Deep Sequential Mosaicking of Fetoscopic Videos — **Source link** ⧉

Sophia Bano, Francisco de Assis Guedes de Vasconcelos, Marcel Tella Amo, George Dwyer ...+6 more authors

Institutions: University College London, Katholieke Universiteit Leuven, King's College London

Related papers:

- Probabilistic visual and electromagnetic data fusion for robust drift-free sequential mosaicking: application to fetoscopy.

- Technical note: probabilistic visual and electromagnetic data fusion for robust drift-free sequential mosaicking: application to fetoscopy

- Pruning strategies for efficient online globally consistent mosaicking in fetoscopy.

- Multi-frame Feature Aggregation for Real-time Instrument Segmentation in Endoscopic Video

- Spatio-temporal Data Augmentation for Visual Surveillance.

Deep Sequential Mosaicking of Fetoscopic Videos

Sophia Bano[1]([✉]), Francisco Vasconcelos[1], Marcel Tella Amo[1], George Dwyer[1], Caspar Gruijthuijsen[2], Jan Deprest[4], Sebastien Ourselin[3], Emmanuel Vander Poorten[2], Tom Vercauteren[3], and Danail Stoyanov[1]

[1] Wellcome/EPSRC Centre for Interventional and Surgical Sciences(WEISS) and Department of Computer Science, University College London, London, UK
sophia.bano@ucl.ac.uk
[2] Department of Mechanical Engineering, KU Leuven University, Leuven, Belgium
[3] School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK
[4] Department of Development and Regeneration, University Hospital Leuven, Leuven, Belgium

Abstract. Twin-to-twin transfusion syndrome treatment requires fetoscopic laser photocoagulation of placental vascular anastomoses to regulate blood flow to both fetuses. Limited field-of-view (FoV) and low visual quality during fetoscopy make it challenging to identify all vascular connections. Mosaicking can align multiple overlapping images to generate an image with increased FoV, however, existing techniques apply poorly to fetoscopy due to the low visual quality, texture paucity, and hence fail in longer sequences due to the drift accumulated over time. Deep learning techniques can facilitate in overcoming these challenges. Therefore, we present a new generalized Deep Sequential Mosaicking (DSM) framework for fetoscopic videos captured from different settings such as simulation, phantom, and real environments. DSM extends an existing deep image-based homography model to sequential data by proposing controlled data augmentation and outlier rejection methods. Unlike existing methods, DSM can handle visual variations due to specular highlights and reflection across adjacent frames, hence reducing the accumulated drift. We perform experimental validation and comparison using 5 diverse fetoscopic videos to demonstrate the robustness of our framework.

Keywords: Sequential mosaicking · Deep learning · Surgical vision · Twin-to-twin transfusion syndrome (TTTS) · fetoscopy

1   Introduction

Twin-to-twin transfusion syndrome (TTTS) can occur during identical twin pregnancies where abnormal vascular anastomoses in the monochorionic placenta result in uneven blood flow between the fetuses [1]. Fetoscopic laser photocoagulation is the most effective treatment for regulating the blood flow. During treatment, the clinician first visually explores the placenta using fetoscopic video to identify vascular anastomoses, building a mental map and treatment plan. Limited FoV, poor visibility and limited maneuverability of the fetoscope introduce

challenges that increase procedural time, can lead to complications and impede verifying completion [14]. Mosaicking can align multiple overlapping images to generate an image with increased FoV. Hence it can provide computer-assisted intervention support to ease the localization of the vascular anastomoses sites.

Mosaicking has recently gained attention to increase the FoV in fetoscopy [10,3,11,12,9]. Totz et al. [13] presented a dynamic view expansion and surface reconstruction approach for minimally invasive surgery by analyzing stereo laparoscopy videos. Reeff at al. [10] and Daga et al. [3] utilized a classical image feature-based matching method for creating mosaics from planar placenta images. The relative transformations between pairs of consecutive fetoscopic images are computed and combined in a chain with respect to a reference frame to generate the mosaic. Error in the relative transformations can propagate to introduce large drift in the overall mosaic, where an electromagnetic tracker (EMT) can be integrated within the fetoscope to minimize any drifting errors [11]. However, integrating an EMT sensor with a fetoscope in-vivo is still an open challenge due to limited form-factor of the fetoscope and due to regulation. To this end, an existing registration technique avoided explicit feature correspondence by utilizing pixel-wise alignment of gradient orientations for a single in-vivo fetoscopic video [9]. Fetoscopic videos are captured from monocular cameras and pose challenges for mosaicking due to varying visual quality due to various types of fetoscopes, occlusions, specular highlights, lack of visual texture, poor visibility due to turbid amniotic fluid and non-planar views [6].

