vertical axis) of predictions that have a DSC with the reference of at least 0.6 (shown on the horizontal axis). When more precise segmentations are considered, i.e. with a DSC > 0.6, the performance of our methodology quickly drops. This is in part explained by the fact that the number of pixels that compose the segmentations of the cartilage is very low with respect to the number of pixels in the slices (as an average computed on the entire dataset, just ∼1% of all pixels belong to the cartilage). This causes the DSC to decrease rapidly if just a few pixels are misclassified by the algorithm.

In terms of speed, our solution runs at ∼90 slices-per-second on the machine detailed in Section 5.5. Since in the computer vision literature, 25-30 frames-per-second are considered real-time performance, we can state that Siam-U-Net is able to run in real-time.

In Figure 10 we present some qualitative results of our proposed solution. In the left block of the figure, going from left to right the three images show respectively the US slice $v_{i,j}^{(t-1)}$, $v_{i,j}^{(t-1)}$ with the reference segmentation $g_{i,j}^{(t-1)}$ (in pink), and $v_{i,j}^{(t-1)}$ with Siam-U-Net's prediction $s_{i,j}^{(t-1)}$ (in green) for the temporal

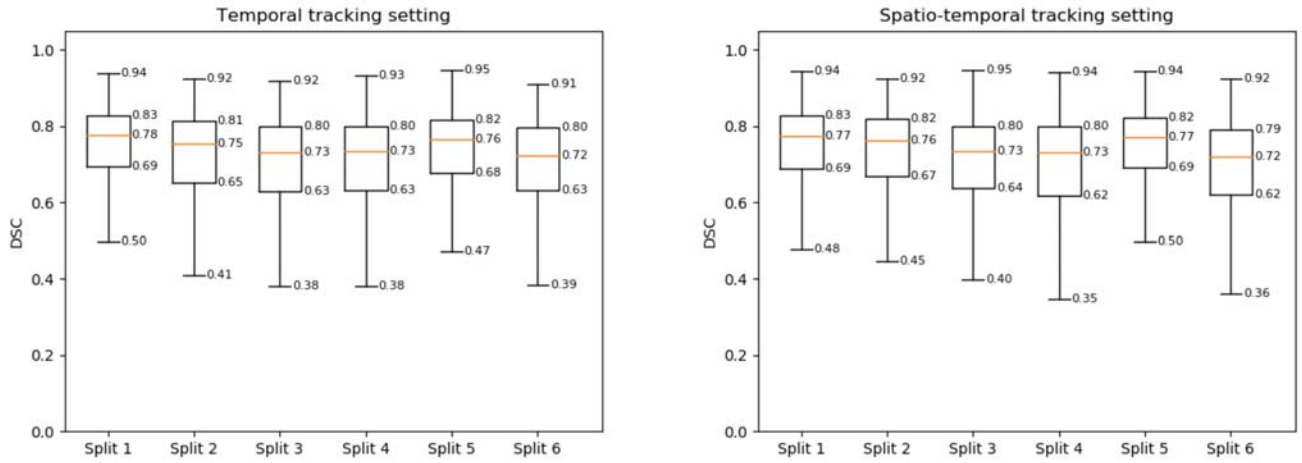

Fig. 8. Boxplots for Evaluation 1. Each boxplot shows the DSC distribution per experiment. On the left, the plots for the temporal tracking setting are presented. On the right, the same plots but for the spatio-temporal setting.

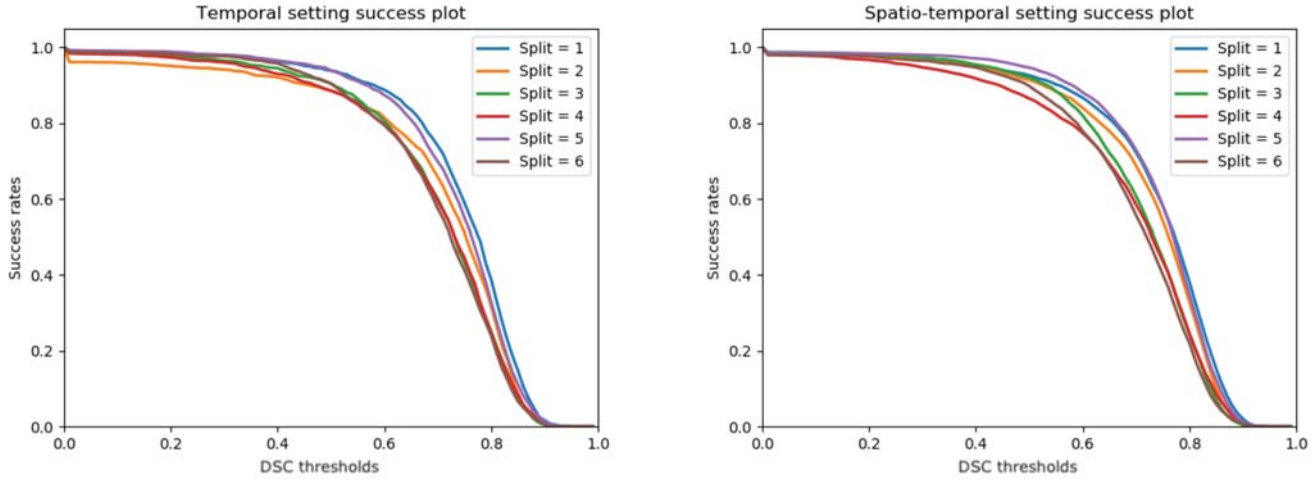

Fig. 9. Success plots, for Evaluation 1, of the temporal (left image) and of the spatio-temporal tracking settings (right image).

step $t-1$. In the right block, each image shows the same elements, but for the next temporal step t. Each row of the figure shows a different US sequence.

Evaluation 2. In Table 5, the results of the temporal tracking consistency evaluation are reported. After the first prediction, Siam-U-Net's DSC performance decreases by 4% on average, showing robustness for tracking. This result also shows that the proposed model has a small performance loss when it uses target patches that are not properly aligned with the actual shape and position of the cartilage, i.e. they propagate some error from previous predictions. With this performance, we can say that Siam-U-Net's tracking ability is also robust to target initialization errors.

