EchoFusion: Tracking and Reconstruction of Objects in 4D Freehand Ultrasound Imaging Without External Trackers

Bishesh Khanal[1,2(✉)], Alberto Gomez[1], Nicolas Toussaint[1], Steven McDonagh[2], Veronika Zimmer[1], Emily Skelton[1], Jacqueline Matthew[1], Daniel Grzech[2], Robert Wright[1], Chandni Gupta[1], Benjamin Hou[2], Daniel Rueckert[2], Julia A. Schnabel[1], and Bernhard Kainz[2]

[1] School of Biomedical Engineering and Imaging Sciences,
King's College London, London, UK
bishesh.khanal@kcl.ac.uk

[2] Department of Computing, Imperial College London, London, UK

Abstract. Ultrasound (US) is the most widely used fetal imaging technique. However, US images have limited capture range, and suffer from view dependent artefacts such as acoustic shadows. Compounding of overlapping 3D US acquisitions into a high-resolution volume can extend the field of view and remove image artefacts, which is useful for retrospective analysis including population based studies. However, such volume reconstructions require information about relative transformations between probe positions from which the individual volumes were acquired. In prenatal US scans, the fetus can move independently from the mother, making external trackers such as electromagnetic or optical tracking unable to track the motion between probe position and the moving fetus. We provide a novel methodology for image-based tracking and volume reconstruction by combining recent advances in deep learning and simultaneous localisation and mapping (SLAM). Tracking semantics are established through the use of a Residual 3D U-Net and the output is fed to the SLAM algorithm. As a proof of concept, experiments are conducted on US volumes taken from a whole body fetal phantom, and from the heads of real fetuses. For the fetal head segmentation, we also introduce a novel weak annotation approach to minimise the required manual effort for ground truth annotation. We evaluate our method qualitatively, and quantitatively with respect to tissue discrimination accuracy and tracking robustness.

1 Introduction

Ultrasound (US) is a very widely used medical imaging modality, well known for its portability, low cost, and high temporal resolution. Although the most popular US imaging is 2D B-mode, 3D mode has become an attractive addition providing a larger field of view at an increased frame rate. There is also growing interest in developing low cost 3D US probes [1]. While 2D mode images are

A. Melbourne and R. Licandro et al. (Eds.): DATRA/PIPPI 2018, LNCS 11076, pp. 117–127, 2018.
https://doi.org/10.1007/978-3-030-00807-9_12

usually of higher resolution, 3D mode has the ability to provide better context of the anatomy with smaller number of images. Thus, 3D images could allow easier compounding and field of view extension to capture all the desired anatomy in a single compounded volume.

Volumetric compounding requires the relative transformation between individual volumes. This can be achieved using image registration if the offset is small and assumptions about the spatial arrangement of the volumes hold, e.g., when performing an imaging sweep at constant speed. For large offsets, or random views of a target volume, image registration alone is insufficient and external tracking such as electromagnetic or optical tracking has to be used to establish localisation coherence. External tracking measures absolute transformations between a fiducial marker on the ultrasound probe and a calibrated world coordinate system. Moving targets within a patient cannot be tracked with fiducial markers, computer vision methods that rely on a direct line of sight, or by tracking the probe via external trackers.

An ability to generate high quality compounded volumes of individual fetuses can be useful for retrospective analysis by experts who might not be available, e.g. in rural areas where the live scanning may be performed by non-experts. High quality compounded volumes can also be important in creating US atlases of different fetal organs. For example, it would be desirable to combine all possible views of the brain of single fetus to maximise the information obtained from individual fetal brains. In fetuses of late Gestational ages (GAs), acquiring images from all possible directions requires probe manipulation, incurring large rotation and translational motion. Registration and tracking of images resulting from such constraint-free probe motions is typically highly challenging. A motion-robust and hardware-lean image-based method to compound a large anatomical RoI in real-time is thus highly desired.

