Placental vessel segmentation and registration in fetoscopy: Literature review and MICCAI FetReg2021 challenge findings

Sophia Bano [a,*,1], Alessandro Casella [b,c,1], Francisco Vasconcelos [a], Abdul Qayyum [i], Abdesslam Benzinou [i], Moona Mazher [j], Fabrice Meriaudeau [k], Chiara Lena [c], Ilaria Anita Cintorrino [c], Gaia Romana De Paolis [c], Jessica Biagioli [c], Daria Grechishnikova [l], Jing Jiao [m], Bizhe Bai [n], Yanyan Qiao [o], Binod Bhattarai [a], Rebati Raman Gaire [p], Ronast Subedi [p], Eduard Vazquez [q], Szymon Płotka [r,s], Aneta Lisowska [r], Arkadiusz Sitek [r,t], George Attilakos [e,f], Ruwan Wimalasundera [e,f], Anna L. David [e,f,g], Dario Paladini [h], Jan Deprest [f,g], Elena De Momi [c], Leonardo S. Mattos [b], Sara Moccia [d], Danail Stoyanov [a]

[a] Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) and Department of Computer Science, University College London, UK
[b] Department of Advanced Robotics, Istituto Italiano di Tecnologia, Italy
[c] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Italy
[d] The BioRobotics Institute and Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Italy
[e] Fetal Medicine Unit, Elizabeth Garrett Anderson Wing, University College London Hospital, UK
[f] EGA Institute for Women's Health, Faculty of Population Health Sciences, University College London, UK
[g] Department of Development and Regeneration, University Hospital Leuven, Belgium
[h] Department of Fetal and Perinatal Medicine, Istituto "Giannina Gaslini", Italy
[i] ENIB, UMR CNRS 6285 LabSTICC, 29238, France
[j] Department of Computer Engineering and Mathematics, University Rovira i Virgili, Spain
[k] ImViA Laboratory, University of Bourgogne Franche-Comté, France
[l] Physics Department, Lomonosov Moscow State University, Russia
[m] Fudan University, China
[n] Medical Computer Vision and Robotics Group, Department of Mathematical and Computational Sciences, University of Toronto, Canada
[o] Shanghai MicroPort MedBot (Group) Co., Ltd, China
[p] NepAL Applied Mathematics and Informatics Institute for Research, Nepal
[q] Redev Technology, UK
[r] Sano Center for Computational Medicine, Poland
[s] Quantitative Healthcare Analysis Group, Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[t] Center for Advanced Medical Computing and Simulation, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States of America

ARTICLE INFO

ABSTRACT

Fetoscopy laser photocoagulation is a widely adopted procedure for treating Twin-to-Twin Transfusion Syndrome (TTTS). The procedure involves photocoagulation pathological anastomoses to restore a physiological blood exchange among twins. The procedure is particularly challenging, from the surgeon's side, due to the limited field of view, poor manoeuvrability of the fetoscope, poor visibility due to amniotic fluid turbidity, and variability in illumination. These challenges may lead to increased surgery time and incomplete ablation of pathological anastomoses, resulting in persistent TTTS. Computer-assisted intervention (CAI) can provide TTTS surgeons with decision support and context awareness by identifying key structures in the scene and expanding the fetoscopic field of view through video mosaicking. Research in this domain has been hampered by the lack of high-quality data to design, develop and test CAI algorithms. Through the *Fetoscopic Placental Vessel Segmentation and Registration (FetReg2021)* challenge, which was organized as part of the MICCAI2021 Endoscopic Vision (EndoVis) challenge, we released the first large-scale multi-center TTTS dataset for the development of generalized and robust semantic segmentation and video mosaicking algorithms with a focus on creating drift-free mosaics from long duration fetoscopy videos. For this challenge, we released a dataset of 2060 images, pixel-annotated for vessels, tool, fetus and background classes, from 18 in-vivo TTTS fetoscopy

---

* Corresponding author.
  E-mail address: sophia.bano@ucl.ac.uk (S. Bano).
[1] Equal contribution.

procedures and 18 short video clips of an average length of 411 frames for developing placental scene segmentation and frame registration for mosaicking techniques. Seven teams participated in this challenge and their model performance was assessed on an unseen test dataset of 658 pixel-annotated images from 6 fetoscopic procedures and 6 short clips. For the segmentation task, overall baseline performed was the top performing (aggregated *mIoU* of 0.6763) and was the best on the vessel class (*mIoU* of 0.5817) while team RREB was the best on the tool (*mIoU* of 0.6335) and fetus (*mIoU* of 0.5178) classes. For the registration task, overall the baseline performed better than team SANO with an overall mean 5-frame SSIM of 0.9348. Qualitatively, it was observed that team SANO performed better in planar scenarios, while baseline was better in non-planner scenarios. The detailed analysis showed that no single team outperformed on all 6 test fetoscopic videos. The challenge provided an opportunity to create generalized solutions for fetoscopic scene understanding and mosaicking. In this paper, we present the findings of the FetReg2021 challenge, alongside reporting a detailed literature review for CAI in TTTS fetoscopy. Through this challenge, its analysis and the release of multi-center fetoscopic data, we provide a benchmark for future research in this field.