In Table 6 we present the results of the consistency assessment in the spatio-temporal setting. Apart for $J = 1, 3$, the performance tend to increase after $J = 5, 10, 15$ slices processed inside the same volume. This demonstrates that tracking through space is easier than tracking through time because of less spatial and appearance changes of the cartilage. After the first processed slice, i.e. $J = 1$, Siam-U-Net's performance decreases by 3.25% across the different volumes, which is consistent with the results presented in Table 5. The lower temporal performance loss, together with the general increase of the average DSC across spatial predictions, suggest that tracking in

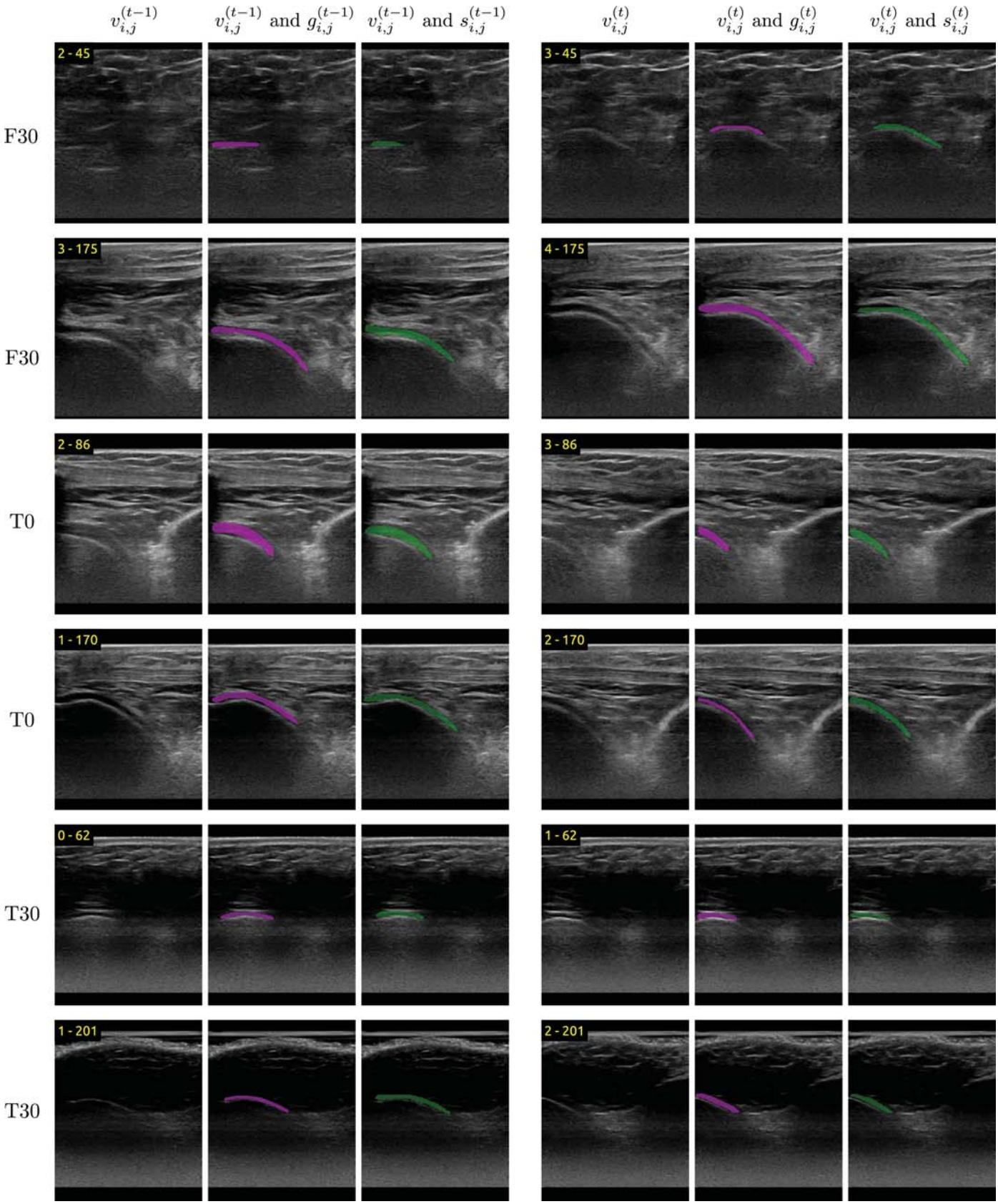

Fig. 10. Qualitative results of our proposed algorithm. The left block, composed of three images, shows respectively the US slice, the US slice with the reference segmentation (in pink) and the US slice with the prediction of our algorithm (in green) for the step $t-1$. In column on the right, the US slice, the US slice with the reference segmentation and prediction for the successive step t are presented. Each row corresponds to a different test sequence. On the left of each row of images, the knee scan modality is reported. The two yellow numbers indicate, respectively, the temporal index t and the slice index j.

Table 5. Results of Evaluation 2. Mean DSC and standard deviation computed at the different temporal steps t in the temporal tracking setting.

	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
DSC	0.73 ± 0.13	0.69 ± 0.18	0.70 ± 0.18	0.68 ± 0.18	0.69 ± 0.18

Table 6. Results of Evaluation 2. Mean DSC and standard deviation computed at the different temporal volume indexes t and different spatial indexes J in the spatio temporal tracking setting.

	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
$J = 1$	0.74 ± 0.13	0.71 ± 0.15	0.69 ± 0.19	0.73 ± 0.16	0.70 ± 0.15
$J = 3$	0.71 ± 0.15	0.71 ± 0.15	0.70 ± 0.18	0.73 ± 0.14	0.71 ± 0.15
$J = 5$	0.70 ± 0.16	0.72 ± 0.15	0.71 ± 0.17	0.73 ± 0.15	0.74 ± 0.11
$J = 10$	0.71 ± 0.16	0.72 ± 0.15	0.70 ± 0.17	0.75 ± 0.14	0.73 ± 0.10
$J = 15$	0.72 ± 0.15	0.72 ± 0.15	0.70 ± 0.17	0.74 ± 0.15	0.73 ± 0.08

space can help to reconstruct better target and searching area patches which in turn can lead to more accurate future predictions.

In general, Siam-U-Net loses some accuracy with the increased length of the sequences, but the results indicate that our proposed network is able to behave well in situations where different kinds of cartilage motion happen. In particular, we can say that Siam-U-Net developed the capability of overcoming both rigid and non-rigid transformations of the cartilage, the former depending on external events such as probe translations, while the latter depending on the changing aspect of the inner anatomical structures while moving the knee. Thus, the proposed solution effectively learned how the cartilage transforms between consecutive slices. This conclusion can be further supported by the performance on the spatio-temporal experimental setting in which Siam-U-Net had to track the cartilage both between temporal consecutive slices (in which the cartilage shape changed due to the events described above) and the spatially nearest slices (the cartilage shape varies within the acquired volumes).

With respect to the latter situation, we believe that our methodology could be also used, as an operator-aided system, to segment US volumes or portions of them. In this scenario, the system could be inputted with just an initial 2D reference segmentation that would be then propagated iteratively to the spatially nearest slices, ultimately producing a volumetric segmentation.

Table 7. Evaluation 3. Mean DSC and standard deviation results of executing Siam-U-Net with the target image patch branch disabled.

Split	Siam-U-Net	Siam-U-Net without target branch
1	0.74 ± 0.15	0.35 ± 0.31
2	0.69 ± 0.20	0.18 ± 0.23
3	0.69 ± 0.16	0.17 ± 0.26
4	0.69 ± 0.17	0.14 ± 0.24
5	0.73 ± 0.14	0.26 ± 0.27
6	0.69 ± 0.15	0.60 ± 0.26
Total	0.70 ± 0.16	0.28 ± 0.26

Evaluation 3. Table 7 displays the results of the quantitative evaluation with the encoder's target patch branch disabled. The high discrepancy with the results of the complete architecture demonstrates that previous visual information embedded by the encoder on the target patch is necessary to provide a correct segmentation of the cartilage. This test shows the significance of the temporal information coming from the target patch in the previous slice, with respect to the appearance information of the cartilage included in the current slice.