Contribution: We propose a novel approach to tackle the tracking problem during 3D fetal US examinations where an application-focused tissue discriminator, based on convolutional neural networks, is integrated into a simultaneous localisation and mapping (SLAM) formulation named EchoFusion. The proposed method yields relative transformations between subsequent volumes, surface reconstruction of the target anatomy, and reconstruction of a compounded volume at the same time. We demonstrate the potential of the proposed approach with experiments for rigid whole body fetal phantom, and for free-hand 4D US covering the head region in real fetuses, without external tracking or a highly restrictive scanning protocol. EchoFusion requires the fetal tissue discriminator to be accurate only in the fetal surface closest to the US probe, allowing the use of: (i) challenging 4D fetal screening US images coming from a very wide range of views, and (ii) weak annotations, enabling large training data at low cost.

Related Work: Extending the FOV by compounding multiple 3D images has been in focus since a wide range of freehand ultrasound probes support 3D images with either matrix array transducers [19] or mechanically steered linear arrays in plane fan mode [5]. Tracking-based methods [3,15] provide good initialisation for a variety of subsequent and task-specific registration methods but often need additional calibration to establish the transformation between object and tracking coordinate system [2]. For rigid non-moving targets, advanced registration strategies can yield good compounding results, given that the acquisition protocol is well defined. For example, [16] uses defined sweeps and multivariate similarity measures in a maximum likelihood framework to mitigate the problem of registration drift observed in earlier, pair-wise registration methods [6]. However, algorithms requiring all the available images simultaneously to estimate transformations cannot be used in real-time applications such as a visual guidance system for non-expert sonographers to receive feedback, during scanning, of the regions already captured.

Recent advances in the robustness of semantic discrimination of tissues in medical images largely enabled by the advent of deep learning, and in SLAM algorithms, provide potential to combine these processes in a reliable fashion. SLAM is known from natural image processing as a powerful tool for indoor [17] and outdoor [8] mapping, location awareness of robots [4] and real-time 3D mesh reconstruction from a stream of RGB images that additionally provide depth information [12]. These techniques have been applied in the medical image analysis community to laparoscopy [19] and movement-based diagnosis [10], but never went beyond RGB (+depth) imaging.

Traditional SLAM methods assume a clear line of sight to map the depth of a scene. However, US images require preprocessing such as segmentation to extract depth of the desired target objects. Convolutional neural networks constitute the state of the art for solving (medical) image segmentation tasks e.g. [9] and have recently shown to be robust for the use in, e.g., fetal screening examinations [20], however only at very young GA when the fetus is fully visible in 3D US volumes. Our work combines fast automatic tissue segmentation that works also on partially visible tissue in later gestation with modern SLAM algorithms. To the best of our knowledge, this is the first time such an approach is proposed.

2 Method

Our approach consists of three main components: **(1)** semantic tissue segmentation, **(2)** transducer to object depth map generation, and **(3)** simultaneous localisation and mapping algorithm. An overview of our approach is shown in Fig. 1.

(1) Semantic tissue discrimination: The objective is to produce a binary segmentation of the target object. For example, for fetal head tracking and reconstruction, the foreground is the fetal head and the remaining structures such as fetal limbs and maternal tissues are background. Fetal segmentation from freehand 4D US can be quite challenging because of the diversity in the image appearance of the same anatomy, cropping due to limited field of view, and the relatively low quality of 4D images compared to 2D images or static 3D volumes. As the images are often corrupted by shadows, fetal body surface at distances far from the transducer cannot be delineated as accurately as surfaces physically nearer to the probe. Thus, in the present work, expert sonographers