1. Introduction

Twin-to-Twin Transfusion Syndrome (TTTS) is a severe complication of monochorionic twin pregnancies. TTTS is characterized by an unbalanced and chronic blood transfer from one twin (the donor twin) to the other (the recipient twin) through placental anastomoses (Baschat et al., 2011). This shared circulation is responsible for serious complications, which may lead to profound fetal hemodynamic and cardiovascular disturbances (Lewi et al., 2013). In 2004, a randomized, controlled trial demonstrated that fetoscopic laser ablation of placental anastomoses in TTTS had a higher survival rate for at least one twin than other treatments, such as serial amnioreduction. Laser ablation further showed a lower incidence of complications, such as cystic periventricular leukomalacia and neurologic complications (Senat et al., 2004). The trial included pregnancy at 16–26 weeks' gestation. Such results were confirmed for pregnancy before 17 and after 26 weeks' gestation (Baud et al., 2013). A description of all the steps that brought laser surgery for coagulation of placental anastomoses to be the elective treatment for TTTS can be found in Deprest et al. (2010).

Fetoscopic laser photocoagulation involves the ultrasound-guided insertion of a fetoscope into the amniotic sac. Through fetoscopic camera, the surgeon identifies abnormal anastomoses and laser ablates them to regulate the blood flow between the two fetuses (as illustrated in Fig. 1(a)). First attempts at laser coagulation included laser ablating all vessels that looked like anastomoses (a non-reproducible and operator-dependent technique), and laser ablating all vessels crossing the inter-fetus membrane (an approach that relies on the assumption that all vessels crossing the dividing membrane are pathological anastomoses) (Quintero et al., 2007). Today, the recognized elective treatment is the *selective laser photocoagulation*, which consists of the precise identification and lasering of placental pathological anastomoses. The selective treatment relies on the identification of the anastomoses (shown in Fig. 1(b)) and their classification into *Arterio-Venous* (from donor to recipient, AVDR, or from the recipient to donor, AVRD), *Arterio-Arterial* (AA) or *Veno-Venous* (VV) anastomoses. The identified AVDR anastomoses are laser ablated to regulate the blood flow between the two fetuses.

Despite all the advancements in instrumentation and imaging for TTTS (Cincotta and Kumar, 2016; Maselli and Badillo, 2016), residual anastomoses after monochorionic placentas treated with fetoscopic laser surgery still represent an issue (Lopriore et al., 2007). This may be explained considering challenges from the surgeon's side, such as limited field of view (FoV), poor visibility and high inter-subject variability. In this complex scenario, computer-assisted intervention (CAI) and surgical data science (SDS) methodologies may be exploited to provide surgeons with context awareness and decision support. However, the research in this field is still in its infancy, and several challenges still have to be tackled (Pratt et al., 2015). These include dynamically changing views with poor texture visibility, low image resolution, non-planar view, especially in the case of the anterior placenta, occlusions due to the fetus and tool, fluid turbidity and specular highlights.

In the context of TTTS fetoscopy, approaches to anatomical landmark segmentation (inter-fetus membrane, vessel) (Casella et al., 2020, 2021; Sadda et al., 2019; Bano et al., 2020a), event detection (Vasconcelos et al., 2018; Bano et al., 2020c) and mosaicking (Gaisser et al., 2018; Tella-Amo et al., 2019; Peter et al., 2018; Bano et al., 2020a,b) exist (Section 2). Even though fetoscopic videos have large inter- and intra-procedure variability, the majority of the segmentation and event detection approaches are validated on a small subset of in-vivo TTTS videos. Existing mosaicking approaches are validated only on a small subset of ex-vivo (Tella-Amo et al., 2019), in-vivo (Peter et al., 2018; Bano et al., 2020a) or underwater phantom sequences (Gaisser et al., 2018). Intensity-based image registration (Bano et al., 2020a; Li et al., 2021) methods rely on placental vessel segmentation maps for registration which facilitated in overcoming some visibility challenges (e.g., floating particles, poor illumination), however, such method fails when the predicted segmentation map is inaccurate, or the vessels are inconsistent across frames or are absent from the view. Deep learning-based flow-field matching for mosaicking (Alabi et al., 2022) has also been proposed, which results in accurate registration even in regions with poor or weak vessels but such an approach fails when the fetoscopic scene is homogenous having poor texture.