Recently, deep image homography estimation methods have been proposed [4,8] that estimate the homography between pairs of image patches extracted from an image. We observe that a full mosaic is generated by computing sequential homographies between adjacent frames, where a fetoscopic video poses challenges such as specular reflections, amniotic fluid particles, and occlusions. This affects the stitching problem, however, such challenges can be tackled when the homography is estimated using multiple pairs of image patches extracted at random from adjacent frames. In this paper, we employ this approach and propose the first generalized Deep Sequential Mosaicking (DSM) framework for creating mosaics with minimum drift from long-range fetoscopic videos captured from various fetoscopes. We adopt the deep image-based homography estimation method [4] to incorporate sequential data by proposing the Controlled Data Augmentation (CDA) and outlier rejection methods. CDA assumes that the transformation between two adjacent frames contains rotation and translation only, and uses a small set of fetoscopic images of varying quality and appearance, for training. To eliminate the error due to varying visual quality and texture paucity between adjacent frames, we propose the outlier rejection method. This increases the robustness by pruning patch-based homography estimates between adjacent frames. CDA along with the outlier rejection minimize the drift without the use of any external sensors and generate reliable mosaics in this challenging application. Comparison with existing methods and validation on 5 datasets verifies the promising generalization capabilities of our method.
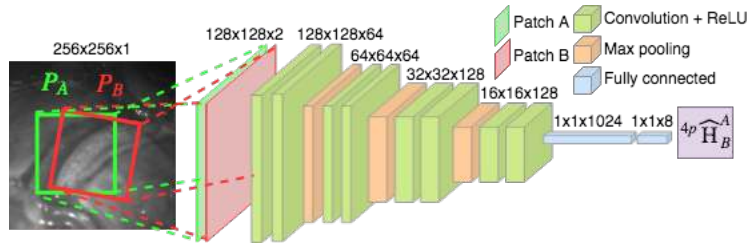
Fig. 1: Deep image homography network with controlled data augmentation.

2   Homography Estimation with Deep Learning

The Deep Image Homography (DIH) model [4] estimates the relative homography between pairs of image patches extracted from a single image. This model uses the 4-point homography parameterization $^{4p}\mathbf{H}$, instead of the 3×3 parameterization $\mathbf{H}$, as the rotation and shear components in $\mathbf{H}$ have smaller magnitude compared to the translation, thus have a small effect on the training loss. Let (u_i, v_i) and (u_i', v_i') denote the four corners of image patch P_A and P_B. Then the 4-point homography $^{4p}\mathbf{H}$ is given by:

$$^{4p}\mathbf{H} = \begin{bmatrix} \Delta u_1 & \Delta u_2 & \Delta u_3 & \Delta u_4 \\ \Delta v_1 & \Delta v_2 & \Delta v_3 & \Delta v_4 \end{bmatrix}^T, \quad \text{where} \begin{array}{l} \Delta u_i = u_i' - u_i, \ \Delta v_i = v_i' - v_i \\ \text{and } i = 1, 2, 3, 4 \end{array} \quad (1)$$

DIH [4] uses a VGG-like architecture, with 8 convolutional and 2 fully connected layers (Fig. 1). The input of the network is P_A and P_B, and output is their relative homography. Note that [4] used the MS-COCO dataset for training, where pair of patches were extracted from a single real image, free of artifacts (e.g. specular highlights, amniotic fluid particles) that appear in sequential data.

DIH [4] generated the training data by randomly selecting P_A from a grayscale image and randomly perturbing its corners to obtain P_B and the Ground-Truth (GT). We observe through experimentation that such data augmentation results in scenarios that are challenging for the network to learn, hence results in a large error (Fig. 3(d) and Fig. 4). While such errors are acceptable in image-based homography [4], for mosaicking even a small error in pairwise homography accumulates over time resulting in increased drift. Therefore, this data generation approach cannot be used as it is for sequential mosaicking.