In Figure 11 the qualitative analysis of the Siam-U-Net's decoder feature activations is shown. While maintaining the same target patch, the original searching area (i.e. the one obtained by the bounding box $b^{(t)}_{search}$) and the vertically down shifted searching area are considered. It can be noticed how the activations and the output mask reflect the shift happening in the searching area. This result suggests that the decoder learned to refine the high level localization map produced by the depth-wise cross correlation operation and thus localize effectively the target cartilage in searching areas.

In contrast to classical statistical approaches for tracking

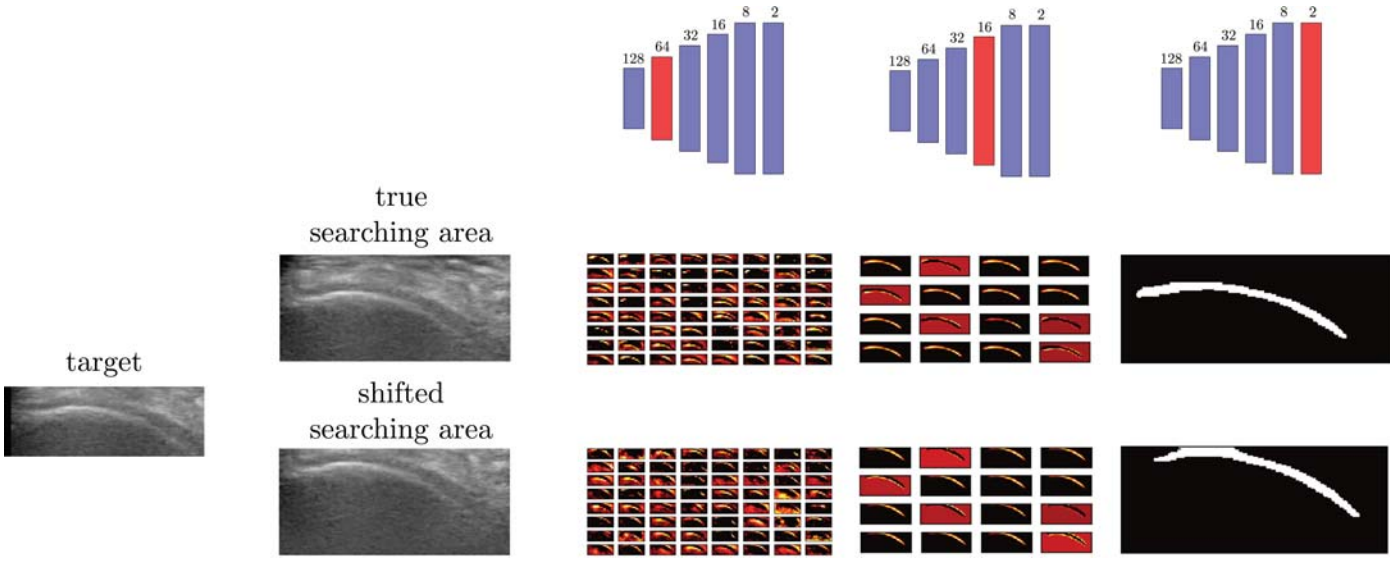

Fig. 11. Qualitative analysis of the Siam-U-Net's decoder feature activations at different positions of the cartilage. For the same target cartilage slice patch, two vertically shifted searching areas are inputted to Siam-U-Net. The intermediate features of the decoder (which belonging layers are highlighted in red in the first row of pictures) and the output mask reflect the shift happening in the searching area, suggesting that our solution effectively learned to localize the target cartilage.

where the trade-off between motion and appearance models are in general controllable, in our setting the balance between the two is learned inherently during training. As pointed by Pflugfelder (2017), SNN-based trackers integrate easily into a single network different tracking-related tasks, such as feature extraction, matching and localization. The proposed Siam-U-Net is an example of that. Although some work has been done from a theoretical point of view (Pflugfelder, 2017), we are not aware of papers that have studied the capabilities of SNN module in VOS. An extensive study to analyze in depth how to control the architectural components of SNN for tracking is out of the scope of this paper, but by presenting the results of Evaluation 5, we tried to provide a preliminary explanation on the impact of the target branch in the segmentation of the object and of the higher level features that are learned by the decoder.

Evaluation 4. In Table 8 we present the comparison between Siam-U-Net trained with the DSC loss and Siam-U-Net trained using the CE loss. The employment of the DSC loss allowed us to produce a more accurate and stable tracking between the different subjects. Through a visual inspection of the resulting segmentations we noticed that the majority of the failure cases of Siam-U-Net trained with the CE loss happened when the hypoechoic and hyperechoic lines of the cartilage were not clearly distinguishable. In these cases, we believe that the CE loss does not produce a learning signal that is meaningful enough for the weak patterns present in these slices. In Figure 12 we show some examples of the described situations.

Table 8. Evaluation 4. Comparison of the results obtained in the temporal tracking setting by training Siam-U-Net with the DSC loss and the CE loss respectively.

Split	Siam-U-Net DSC Loss average DSC	Siam-U-Net CE Loss average DSC
1	0.74 ± 0.15	0.61 ± 0.24
2	0.69 ± 0.20	0.65 ± 0.21
3	0.69 ± 0.16	0.69 ± 0.16
4	0.69 ± 0.17	0.68 ± 0.18
5	0.73 ± 0.14	0.67 ± 0.18
6	0.69 ± 0.15	0.68 ± 0.18
Total	0.70 ± 0.16	0.66 ± 0.19

Evaluation 5. The DSC between the reference and the new segmentations annotated by Operator 1 resulted in 0.63 ± 0.30 and median DSC of 0.77. This result was consistent with Operator 2 that had a mean DSC of 0.61 ± 0.25 and median DSC of 0.69. In Figure 13 the boxplots for the two observer evaluations are given. It can be easily seen how widespread the two DSC distributions are. The p-value of the two-sample test between the DSC distributions of the experts resulted in 0.242, suggesting a correlation between the two. The comparison between Siam-U-