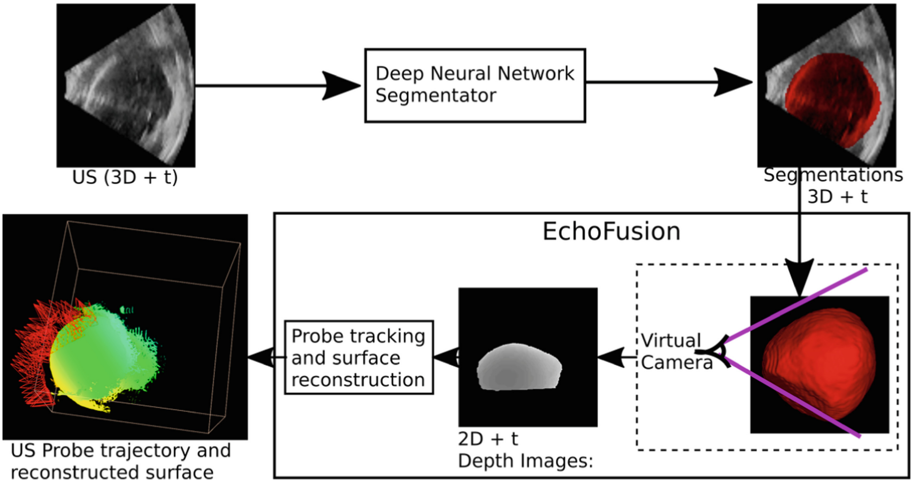

US (3D + t)
Deep Neural Network Segmentator
Segmentations 3D + t
EchoFusion
Virtual Camera
2D + t Depth Images:
Probe tracking and surface reconstruction
US Probe trajectory and reconstructed surface

Fig. 1. Overview: Residual 3D U-Net segments each incoming 3D US from which target fetal organ's surface depth is extracted by a virtual camera located at the ultrasound probe. EchoFusion estimates the camera transformation w.r.t previous frame using the incoming depth image and updates the dense surface model.

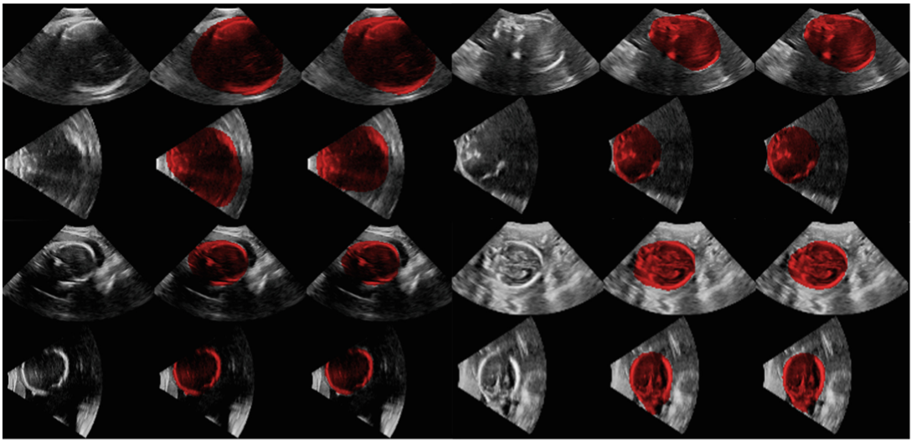


Fig. 2. Four US volumes with input, GT, and predicted volumes (left to right) of two central orthogonal slices. This shows the typical diversity of the input sizes, view direction, partial head views, shadows and US artifacts in the dataset used in our experiments.

delineated the closest surface accurately but approximated the shape of the RoI in the surface further from the probe as shown in Fig. 2.

For semantic segmentation, we use a Residual 3D U-Net architecture which has U-Net structure [13] and is similar to V-net [11] with all convolution layers being replaced by residual-units [7] known to make training more stable. We follow the common strategy whereby skip connections are implemented via concatenation in the up-sampling component of the network and down-sampling is performed with strided convolution (*cf.* max pooling). Each convolutional layer of the original architecture is additionally augmented by a residual block containing two convolutions in a similar fashion to [9]. We employ $[16, 32, 64, 128]$ feature maps per layer and all kernels and feature maps are 3D. Each layer additionally utilizes batch normalization, ReLUs and zero-padding.