3   Deep Sequential Mosaicking (DSM)

Mosaic from an image sequence can be generated by finding the pairwise homographies between adjacent frames, followed by computing the relative homographies with respect to a reference frame. The GT pairwise homographies are unknown in fetoscopic videos since they are captured from a monocular camera. Therefore, only through visualization, we can observe the error accumulated over
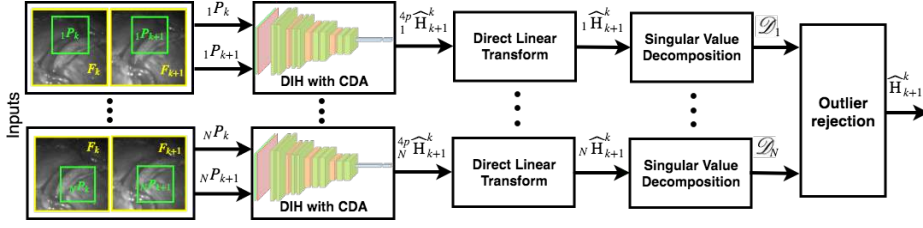
Fig. 2: Overview of the proposed Deep Sequential Mosaicking(DSM) method that uses Controlled Data Augmentation (CDA) for training Deep Image Homography (DIH) model and outlier rejection for pruning the homography estimates.

time. For minimizing this error, in our proposed DSM, the relative homography is learned between patches that are extracted from a single image following the CDA (Sec. 3.1). Unlike [4], in practice homography is computed between two adjacent frames, having specular highlights and lack of texture, in fetoscopic videos. Therefore, testing by using pairs of patches from two adjacent frames results in varying $\mathbf{H}$. To overcome this error, we propose an outlier rejection step (Sec. 3.2) to improve the estimation. During testing (Fig. 2), we compute homographies between pairs of adjacent frames N times by randomly selecting the location of P_A. The estimated $^{4p}\mathbf{H}$ is converted to $\mathbf{H}$ by applying Direct Linear Transform (DLT), followed by its decomposition using Singular Value Decomposition (SVD) and outlier rejection for removing inaccurate estimations.

3.1 Controlled Data Augmentation (CDA)

Pairwise homography between two consecutive frames F_k and F_{k+1} are related by affine transformations including rotation, translation, scale, and shear. A TTTS procedure is performed at a fixed distance from the placenta, hence the scale remains constant. Fetoscope motion is physically constrained by the incision point (remote center of motion), which makes shear very small in consecutive frames, compared to rotation and translation. Therefore, we neglect the scale and shear components and assume that F_k and F_{k+1} are related by translation and rotation only. This helps to minimize the error in relative homography and consequently reduce the drift in mosaicking. For CDA, given a grayscale image I, an image patch P_A is extracted at a random location with corner points (u_i, v_i), where $i = 1, 2, 3, 4$. Rotation by β and translation by (d_x, d_y) is applied:

$$\begin{bmatrix} u_i' \\ v_i' \end{bmatrix} = \begin{bmatrix} cos\beta & sin\beta \\ -sin\beta & cos\beta \end{bmatrix} \begin{bmatrix} u_i \\ v_i \end{bmatrix} + \begin{bmatrix} d_x \\ d_y \end{bmatrix}, \tag{2}$$

to obtain P_B, where β, d_x and d_y are empirically selected. During training, the relative homography is learned between patches that are extracted from a single image following the CDA. Due to lack of texture and poor contrast in fetoscopic videos, homography between two consecutive frames may not be accurate.
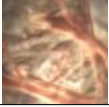
Representative frame					
Data type	Synthetic (SYN)	Ex-vivo in water (EX)	Phantom without fetus (PHN1)	TTTS Phantom in water (PHN2)	Invivo TTTS procedure (INVI)
Imaging source	-	Stereo	Rigid 30° scope	Rigid scope	Rigid scope
No. of frames	811	404	681	400	200
Resolution (pixels)	385×385	250×250	1280×960	720×720	470×470
Crop resolution (pixels)	260×260	250×250	834×834	442×442	312×312
Camera view	Planar	Planar	Non-planar	Non-planar heavy occlusions	Non-planar heavy occlusions
Motion type	Circular	Spiral	Circular freehand	Exploratory freehand	Exploratory freehand

Table 1: Main characteristics of the datasets used for the experimental analysis.