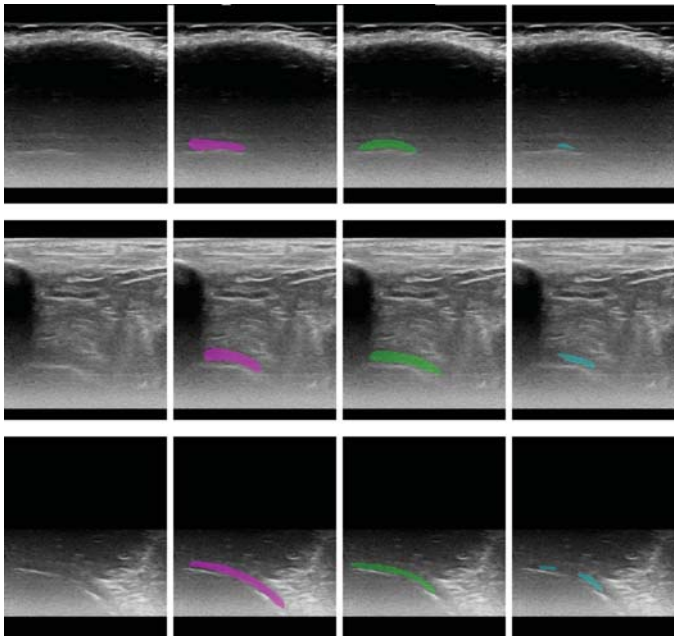

Fig. 12. Qualitative comparison of Siam-U-Net trained with either the DSC loss or the CE loss. From left to right, the first column of images shows the original US slices; the second the US slices with the reference segmentations; the third the predictions of Siam-U-Net trained with the DSC loss and the last column on the right the predictions of Siam-U-Net trained with the CE loss.

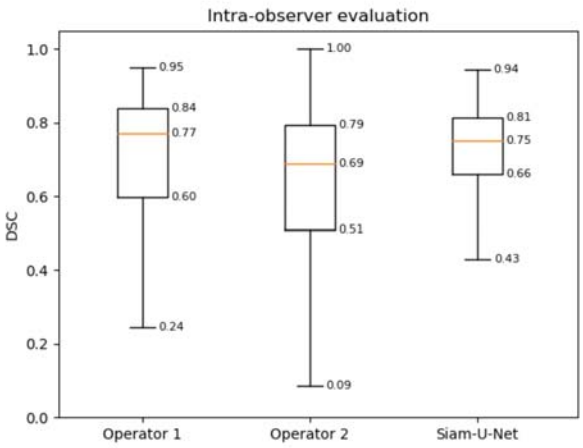

Fig. 13. Boxplots for the intra-observer evaluation (Evaluation 5) on the two expert operators and for Siam-U-Net. The boxplot for Siam-U-Net was obtained by considering all the predictions across the six dataset splits.

Net's and Operator 1's and Operator 2's performance achieved p-values of $3.41 \cdot 10^{-9}$ and $6.35 \cdot 10^{-15}$ respectively. This shows that there is no correlation between the performance of Siam-U-Net and the one of the experts. Given these results, we can say that Siam-U-Net has an average localization ability that is higher and more robust than the expert operators. The high intra-observer variability can be motivated by the effect of US physics on the knee cartilage, making its localization difficult. Due to US physics, the US beam has a better reflection when it perpendicularly intercepts the part of the cartilage which is flat and consequently it allows to produce an image with better quality in those regions. These situations make easier the distinction of the cartilage hypoechoic and hyperechoic lines. However, it is not the case when the beam intercepts the left and right extremes of the cartilage. Due to the non-perpendicularity of the cartilage walls in those areas, the transmitted US beam are subject to scattering. This leads to images where the cartilage structure is, partially or sometimes totally, not visible.

Evaluation 6. U-Net's mean and standard deviation DSC values are reported in Table 9 for the temporal tracking scenario while a boxplot is represented on the left plot of Figure 14. The average performance is 6% lower than Siam-U-Net, with widespread distributions resembling the expert operators' outcome. This worse performance can be in part explained by the class imbalance of pixel masks. Since U-Net has to predict more pixel probabilities (i.e. prediction masks have bigger dimensions than the ones of Siam-U-Net), it is more susceptible to mislabeling. This situation, together with the small percentage of pixels belonging to the cartilage, makes it easier to missegment the cartilage, increasing the spread of the distribution and decreasing the average performance. Similar conclusions can be reached for the solution by Léger et al. (2018). Regarding the processing time, U-Net predicts segmentations with an average speed of 45 slices-per-second, half the speed of Siam-U-Net, while the solution of Léger et al. (2018) runs at 35 slices-per-second. In summary, with respect to a tracking-by-segmentation approach used by the compared works, the use of previous temporal or spatial information and Siam-U-Net's architecture is definitely useful to speed up the tracking process and to provide a more accurate and consistent segmentation of the femoral condyle cartilage.

In Table 10 the results of Siam-U-Net against OSVOS and RGMP are reported. We suggest that the lower performance of both OSVOS and RGMP are caused by overfitting, due to the relatively small dataset used and the high capacity of the

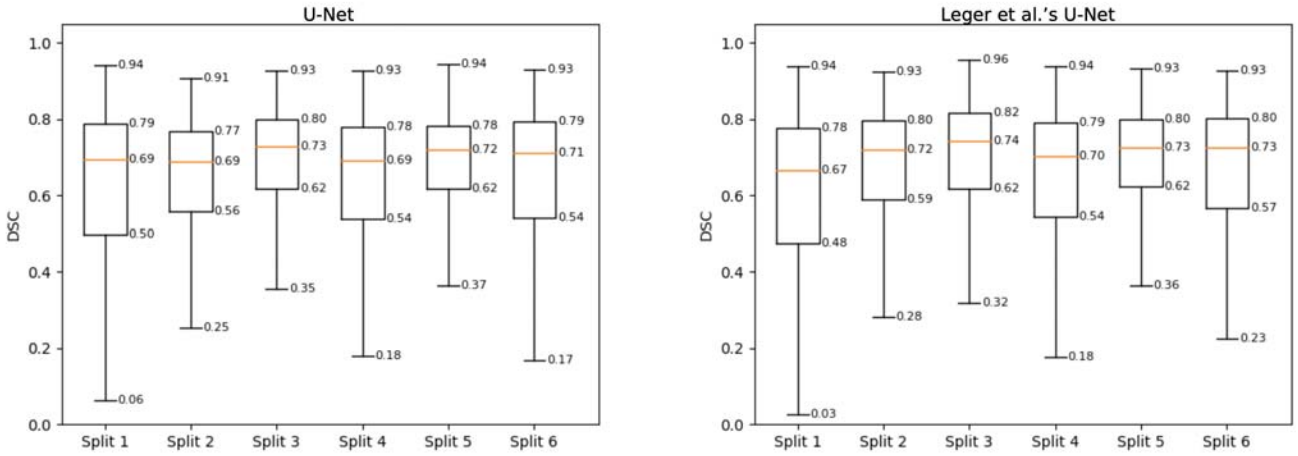

Fig. 14. Boxplots for the temporal tracking performance of U-Net (on the left) and of the solution of Léger et al. (2018) (on the right).

Table 9. Results of Evaluation 6. Comparison of Siam-U-Net performance against U-Net (Ronneberger et al., 2015) and the model proposed by Léger et al. (2018).