For training we draw input training patches of size $64 \times 64 \times 64$ voxels with an equal probability of patches being centered around a voxel from the foreground or background label class. We train to minimize a standard cross-entropy loss using Adam optimization with learning rate of 0.001 and l_2 regularization. Our training imagery is augmented via Gaussian additive noise ($\sigma = 0.02$) with image flipping in each axis.

(2) Transducer distance field generation: Depth images can be generated using a virtual pinhole camera that looks into the 3D segmented model from the same direction as the US probe. All voxels in the output segmentation have known physical co-ordinates with respect to an arbitrary reference point, set as the origin of the world co-ordinate. In the input image volumes, the origin was set to a central point in xz-plane at y = 0 making the US probe directed towards positive y-direction and placed $y < 0$. We set a virtual camera that looks towards positive y-axis and along the line $x = z = 0$. The exact position and the view angle of the camera depends on the sector width and sector height of the input 3D US volume. If the camera is too far away, it sees the flat surface at the edge of the US sector. Similarly, if the camera is too close, the FoV is not wide enough and some parts of the tissue region may be missed. In order to estimate an optimal camera position, first we separately compute the intersection and angle between sector lines for the central slices in yz-plane and the central slices in xy-plane as follows:

1. Extract sector mask using thresholding, morphological closing to remove holes.
2. Extract edges using Canny edge detection on the sector mask.
3. Use Hough transform to detect the two sector lines.
4. Compute intersection and angle between the lines found in 3.

Then, the camera distance is set to be the minimum of the two intersection points, and the view angle is chosen to be the wider of the two angles.

(3) Tracking and Reconstruction with EchoFusion: In SLAM [12], a sequence of partial views of a 3D scene captured as 2D RGB images and/or depth images is used to estimate all the relative poses of the camera and reconstruct the 3D scene. Like all SLAM algorithms, we also use only the frontal surface of the 3D scene that are not occluded from the camera view to track and build the 3D scene incrementally. Thus, we use a volumetric surface representation to store global 3D scene as a truncated signed distance function (TSDF) [12] in a predetermined 3D voxel grid. This 3D model is updated with each new incoming depth image by estimating the camera transformation with respect to the previous frame. The algorithm can be outlined as follows:

1. From the generated depth image compute the 3D vertex and normals in camera co-ordinate space.
2. The 3D vertex and normals from the previous frame are estimated by ray casting the 3D model built so far from the global camera position estimated from the previous frame.
3. The relative camera transformation is then estimated using Iterative Closest Point (ICP) of the two point sets from the current and the previous frames.

The 3D model gets better and smoother as more consistent data becomes available.

Implementation Details: We adapted an open source implementation[1] of Kinect Fusion [12]. The focal length of the virtual camera can be computed as $f = \frac{w/2}{\tan(\alpha/2)}$, where α is the view angle and w is the image width in pixel co-ordinates. We set depth and RGB image sizes to 480 × 480. The discriminator model is trained on a Nvidia Titan X GPU with 12 GB of memory. During runtime, the same GPU can be used for inference and EchoFusion, as the inference from the network does not require large resources like in training time. The network was implemented in tensorflow.

3 Experiments and Results

Phantom Data: We use data from a fetal phantom Kyotokagaku UTU-1 at a gestational age of about 20 weeks. The GT segmentation consists of fetal vs. maternal tissue delineation in 28 3D volumes which is randomly split into 24 training samples and 4 validation samples. The GT segmentations include both the fetal head and body as foreground.

Fetal Screening Data: Two expert sonographers delineated 192 US fetal head volumes for training and validation of fetal head segmentation. These 3D images were selected from 4D freehand scanning of 19 different fetuses having GAs in the range of 23–34 weeks with mean (std) age of 30 (2.842) weeks.