3.2 Homography Matrix Decomposition and Outlier Rejection

To obtain the most consistent homography matrix, we first decompose the homography matrix by applying SVD [7]:

$$\begin{bmatrix} \widehat{h}_{11} & \widehat{h}_{12} \\ \widehat{h}_{21} & \widehat{h}_{22} \end{bmatrix} = \begin{bmatrix} cos\widehat{\theta} & sin\widehat{\theta} \\ -sin\widehat{\theta} & cos\widehat{\theta} \end{bmatrix} \begin{bmatrix} \widehat{s}_g & 0 \\ 0 & \widehat{s}_h \end{bmatrix} \begin{bmatrix} cos\widehat{\gamma} & sin\widehat{\gamma} \\ -sin\widehat{\gamma} & cos\widehat{\gamma} \end{bmatrix}, \tag{3}$$

and $\widehat{t}_x = \widehat{h}_{13}$, $\widehat{t}_y = \widehat{h}_{23}$ are the translation components. By solving eq. 3, we obtain the decomposed parameters, $\mathscr{D} = \{\widehat{\theta}, \widehat{\gamma}, \widehat{s}_g, \widehat{s}_h\}$ [7]. Next, for F_k and F_{k+1}, we compute $_n\widehat{\mathbf{H}}^k_{k+1}$ for $N = 99$ iterations by selecting a new random patch pair $_nP_k$ and $_nP_{k+1}$ at each iteration and obtain N decompose parameters, represented for example as $(\widehat{\theta}_n)^N_{n=1}$. The variations in $(\widehat{s}_{gn})^N_{n=1}$ and $(\widehat{s}_{hn})^N_{n=1}$ are very small due to fixed scale assumption, but are significant in $(\widehat{\theta}_n)^N_{n=1}$ and $(\widehat{\gamma}_n)^N_{n=1}$. Since the first and third matrices in eq. 3 are orthogonal, $\widehat{\theta}_n = -\widehat{\gamma}_n$, filtering either of the two has the same effect. We apply median filtering, since it is useful for mitigating the effect of the outliers, to $(\widehat{\theta}_n)^N_{n=1}$ to get its argument i, giving the most consistent value for θ. This argument is used to obtain $\widehat{\gamma}_i$, $\widehat{s}_{xi}$, $\widehat{s}_{yi}$, $\widehat{t}_{xi}$ and $\widehat{t}_{yi}$, that are then plugged into eq. 3 to get the consistent $_i\widehat{\mathbf{H}}^k_{k+1}$.

4 Experimental Setup and Evaluation Protocol

For experimental analysis, we use 5 fetoscopic videos (Table. 1), which include a synthetic video (SYN) - a discontinuous version of this sequence was used in [11], an ex-vivo in water (EX) data reported in [5], a placenta phantom (PHN1), a TTTS phantom[5] in water (PHN2) depicting an in-vivo procedure and an in-vivo TTTS procedure (INVI). Note from Table. 1 the variability in visual quality, appearance, resolution, imaging source, camera views and captured motion. These variations pose challenging scenarios for mosaicking methods.

---

[5] TTTS phantom from Surgical Touch Simulator: `https://www.surgicaltouch.com/`

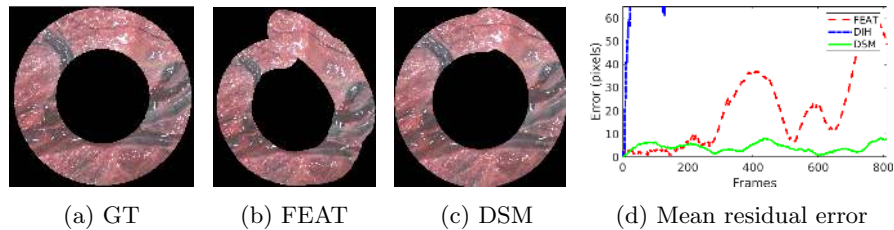(a) GT   (b) FEAT   (c) DSM   (d) Mean residual error

Fig. 3: (a-c) Visualization of mosaics for one circular loop (360 frames) of the SYN sequence. (d) Quantitative comparison of FEAT, DIH and DSM.

For training, we use 600 frames extracted at random from SYN, PHN1, PHN2, INVI and another ex-vivo still images dataset (not used in testing as it is not a video sequence). EX (Table 1) is not use during training, hence it is an unseen data for testing. We extract square frames, from the circular FoV of fetoscopic videos, to be used as the input to DSM. All images are converted to grayscale and resized to 256×256 pixels. We use Keras with Tensorflow backend for the implementation and train our network for about 15 hours on a Tesla V100 (32GB) using learning rate of 10^{-4} and ADAM optimizer. DIH with CDA is trained for 60,000 epochs with a batch size of 32. In each epoch, pairs of patches are generated by randomly selecting β between $(-5, +5)$ *degrees*, and d_x and d_y between $(-16, 16)$. Same training settings are used for DIH without CDA where each corner point of P_A is perturbed at random between $(-16, 16)$.