Split	Siam-U-Net average DSC	U-Net average DSC	Léger et al. (2018)'s U-Net average DSC
1	0.74 ± 0.15	0.61 ± 0.24	0.60 ± 0.23
2	0.69 ± 0.20	0.62 ± 0.22	0.65 ± 0.23
3	0.69 ± 0.16	0.68 ± 0.18	0.68 ± 0.21
4	0.69 ± 0.17	0.62 ± 0.23	0.64 ± 0.22
5	0.73 ± 0.14	0.66 ± 0.21	0.67 ± 0.20
6	0.69 ± 0.15	0.63 ± 0.23	0.62 ± 0.28
Total	0.70 ± 0.16	0.64 ± 0.22	0.64 ± 0.23

Table 10. Results of Evaluation 4. Comparison of Siam-U-Net performance against the state-of-the-art methods, OSVOS and RGMP, in the temporal tracking setting.

Split	Siam-U-Net average DSC	OSVOS average DSC	RGMP average DSC
1	0.74 ± 0.15	0.50 ± 0.30	0.24 ± 0.29
2	0.69 ± 0.20	0.43 ± 0.27	0.53 ± 0.24
3	0.69 ± 0.16	0.45 ± 0.27	0.49 ± 0.20
4	0.69 ± 0.17	0.45 ± 0.28	0.55 ± 0.23
5	0.73 ± 0.14	0.44 ± 0.27	0.51 ± 0.28
6	0.69 ± 0.15	0.50 ± 0.26	0.49 ± 0.25
Total	0.70 ± 0.16	0.46 ± 0.28	0.47 ± 0.25

models, that are composed by very deep CNNs. In terms of processing speed, the test revealed that RGMP had an average running time of around 38 slices-per-second, about two times slower than Siam-U-Net. OSVOS processed around 7 slices-per-second, with an additional time of 3 minutes for the on-line training that is performed before processing every 2D+time sequence. Siam-U-Net instead is trained solely offline and it can be applied straight away to any given sequence of images. Additionally, the end-to-end strategy employed by our solution permits also to simplify the training process and so to reduce its required time, since the pre-training phase done on ImageNet (Deng et al., 2009) by OSVOS and RGMP is not more necessary.

6.1. Limitations and Future Work

One of the main drawbacks of this work is the processing of 2D US images. An experienced clinician, when provided with volumetric data, usually exploits the information contained in neighbouring slices to interpret a 2D image. Siam-U-Net does not take advantage of this process, which has the potential to include more information and consequently allow a more accurate tracking of the cartilage. In the future, it could be interesting to adapt Siam-U-Net to work with 3D+time data, by combining a volumetric segmentation model like V-Net (Milletari et al., 2016) with the siamese tracking framework.

By a qualitative evaluation of Siam-U-Net's failure cases, we discovered some situations like shown in Figure 15. In these cases, the upper hyperechoic line of the cartilage is not clearly defined and causes Siam-U-Net to produce segmentations where similar cartilage patterns are present (in the area identified by the mid-left green segmentation of Figure 15). Since this wrong output becomes the input for next step, the error could be ulteriorly propagated. To resolve these circum-

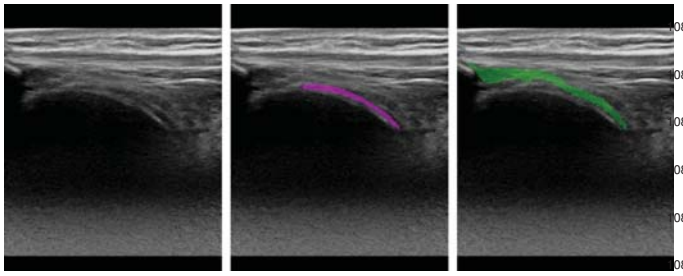

Fig. 15. A failure example of Siam-U-Net (depicted in green in the right image). In the left US image, it can be seen that the upper hyperechoic line of the cartilage is not clearly defined.

stances, since Siam-U-Net utilizes dropout layers, we could investigate the implementation of uncertainty estimations, in a similar fashion as done by Kendall et al. (2015). In the best case, with an high rate of segmentation uncertainty, Siam-U-Net could integrate some mechanism to ask for reinitialisation.

An in-depth analysis of the architectural components and the tracking capabilities of our proposed solution is a valuable reference for SNN-based trackers that we are planning to work in the next future. Another interesting future direction is the adaption of Siam-U-Net for user-aided segmentation of 3D volumes (US, CT, MRI).

From a clinical point of view, the acquired US data represents several possible scenarios in robotic knee arthroscopy, but not all of them. In particular, in this proof-of-concept work the most difficult and critical situations were replicated. Future studies will include temporally longer sequences and more angles of knee flexion. Furthermore, differently from the actual surgery, the image acquisition has been performed in water. In the future, a coupling device needs to be developed to avoid the presence of air gaps at the interface between the probe and the knee surface.

7. Conclusions

As the knee cartilage is one of the structures that is most at risk during MIPs, we demonstrated the feasibility of using a novel DL architecture to track in real-time the femoral condyle cartilage imaged with US, under simulated surgical conditions. The proposed DL architecture, Siam-U-Net, is the combination of neural networks for medical image segmentation and the siamese framework for visual tracking. We evaluated the proposed solution using the DSC against an expert surgeon and we obtained an average performance of 0.70 ± 0.16 in the temporal tracking setting. We also present experimental results for a spatio-temporal tracking setting, showing that our solution is robust to the high variability of the cartilage aspect under the considered conditions. The high intra-operator variability (intra-operator DSC of 0.63 ± 0.30 and 0.61 ± 0.25) suggests that there are some limitations in the maximum performance that can be achieved by the network. This can be attributed to the uncertainty in the ground-truth segmentations that is dependent to the physics of the US beam. Regarding the processing speed, our network is able to run at 90 slices-per-second on a GPU-provided machine. Given its speed and accuracy, we believe that Siam-U-Net has the potential for guiding surgeons or future autonomous robotic systems during MIPs.

Acknowledgements

This work was partially supported by the Australia-India strategic research fund AISRF53820 (Intelligent Robotic Imaging System for keyhole surgeries) and by the Australian Research Council project (DP180103232). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

Akgul, Y., Kambhamettu, C., Stone, M., 1999. Automatic extraction and tracking of the tongue contours. IEEE Transactions on Medical Imaging 18, 1035–1045. URL: http://ieeexplore.ieee.org/document/811315/, doi:10.1109/42.811315.

Antico, M., Sasazawa, F., Wu, L., Jaiprakash, A., Roberts, J., Crawford, R., Pandey, A.K., Fontanarosa, D., 2019. Ultrasound guidance in minimally invasive robotic procedures. Medical Image Analysis URL: https://www.sciencedirect.com/science/article/pii/S1361841519300027, doi:10.1016/J.MEDIA.2019.01.002.

Ben-Cohen, A., Diamant, I., Klang, E., Amitai, M., Greenspan, H., 2016. Deep learning and data labeling for medical applications, in: Proceedings of the International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. In: Lecture Notes in Computer Science, pp. 77–85.

Bertinetto, L., Henriques, J.F., Valmadre, J., Torr, P.H.S., Vedaldi, A., 2016a. Learning feed-forward one-shot learners. NEURAL INFO PROCESS SYS F. URL: https://dl.acm.org/citation.cfm?id=3157155.

Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S., 2016b. Fully-Convolutional Siamese Networks for Object Tracking URL: http://arxiv.org/abs/1606.09549, arXiv:1606.09549.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R., 1993. Signature Verification Using a "Siamese" Time Delay Neural Network, in: Proceedings of the 6th International Conference on Neural Information Processing Systems, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. pp. 737–744. URL: http://dl.acm.org/citation.cfm?id=2987189.2987282.

Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixe, L., Cremers, D., Gool, L.V., 2017. One-Shot Video Object Segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 5320–5329. URL: http://ieeexplore.ieee.org/document/8100048/, doi:10.1109/CVPR.2017.565.

Carneiro, G., Nascimento, J.C., 2013. Combining multiple dynamic models and deep learning architectures for tracking the left ventricle endocardium in ultrasound data. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 2592–2607. URL: http://ieeexplore.ieee.org/document/6517436/, doi:10.1109/TPAMI.2013.96.

Ce Liu, Yuen, J., Torralba, A., 2011. SIFT Flow: Dense Correspondence across Scenes and Its Applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 978–994. URL: http://ieeexplore.ieee.org/document/5551153/, doi:10.1109/TPAMI.2010.147.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, Springer, Cham, pp. 424–432. URL: http://link.springer.com/10.1007/978-3-319-46723-8{_}49, doi:10.1007/978-3-319-46723-8_49.

De Luca, V., Benz, T., Kondo, S., König, L., Lübke, D., Rothlübbers, S., Somphone, O., Allaire, S., Lediju Bell, M.A., Chung, D.Y.F., Cifor, A., Grozea, C., Günther, M., Jenne, J., Kipshagen, T., Kowarschik, M., Navab, N., Rühaak, J., Schwaab, J., Tanner, C., 2015. The 2014 liver ultrasound tracking benchmark. Physics in medicine and biology 60, 5571–99. URL: http://www.ncbi.nlm.nih.gov/pubmed/26134417http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5454593, doi:10.1088/0031-9155/60/14/5571.

Deng, J., Dong, W., Socher, R., Li, L.J., Kai Li, Li Fei-Fei, 2009. ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 248–255. URL: http://ieeexplore.ieee.org/document/5206848/, doi:10.1109/CVPR.2009.5206848.

Dice, L.R., 1945. Measures of the Amount of Ecologic Association Between Species. Ecology 26, 297–302. URL: http://doi.wiley.com/10.2307/1932409, doi:10.2307/1932409.

Faisal, A., Ng, S.C., Goh, S.L., George, J., Supriyanto, E., Lai, K.W., 2015. Multiple LREK Active Contours for Knee Meniscus Ultrasound Image Segmentation. IEEE Transactions on Medical Imaging 34, 2162–2171. URL: http://ieeexplore.ieee.org/document/7091031/, doi:10.1109/TMI.2015.2425144.

Faisal, A., Ng, S.C., Goh, S.L., Lai, K.W., 2018a. Knee cartilage segmentation and thickness computation from ultrasound images. Medical & Biological Engineering & Computing 56, 657–669. URL: http://www.ncbi.nlm.nih.gov/pubmed/28849317http://link.springer.com/10.1007/s11517-017-1710-2, doi:10.1007/s11517-017-1710-2.

Faisal, A., Ng, S.C., Goh, S.L., Lai, K.W., 2018b. Knee Cartilage Ultrasound Image Segmentation Using Locally Statistical Level Set Method, Springer, Singapore, pp. 275–281. URL: http://link.springer.com/10.1007/978-981-10-7554-4{_}48, doi:10.1007/978-981-10-7554-4_48.

Farabet, C., Couprie, C., Najman, L., LeCun, Y., 2013. Learning Hierarchical Features for Scene Labeling. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 1915–1929. URL: http://ieeexplore.ieee.org/document/6338939/, doi:10.1109/TPAMI.2012.231.

Giraldo, J.J., Alvarez, M.A., Orozco, A.A., 2015. Peripheral nerve segmentation using Nonparametric Bayesian Hierarchical Clustering, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 3101–3104. URL: http://www.ncbi.nlm.nih.gov/pubmed/26736948http://ieeexplore.ieee.org/document/7319048/, doi:10.1109/EMBC.2015.7319048.

Gomariz, A., Li, W., Ozkan, E., Tanner, C., Goksel, O., 2019. Siamese Networks with Location Prior for Landmark Tracking in Liver Ultrasound Sequences URL: http://arxiv.org/abs/1901.08109, arXiv:1901.08109.

Grundmann, M., Kwatra, V., Han, M., Essa, I., 2010. Efficient hierarchical graph-based video segmentation, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE. pp. 2141–2148. URL: http://ieeexplore.ieee.org/document/5539893/, doi:10.1109/CVPR.2010.5539893.

Guerrero, J., Salcudean, S., McEwen, J., Masri, B., Nicolaou, S., 2007. Real-Time Vessel Segmentation and Tracking for Ultrasound Imaging Applications. IEEE Transactions on Medical Imaging 26, 1079–1090. URL: http://ieeexplore.ieee.org/document/4280891/, doi:10.1109/TMI.2007.899180.

Guo, Q., Feng, W., Zhou, C., Huang, R., Wan, L., Wang, S., 2017. Learning Dynamic Siamese Network for Visual Object Tracking, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE. pp. 1781–1789. URL: http://ieeexplore.ieee.org/document/8237458/, doi:10.1109/ICCV.2017.196.

Hackel, J.G., Khan, U., Loveland, D.M., Smith, J., 2016. Sonographically Guided Posterior Cruciate Ligament Injections: Technique and Validation. PM&R 8, 249–253. URL: http://www.ncbi.nlm.nih.gov/pubmed/26247162http://doi.wiley.com/10.1016/j.pmrj.2015.07.008, doi:10.1016/j.pmrj.2015.07.008.

Held, D., Thrun, S., Savarese, S., 2016. Learning to Track at 100 {FPS} with Deep Regression Networks, in: European Conference on Computer Vision. URL: http://arxiv.org/abs/1604.01802, arXiv:1604.01802.

Hirahara, A.M., Andersen, W.J., 2016. Ultrasound-Guided Percutaneous Reconstruction of the Anterolateral Ligament: Surgical Technique and Case Report. American journal of orthopedics (Belle Mead, N.J.) 45, 418–460. URL: http://www.ncbi.nlm.nih.gov/pubmed/28005093.

Huang, X., Dione, D.P., Compas, C.B., Papademetris, X., Lin, B.A., Bregasi, A., Sinusas, A.J., Staib, L.H., Duncan, J.S., 2014. Contour tracking in echocardiographic sequences via sparse representation and dictionary learning. Medical image analysis 18, 253–71. URL: http://www.ncbi.nlm.nih.gov/pubmed/24292554http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3946038, doi:10.1016/j.media.2013.10.012.

Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift URL: http://arxiv.org/abs/1502.03167, arXiv:1502.03167.