The sonographers used MITK [18] to segment six to seven representative slices manually, then performed 3D interpolation from these slices to create a 3D shape. Many of these images contained shadows on the far-field surface, so the manual delineation was done empirically based on the sonographers' anatomical knowledge of the head shape. We split 192 GT data into 184 training and 8 validation images. We then test the trained network only once on a set containing GT segmentations from five fetuses not used in training-validation set.

Evaluation: We use Dice score to evaluate the performance of segmentation quantitatively. Evaluating tracking accuracy is challenging without a ground truth. Surface reconstruction which can be qualitatively observed depends on

[1] https://github.com/Nerei/kinfu_remake.

the tracking obtained from the SLAM. To assess the tracking robustness on freehand 4D US stream of the real fetal heads, we test our framework on 37 fetuses and compute the number of tracking losses (i.e. reset of the tracked pose) and the longest sequence without any resets.

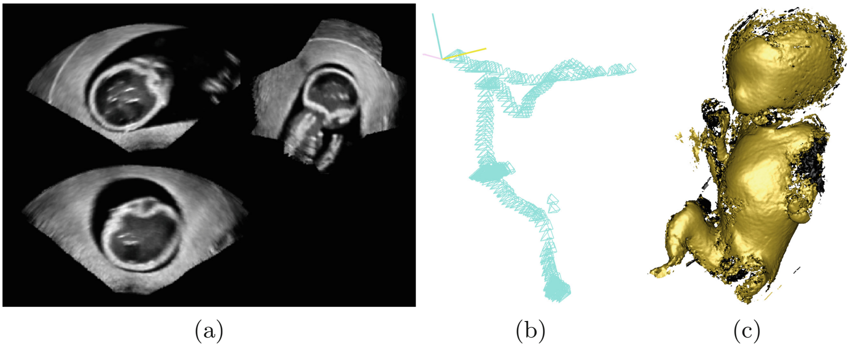


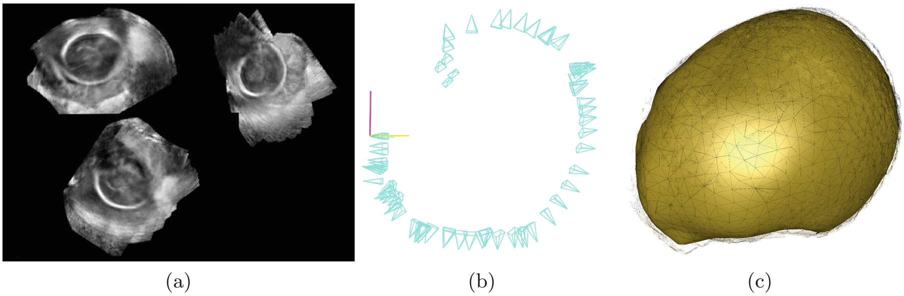
Fig. 3. Orthogonal slices through examples of compounded volumes (a), EchoFusion tracking trajectory (b) and TSDF iso-surface reconstruction (c) for sequences from whole body fetus phantom. The sequence of images of the static phantom were taken with a very wide range of probe directions as seen in the top right slice in (a), and from the trajectory in (b). Limbs are not reconstructed faithfully due to limb information being purposefully discarded at segmentation time.

Results: Table 1 shows quantitative results for segmentation performance on both the phantom and the real fetuses. Since there was only one phantom available which was used to create training and validation set, there is no test set for the phantom. For the real fetuses, test set was created using the same protocol as the training sets but from the fetuses that were not used for training or validation. Although the number of images used for training on the phantom is much smaller than for the real fetuses, the validation set accuracy is higher for the phantom. This is not surprising because the images from the phantom are much less challenging than the real fetuses.

Table 1. Dice scores for real-time semantic tissue discrimination.