We perform comparison of DSM with a feature-based (FEAT) [2] and DIH [4] methods. FEAT extract SURF features from a pair of images and performs an exhaustive search for feature matching to estimate the homography. We report the mean residual error (as detailed in [11]) between the GT and estimated relative homographies for SYN (the only sequence with known GT homographies). For quantitative evaluation, we report the average Root Mean Square Error (RMSE) between pair of image patches with known GT homograpies obtained from data augmentation, and average pixel-wise photometric error computed by taking the L1-distance between frame F_{k+1} and reprojected F_k using the estimated homography. We also report qualitative results through visualization.

5   Results and Discussion

The visualization and comparison results on one circular loop (360 frames) of the SYN sequence are shown in Fig. 3(a)-(c). Note the small drift in DSM compared to FEAT. Similar behavior is observed from the mean residual error in Fig. 3(d) where the errors are reported for FEAT, DIH and DSM for the complete length of the sequence (811 frames). It can be seen that the error for FEAT starts increasing after approximately 300 frames and the mosaic starts drifting away. DIH error explodes within a few frames due to the random perturbation during training (Sec. 2). On the other hand, the error for DSM is very small

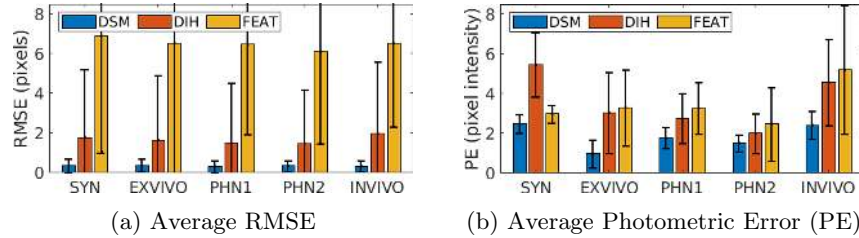(a) Average RMSE   (b) Average Photometric Error (PE)

Fig. 4: Quantitative evaluation and comparison on five diverse fetoscopic videos.

and remains bounded. This is further verified from the low RMSE (0.36) and photometric (2.48) errors for DSM (Fig. 4). Comparison of our proposed DSM with FEAT and DIH is presented in Fig. 4. Overall the pairwise homography errors are high for FEAT for all five sequences due to poor visual quality and lack of texture in the fetoscopic videos. The RMSE and photometric errors for DIH are low compared to FEAT but are always higher compared to DSM (e.g. RMSE on EX for DIH (1.64) and DSM (0.38)). In DIH, this error accumulated over time during mosaic generation and resulted in a large drift. For EX, PHN1, PHN2 and INVI sequences, the average RMSE errors are 0.38, 0.32, 0.35 and 0.34, and photometric errors are 0.98, 1.76, 1.52, 2.42, respectively.

Mosaics generated using the proposed DSM for the EX, PHN1, PHN2 and INVI sequences are shown in Fig. 5. These mosaics are best assessed in the supplemental video that shows the qualitative comparison with respect to FEAT and DIH. DSM created a meaningful mosaic for EX (unseen data) with minimum drift accumulation over time which can be observed from the start and end frames in Fig. 5(a). PHN1 contained non-planar views without occlusions with a freehand circular trajectory. DSM generated reliable mosaics with minimum drift (Fig. 5(b)), however FEAT drifted away due to non-planar views, insufficient feature matches and long-range videos. PHN2 and INVI represent the most challenging scenarios containing highly non-planar views with heavy occlusions, low resolution and texture paucity. We observe from Fig. 5(c)(d) that although the generated mosaics can serve well for increasing the FoV, yet there is a noticeable drift due to highly challenging conditions. Such errors may be corrected by end-to-end training using the photometric loss [8].