In fetoscopy, a major effort is needed to collect large, high-quality, multi-center datasets that can capture the variability of fetoscopic video. This reflects a well-known problem in the medical image analysis community (Litjens et al., 2017) that is currently addressed by organizing international initiatives such as Grand Challenge.[2]

1.1. Our contributions

Placental Vessel Segmentation and Registration for Mosaicking (FetReg2021)[3] challenge is a crowdsourcing initiative to address key problems in fetoscopy towards developing CAI techniques for providing TTTS with decision support and context awareness. With FetReg2021, we collected a large multi-center dataset to better capture not only inter- and intra-procedure variability but also inter-domain (data captured in two different clinical sites) variability. The FetReg2021 dataset can support developing robust and generalized models, paving the way for the translation of deep-learning methodologies in the actual surgical practice. The dataset is available to the research community,[4] under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0), to foster research in the field. FetReg2021 was organized as part of the MICCAI 2021 Endoscopic Vision (EndoVis)[5] challenge, and aimed at solving two tasks: placental scene segmentation and frame registration for mosaicking.

---

[2] https://grand-challenge.org/
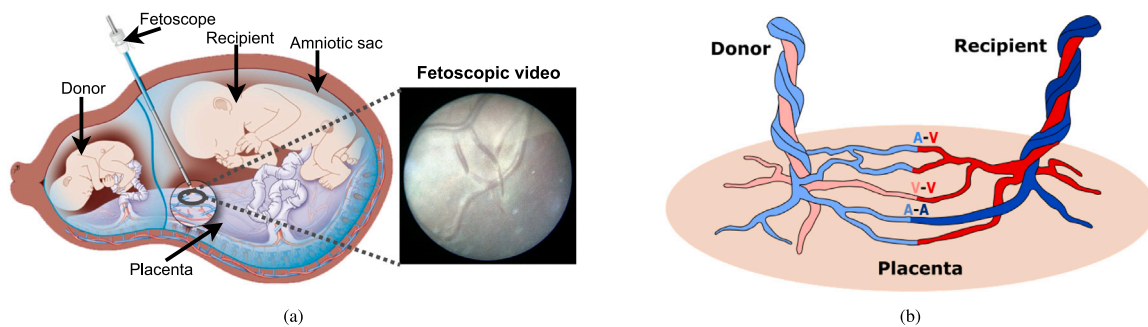
[3] FetReg challenge website: https://www.synapse.org/#!Synapse:syn25313156/.

[4] FetReg dataset: https://www.ucl.ac.uk/interventional-surgical-sciences/weiss-open-research/weiss-open-data-server/fetreg.

[5] EndoVis Challenges: https://endovis.grand-challenge.org/.

Fig. 1. Illustrations of Twin-to-Twin Transfusion Syndrome. (a) shows the fetoscopic laser photocoagulation procedure, where the field of view of the fetoscope is extremely narrow. (b) shows the types of anastomoses (i) A-V: arterio-venous, (ii) V-V: veno-venous, and (iii) A-A: arterio-arterial. In the placenta, conversely from body circulatory system, arteries carry deoxygenated blood (in blue), and veins carry oxygenated blood (in red).

Table 1

Overview of the existing segmentation (Section 2.1–2.1.2, event detection (Section 2.3) and video mosaicking methods (Section 2.2) for fetoscopy. The type of dataset used in each method is also reported. Key: IFM - inter-fetus membrane; GMS - grid-based motion statistics; EMT - electromagnetic tracker.