Set	images(real)	mean(std)	images(Phantom)	mean(std)
Train	178	0.9408(0.0389)	24	0.9735(0.0125)
Validation	8	0.9217(0.0212)	4	0.9267(0.0074)
Test	26	0.8942(0.0671)	-	-

Figures 3 and 4 show qualitative results after compounding a series of 10–20 EchoFusion-tracked consecutive 3D volume acquisitions from different locations. 3D surface reconstruction in Fig. 3 shows that both the phantom face and body


(a)

(b)

(c)

Fig. 4. Orthogonal slices through examples of compounded volumes (a), EchoFusion tracking trajectory (b) and TSDF iso-surface reconstruction (c) for sequences of a real fetal head. Note that the tracking is relative only to the fetal head, and not the other moving maternal and fetal tissues.

which were selected as foreground objects for the segmentation are nicely reconstructed. Similarly, the fetal head in compounded volume in Fig. 4 shows that the sequence of images registered reasonably well although they were taken from a wide range of angles. Table 2 shows EchoFusion tracking performance on 37 fetal sequences of volumes. On average, there were approximately 98 total frames for which the SLAM algorithm lost tracking approximately 5 times. These sequences were obtained by moving the probe in different directions trying to cover the head (skull and face) from all possible directions. The sequences were used as they were acquired without data cleaning, thus containing views which do not show the fetal head and many frames with only partial views of the head region.

Table 2. Robustness with respect to continuously tracked frames for 37 fetuses.

	mean(std)	median	range
Total frames	98.11(54.65)	91	[21, 277]
No. of tracking losses	5.16(3.67)	5	[0, 15]
Longest sequence without tracking loss	40.86(30.85)	31	[4, 152]

4 Discussion

The key contribution of this work is the novel approach to the tracking and compounding problem in freehand 4D US, which constitutes combining the powerful semantic segmentation neural networks with modern SLAM algorithms. Since both of them are very active fields of research, there is a lot of potential to improve EchoFusion for a multitude of applications including compounding, image reconstruction, artefact reduction, super resolution and fetal face biometrics using the resulting dense surface model. Moreover, this method could also allow non-expert to acquire dense data for retrospective evaluation.

The goal of this work was to provide a proof of concept, but clinical translation of this method would require a more extensive quantitative validation of the tracking accuracy, drift over the long sequence, and compare how segmentation accuracy impacts the overall tracking accuracy.

The use of whole body phantom vs fetal head also demonstrates that the top level approach generalises across organs and anatomy as we can train the segmentation network for a desired RoI. However, the current implementation of the SLAM algorithm works only for largely rigid body motion; the static phantom and the fetal head can be reasonably assumed to have mostly rigid body movement with respect to the probe at the semantic level. For non-rigid movements of the fetus such as the whole body or abdomen, the current SLAM component must be replaced with the methods that take dynamic scene changes into account [14]. However, such approaches would still not be robust to sudden movements (*e.g.* kicks) and introduce a significant computational overhead, potentially jeopardizing hard real-time constraints. One approach to tackle this problem is to consider such suddenly moving limbs as background in segmentation so that they are ignored during the tracking and reconstruction. There can still be challenges, (*e.g.* turning the head in the opposite direction and staying there, when reconstructing head/shoulders/torso at once), and is more of an open problem at present. However, being able to focus and compound on quasi-rigid areas like only the head or only abdomen and changing the model depending on target application would already be very valuable e.g., for the creation of fetal brain or abdomen atlases.

5 Conclusion

We have developed a novel approach demonstrating a promising potential for robust segmentation and tracking of fetuses in utero. EchoFusion is versatile and could be applied in any situation where an independently moving target object is occluded by other tissue or material. We have also introduced a way to learn a tissue discriminator from weak annotations in fetal 3D US images and discussed the performance of a Residual 3D U-Net tissue discriminator learning from this data. This discriminator is key to establishing semantics for SLAM-based tracking, which we evaluated on 4D freehand US of a fetal phantom and on real fetuses from screening examinations. In the future, we will perform a more extensive validation of the tracking accuracy, and also find a way to derive robust SDFs from tissue probabilities to exploit the possibilities of dynamic fusion approaches.