The experimental results show that DSM is capable of handling varying visual quality (varying illumination, specular highlights and low resolution), planar and non-planar views with heavy occlusions. Qualitative evaluation on the unseen EX dataset verified the robustness and generalization capabilities of the proposed DSM. Unlike the existing methods that use external sensors for minimizing the drift [11], DSM relied only on image data and generated meaningful mosaics with minimum drift even for non-planar sequences.
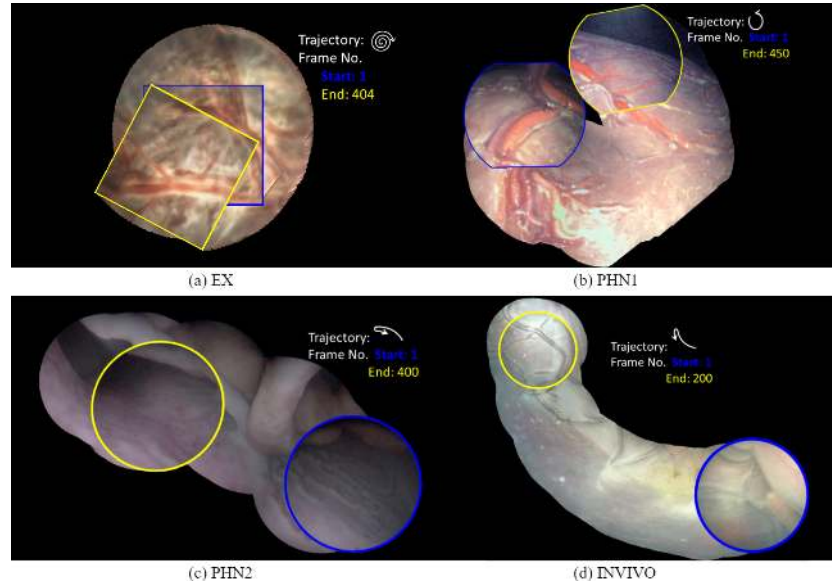
Fig. 5: Qualitative results of the proposed DSM on four diverse fetoscopic videos. The motion trajectories, start and end frames are marked for visualization.

6   Conclusion

We proposed a deep sequential mosaicking method for fetoscopic videos acquired through various sources which to our knowledge is a first. Our approach used an existing deep image homography network as a backbone for training but performed controlled data augmentation by assuming that there is only a small change in rotation and translation between two consecutive frames. Due to the lack of texture in fetoscopic sequences, varying specular highlights and turbid amniotic fluid, homography estimation varies between consecutive frames when selecting patch location randomly during testing. To overcome this problem, we proposed an outlier rejection step to obtain a reliable prediction in the least squares sense. Experimental evaluation on five diverse fetoscopic sequences showed that, unlike existing methods that drift rapidly in just a few frames, our method produced mosaics with less drift even for long-range sequences.

References

1. Baschat, A., et al.: Twin-to-twin transfusion syndrome (TTTS). Journal of Perinatal Medicine **39**(2), 107–112 (2011)

2. Brown, M., Lowe, D.G.: Automatic panoramic image stitching using invariant features. International Journal of Computer Vision **74**(1), 59–73 (2007)
3. Daga, P., et al.: Real-time mosaicing of fetoscopic videos using SIFT. In: Medical Imaging: Image-Guided Procedures (2016)
4. DeTone, D., Malisiewicz, T., Rabinovich, A.: Deep image homography estimation. RSS Workshop on Limits and Potentials of Deep Learning in Robotics (2016)
5. Dwyer, G., et al.: A continuum robot and control interface for surgical assist in fetoscopic interventions. IEEE robotics & automation letters **2**(3), 1656–1663 (2017)
6. Gaisser, F., et al.: Stable image registration for in-vivo fetoscopic panorama reconstruction. Journal of Imaging **4**(1), 24 (2018)
7. Malis, E., Vargas, M.: Deeper understanding of the homography decomposition for vision-based control. Ph.D. thesis, INRIA (2007)
8. Nguyen, T., et al.: Unsupervised deep homography: A fast and robust homography estimation model. IEEE Robotics and Automation Letters **3**(3), 2346–2353 (2018)
9. Peter, L., et al.: Retrieval and registration of long-range overlapping frames for scalable mosaicking of in vivo fetoscopy. IJCARS **13**(5), 713–720 (2018)
10. Reeff, M., Gerhard, F., Cattin, P., Gábor, S.: Mosaicing of endoscopic placenta images. INFORMATIK 2006–Informatik für Menschen, Band 1 (2006)
11. Tella-Amo, M., et al.: Probabilistic visual and electromagnetic data fusion for robust drift-free sequential mosaicking: application to fetoscopy. Journal of Medical Imaging **5**(2), 021217 (2018)
12. Tella-Amo, M., et al.: Pruning strategies for efficient online globally-consistent mosaicking in fetoscopy. Journal of Medical Imaging (2019)
13. Totz, J., et al.: Dense surface reconstruction for enhanced navigation in MIS. In: MICCAI. pp. 89–96. Springer (2011)
14. Vasconcelos, F., et al.: Towards computer-assisted ttts: Laser ablation detection for workflow segmentation from fetoscopic video. IJCARS **13**(10), 1661–1670 (2018)