Jaiprakash, A., O'Callaghan, W.B., Whitehouse, S.L., Pandey, A., Wu, L., Roberts, J., Crawford, R.W., 2017. Orthopaedic surgeon attitudes towards current limitations and the potential for robotic and technological innovation in arthroscopic surgery. Journal of Orthopaedic Surgery 25, 230949901668499. URL: http://www.ncbi.nlm.nih.gov/pubmed/28142353http://journals.sagepub.com/doi/10.1177/2309499016684993, doi:10.1177/2309499016684993.

Kanaan, Y., Jacobson, J.A., Jamadar, D., Housner, J., Caoili, E.M., 2013. Sonographically Guided Patellar Tendon Fenestration: Prognostic Value of Preprocedure Sonographic Findings. Journal of Ultrasound in Medicine 32, 771–777. URL: http://www.ncbi.nlm.nih.gov/pubmed/23620318http://www.jultrasoundmed.org//cgi/doi/10.7863/ultra.32.5.771, doi:10.7863/ultra.32.5.771.

Kendall, A., Badrinarayanan, V., Cipolla, R., 2015. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding URL: http://arxiv.org/abs/1511.02680, arXiv:1511.02680.

Kingma, D.P., Ba, J., 2014. Adam: {A} Method for Stochastic Optimization. CoRR abs/1412.6. URL: http://arxiv.org/abs/1412.6980, arXiv:1412.6980.

Köroğlu, M., Çallıoğlu, M., Eriş, H.N., Kayan, M., Çetin, M., Yener, M., Gürses, C., Erol, B., Türkbey, B., Parlak, A.E., Akhan, O., 2012. Ultrasound guided percutaneous treatment and follow-up of Baker's cyst in knee osteoarthritis. European Journal of Radiology 81, 3466–3471. URL: http://www.ncbi.nlm.nih.gov/pubmed/22726355https://linkinghub.elsevier.com/retrieve/pii/S0720048X12002446, doi:10.1016/j.ejrad.2012.05.015.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks, in: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 25. Curran Associates, Inc., pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networ pdf.

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. doi:10.1038/nature14539.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition, in: Proceedings of the IEEE, pp. 2278–2324.

Léger, J., Brion, E., Javaid, U., Lee, J., De Vleeschouwer, C., Macq,

B., 2018. Contour Propagation in CT Scans with Convolutional Neural Networks, in: Advanced Concepts for Intelligent Vision Systems, pp. 380–391. URL: http://link.springer.com/10.1007/978-3-030-01449-0{_}32, doi:10.1007/978-3-030-01449-0_32.

Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J., 2018a. SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks URL: http://arxiv.org/abs/1812.11703, arXiv:1812.11703.

Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X., 2018b. High Performance Visual Tracking with Siamese Region Proposal Network, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE. pp. 8971–8980. URL: https://ieeexplore.ieee.org/document/8579033/, doi:10.1109/CVPR.2018.00935.

Long, J., Shelhamer, E., Darrell, T., 2014. Fully Convolutional Networks for Semantic Segmentation URL: http://arxiv.org/abs/1411.4038, arXiv:1411.4038.

Lueders, D.R., Smith, J., Sellon, J.L., 2016. Ultrasound-Guided Knee Procedures. Physical Medicine and Rehabilitation Clinics of North America 27, 631–648. URL: http://www.ncbi.nlm.nih.gov/pubmed/27468670https://linkinghub.elsevier.com/retrieve/pii/S1047965116300213, doi:10.1016/j.pmr.2016.04.010.

Maninis, K.K., Caelles, S., Chen, Y., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L., 2018. Video Object Segmentation Without Temporal Information. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) .

Marki, N., Perazzi, F., Wang, O., Sorkine-Hornung, A., 2016. Bilateral Space Video Segmentation, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 743–751 URL: http://ieeexplore.ieee.org/document/7780456/, doi:10.1109/CVPR.2016.87.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), IEEE. pp. 565–571. URL: http://ieeexplore.ieee.org/document/7785132/, doi:10.1109/3DV.2016.79.

Morvan, G., Vuillemin, V., Guerini, H., 2012. Interventional musculoskeletal ultrasonography of the lower limb. Diagnostic and Interventional Imaging 93, 652–664. URL: http://www.ncbi.nlm.nih.gov/pubmed/22921690https://linkinghub.elsevier.com/retrieve/pii/S221156841200280X, doi:10.1016/j.diii.2012.07.007.

Nouri, D., Rothberg, A., 2015. Liver Ultrasound Tracking using a Learned Distance Metric, in: Proceedings of MICCAI workshop: Challenge on Liver Ultrasound Tracking, pp. 5–12.

Oh, S.W., Lee, J.Y., Sunkavalli, K., Kim, S.J., 2018. Fast Video Object Segmentation by Reference-Guided Mask Propagation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE. pp. 7376–7385. URL: https://ieeexplore.ieee.org/document/8578868/, doi:10.1109/CVPR.2018.00770.

Oktay, O., Bai, W., Lee, M., Guerrero, R., Kamnitsas, K., Caballero, J., de Marvao, A., Cook, S., O'Regan, D., Rueckert, D., 2016. Multi-input Cardiac Image Super-Resolution Using Convolutional Neural Networks, pp. 246–254. URL: http://link.springer.com/10.1007/978-3-319-46726-9{_}29, doi:10.1007/978-3-319-46726-9_29.

Oshima, T., Nakase, J., Numata, H., Takata, Y., Tsuchiya, H., 2016. Ultrasonography imaging of the anterolateral ligament using real-time virtual sonography. The Knee 23, 198–202. URL: http://www.ncbi.nlm.nih.gov/pubmed/26810600https://linkinghub.elsevier.com/retrieve/pii/S0968016015002331, doi:10.1016/j.knee.2015.10.002.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library, in: Wallach, H., Larochelle, H., Beygelzimer, A., dAlché Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035.

Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., Sorkine-Hornung, A., 2017. Learning Video Object Segmentation from Static Images, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 3491–3500. URL: http://ieeexplore.ieee.org/document/8099855/, doi:10.1109/CVPR.2017.372.

Pflugfelder, R., 2017. An In-Depth Analysis of Visual Tracking with Siamese Neural Networks URL: http://arxiv.org/abs/1707.00569, arXiv:1707.00569.

Pinheiro, P.O., Collobert, R., 2014. Recurrent Convolutional Neural Networks for Scene Labeling, in: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, JMLR.org. pp. I–82—-I–90. URL: http://dl.acm.org/citation.cfm?id=3044805.3044816.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks URL: http://arxiv.org/abs/1506.01497, arXiv:1506.01497.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Roussos, A., Katsamanis, A., Maragos, P., 2009. Tongue tracking in Ultrasound images with Active Appearance Models, in: 2009 16th IEEE International Conference on Image Processing (ICIP), IEEE. pp. 1733–1736. URL: http://ieeexplore.ieee.org/document/5414520/, doi:10.1109/ICIP.2009.5414520.