Acknowledgements. This work was supported by the Wellcome/EPSRC Centre for Medical Engineering [WT 203148/Z/16/Z], Wellcome Trust IEH Award [102431]. The authors thank Nvidia Corporation for the donation of a Titan Xp GPU.

References

1. Angiolini, F., et al.: 1024-Channel 3D ultrasound digital beamformer in a single 5W FPGA. In: Proceedings of the Conference on Design, Automation & Test in Europe, pp. 1225–1228. European Design and Automation Association (2017)
2. Blackall, J.M., Rueckert, D., Maurer, C.R., Penney, G.P., Hill, D.L.G., Hawkes, D.J.: An image registration approach to automated calibration for freehand 3D ultrasound. In: Delp, S.L., DiGoia, A.M., Jaramaz, B. (eds.) MICCAI 2000. LNCS, vol. 1935, pp. 462–471. Springer, Heidelberg (2000). https://doi.org/10.1007/978-3-540-40899-4_47
3. Octorina Dewi, D.E., Mohd. Fadzil, M., Mohd. Faudzi, A.A., Supriyanto, E., Lai, K.W.: Position tracking systems for ultrasound imaging: a survey. In: Lai, K.W., Octorina Dewi, D.E. (eds.) Medical Imaging Technology. LNB, pp. 57–89. Springer, Singapore (2015). https://doi.org/10.1007/978-981-287-540-2_3
4. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping: part I. IEEE Robot. Autom. Mag. **13**(2), 99–110 (2006)
5. Fenster, A., Downey, D.B.: 3-D ultrasound imaging: a review. IEEE Eng. Med. Biol. Mag. **15**(6), 41–51 (1996)
6. Gee, A.H., et al.: Rapid registration for wide field of view freehand three-dimensional ultrasound. IEEE Trans. Med. Imaging **22**(11), 1344–1357 (2003)
7. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
8. Heng, L., et al.: Self-calibration and visual SLAM with a multi-camera system on a micro aerial vehicle. Auton. Robot. **39**(3), 259–277 (2015)
9. Kamnitsas, K., et al.: Ensembles of multiple models and architectures for robust brain tumour segmentation. CoRR abs/1711.01468 (2017)
10. Kontschieder, P., et al.: Quantifying progression of multiple sclerosis via classification of depth videos. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8674, pp. 429–437. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10470-6_54
11. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
12. Newcombe, R.A., et al.: KinectFusion: real-time dense surface mapping and tracking. In: ISMAR, pp. 127–136. IEEE (2011)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Slavcheva, M., Baust, M., Ilic, S.: SobolevFusion: 3D reconstruction of scenes undergoing free non-rigid motion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2646–2655 (2018)
15. Solberg, O.V., et al.: Freehand 3D ultrasound reconstruction algorithmsa review. Ultrasound Med. Biol. **33**(7), 991–1009 (2007)
16. Wachinger, C., Wein, W., Navab, N.: Three-dimensional ultrasound mosaicing. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007. LNCS, vol. 4792, pp. 327–335. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-75759-7_40

17. Whelan, T., et al.: ElasticFusion: dense SLAM without a pose graph. In: Robotics: Science and Systems (2015)
18. Wolf, I., et al.: The Medical Imaging Interaction Toolkit (MITK) a toolkit facilitating the creation of interactive software by extending VTK and ITK
19. Wygant, I.O., et al.: Integration of 2D CMUT arrays with front-end electronics for volumetric ultrasound imaging. IEEE Trans. Ultrason. Ferroelect. Freq. Control **55**(2), 327–342 (2008)
20. Yang, X., et al.: Towards automatic semantic segmentation in volumetric ultrasound. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 711–719. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_81