Sørensen, T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol. Skr. 5, 1–34.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 15, 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

Tao, R., Gavves, E., Smeulders, A.W.M., 2016. Siamese Instance Search for Tracking, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 1420–1429. URL: http://ieeexplore.ieee.org/document/7780527/, doi:10.1109/CVPR.2016.158.

Tighe, J., Lazebnik, S., 2013. Finding Things: Image Parsing with Regions and Per-Exemplar Detectors, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 3001–3008. URL: http://ieeexplore.ieee.org/document/6619230/, doi:10.1109/CVPR.2013.386.

Tsai, D., Flagg, M., Nakazawa, A., Rehg, J.M., 2012. Motion Coherent Tracking Using Multi-label MRF Optimization. International Journal of Computer Vision 100, 190–202. URL: http://link.springer.com/10.1007/s11263-011-0512-5, doi:10.1007/s11263-011-0512-5.

Tyryshkin, K., Mousavi, P., Beek, M., Ellis, R.E., Pichora, D.R., Abolmaesumi, P., 2007. A navigation system for shoulder arthroscopic surgery. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine 221, 801–812. URL: http://www.ncbi.nlm.nih.gov/pubmed/18019466http://journals.sagepub.com/doi/10.1243/09544119JEIM281, doi:10.1243/09544119JEIM281.

Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., Torr, P.H.S., 2017. End-to-End Representation Learning for Correlation Filter Based Tracking, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 5000–5008. URL: http://ieeexplore.ieee.org/document/8100014/, doi:10.1109/CVPR.2017.531.

Voigtlaender, P., Leibe, B., 2017. Online Adaptation of Convolutional Neural Networks for Video Object Segmentation URL: http://arxiv.org/abs/1706.09364, arXiv:1706.09364.

Wang, Q., Gao, J., Xing, J., Zhang, M., Hu, W., 2017. DCFNet: Discriminant Correlation Filters Network for Visual Tracking URL: http://arxiv.org/abs/1704.04057, arXiv:1704.04057.

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.S., 2018. Fast Online Object Tracking and Segmentation: A Unifying Approach URL: http://arxiv.org/abs/1812.05050, arXiv:1812.05050.

Welch, B.L., 1947. The Generalization Of Student's Problem When Several Different Population Variances Are Involved. Biometrika 34, 28–35. URL: https://doi.org/10.1093/biomet/34.1-2.28, doi:10.1093/biomet/34.1-2.28, arXiv:http://oup.prod.sis.lan/biomet/article-pdf/34/1-2/28/553093

Wong-On, M., Til-Pérez, L., Balius, R., 2015. Evaluation of MRI-US Fusion Technology in Sports-Related Musculoskeletal Injuries. Advances in Therapy 32, 580–594. URL: http://www.ncbi.nlm.nih.gov/pubmed/26093660http://link.springer.com/10.1007/s12325-015-0217-1, doi:10.1007/s12325-015-0217-1.

Wu, L., Jaiprakash, A., Pandey, A.K., Fontanarosa, D., Jonmohamadi, Y., Antico, M., Strydom, M., Razjigaev, A., Sasazawa, F., Roberts, J., Crawford, R., 2018. Robotic and image-guided knee arthroscopy, in: Elsevier's Hand-

book of Robotic and Image-Guided Surgery.
Wu, Y., Lim, J., Yang, M.H., 2013. Online Object Tracking: A Benchmark., in: CVPR, IEEE Computer Society. pp. 2411–2418. URL: http://dblp.uni-trier.de/db/conf/cvpr/cvpr2013.html{#}WuLY13.
Yu, L., Yang, X., Chen, H., Qin, J., Heng, P.A., 2017. Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images. URL: https://dl.acm.org/citation.cfm?id=3298250.

Vitae

Matteo Dunnhofer. Matteo Dunnhofer received the B. Sc. and M. Sc. in computer science from the University of Udine (Udine, Italy) in 2016 and 2018 respectively. Currently he is a PhD student in Industrial and Information Engineering at the same institute. His research is focused on the application of deep learning techniques to different problems in computer vision and medical image analysis.

Maria Antico. Maria Antico received the BEng. in engineering sciences from the University of Rome Tor Vergata, Italy, in 2014 and MEng. in biomechanical engineering from the Technical University of Delft (The Netherlands) in 2016. She is currently a PhD candidate at Queensland University of Technology (Australia). Her research is focused on advanced tissue recognition techniques for fully automated robotic surgery.

Fumio Sasazawa. Dr. Fumio Sasazawa is an orthopaedic surgeon specializing in lower extremities including hip and knee joint. He graduated from University of Tokyo, Faculty of Engineering (Tokyo, Japan) in 1997, and then graduated from Shinshu University School of Medicine (Matsumoto, Japan) to obtain medical license in 2004. He obtained a doctor's degree in cellular and molecular biology at Hokkaido University Graduate School of Medicine in 2014. He worked as a visiting researcher in the medical robotics team of Queensland University of Technology (Brisbane, Australia) in 2017-18.

Yu Takeda. Dr. Yu Takeda is an orthopaedic surgeon. He studied medicine at the Hyogo College of Medicine (Japan) between 2003 and 2009 and was awarded a Ph.D. degree by the Hyogo College of Medicine in 2018. He is currently working as a researcher at the Queensland University of Technology (Australia) in the field of ultrasound-guided autonomous surgery robotic applications.

Saskia Camps. Saskia Camps received both her Bachelor's and Master's degree in Biomedical engineering with a focus on medical image processing from the Eindhoven University of Technology, the Netherlands. Currently, she is finalizing her PhD thesis on ultrasound guidance for radiotherapy of prostate cancer patients in a collaborative project between Philips Research and the MAASTRO Clinic, the Netherlands. In the meantime, she started a new position at EBAMed, a Swiss startup that aims at treating cardiac arrhythmias by means of external beam therapy.

Niki Martinel. Niki Martinel received the M.Sc. (with honors) and the Ph.D. from the University of Udine, Italy in 2010 and 2014, respectively. He is an assistant professor at the Department of Mathematics, Computer Science and Physics at the University of Udine. His research interests include machine learning, wide area scene analysis, deep/hierarchical learning architectures, unsupervised learning.

Christian Micheloni. Christian Micheloni received the M.Sc. and Ph.D. degrees from the University of Udine, Udine, Italy, in 2002 and 2006, respectively. He is Associate Professor with the Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy. His current interests include active vision for the wide area scene analysis, resource aware camera networks, pattern recognition, camera network self reconfiguration, person Re-Identification and machine learning.

Gustavo Carneiro. Gustavo Carneiro received his Ph.D. degree from the University of Toronto, Canada, in 2004. He is a full Professor at the School of Computer Science and the Australian Institute for Machine Learning of the University of Adelaide. His current research interests include machine learning, computer vision and medical image analysis.

Davide Fontanarosa. Dr. Davide Fontanarosa is a physicist with a solid background in ultrasound imaging and medical physics. He worked in one of the top institutions for radiation therapy (MAASTRO Clinic, in the Netherlands) and in one of the largest industrial research laboratories in the world, Philips