Reference	Task	Methodology	Imaging type
Almoussa et al. (2011)	Vessel segmentation	Hessian filter and Neural Network trained on handcrafted features	Ex-vivo
Chang et al. (2013)	Vessel segmentation	Combined Enhancement Filters	Ex-vivo (150 images)
Sadda et al. (2019)	Vessel segmentation	Convolutional Neural Network (U-Net)	In-vivo (345 frames from 10 TTTS procedures)
Bano et al. (2019)	Vessel segmentation	Convolutional Neural Network	In-vivo (483 frames from 6 TTTS procedures)
Casella et al. (2020)	IFM segmentation	Adversarial Neural Network (ResNet)	In-vivo (900 frames from 6 TTTS procedures)
Casella et al. (2021)	IFM segmentation	Spatio-temporal Adversarial Neural Network (3D DenseNet)	In-vivo (2000 frames from 20 TTTS procedures)[a]
Reeff et al. (2006)	Mosaicking	Hybrid feature and intensity-based	In water ex-vivo placenta
Daga et al. (2016)	Mosaicking	Feature-based with GPU for real time computation	Ex-vivo, Phantom placenta
Tella et al. (2016)	Mosaicking	Combined EM and visual tracking probablistic model	Ex-vivo w/laparoscope& EMT
Gaisser et al. (2016)	Mosaicking	Deep-learned features through contrastive loss	Ex-vivo and Phantom placenta video frames
Yang et al. (2016)	Mosaicking	SURF features matching and RANSAC for transformation estimation	Ex-vivo and monkey placentas w/laparoscope
Gaisser et al. (2017)	Mosaicking	Handcrafted features and LMedS for transformation estimation	Ex-vivo, In water placenta phantom
Tella-Amo et al. (2018)	Mosaicking	Combined EM and visual tracking with bundle adjustment	Ex-vivo placenta w/laparoscope & EMT
Gaisser et al. (2018)	Mosaicking	Extended (Gaisser et al., 2016) to detect stable vessel regions	In water placenta phantom
Sadda et al. (2018)	Mosaicking	AGAST detector with SIFT followed by GMS matching	In-vivo (# frames/clips)
Peter et al. (2018)	Mosaicking	Direct pixel-wise alignment of image gradient orientations	In-vivo (# frames/clips)
Tella-Amo et al. (2019)	Mosaicking	Pruning through EM and super frame generation	Ex-vivo placenta w/laparoscope & EMT
Bano et al. (2019, 2020a)	Mosaicking	Deep learning-based four point registration in consecutive images	Synthetic, Ex-vivo, Phantom, In-vivo phantom)
Bano et al. (2020a)	Mosaicking	Direct alignment of predicted vessel maps	In-vivo fetoscopy placenta data (6 procedures) [b]
Li et al. (2021)	Mosaicking	Direct alignment of predicted vessel with graph optimization	In-vivo fetoscopy placenta data (3 procedures) [b]
Alabi et al. (2022)	Mosaicking	FlowNet 2.0 with robust estimation for direct registration	Extended in-vivo fetoscopy data (6 procedures) [b]
Casella et al. (2022)	Mosaicking	Learning-based keypoint matching for registration	Extended in-vivo placenta data (6 procedures) [b]
Bano et al. (2022)	Mosaicking	Placental vessel-guided detector-free matching for registration	Extended in-vivo fetoscopy data (6 procedures) [b]
Vasconcelos et al. (2018)	Ablation detection	Binary classification using ResNet	In-vivo fetoscopy videos (5 procedures)
Bano et al. (2020c)	Event detection	Spatio-temporal model for multi-label classification	In-vivo fetoscopy videos (7 procedures)
This work which includes the FetReg dataset (Bano et al., 2021)	Segmentation and Mosaicking	Comparison of segmentation and mosaicking methods submitted to the FetReg challenge	Multi-center data (2718 annotated frames from 24 TTTS procedures and 9616 unannotated video clips

[a] Inter-Fetus Membrane Segmentation Dataset: https://zenodo.org/record/7259050.

[b] Fetoscopy Placenta Dataset: https://www.ucl.ac.uk/interventional-surgical-sciences/fetoscopy-placenta-data.

In this paper, we present the results and findings of the FetReg2021 challenge, in which 7 teams participated. We further provide a detailed review of the relevant literature on CAI for fetoscopy. To conclude, we benchmark FetReg2021 participants' methods against the existing state-of-the-art in fetoscopic scene segmentation and mosaicking method.

2. Related work

This section surveys the most relevant CAI methods developed in the field of TTTS surgery (see Table 1). This includes anatomical structure segmentation (Section 2.1), mosaicking and navigation (Section 2.2), and surgical event recognition (Section 2.3).

2.1. Anatomical structure segmentation

Image segmentation is one of the most explored tasks in medical image analysis. Segmentation from intra-operative images aims at supporting surgeons by enhancing the visibility of relevant structures (e.g., blood vessels) but presents additional challenges over anatomical image analysis due to poor texture and uncertain contours. Segmentation algorithms for TTTS partition mainly focus on vessel (Section 2.1.1) and placenta (Section 2.1.2) segmentation, as reference anatomical structures to provide surgeons with context awareness.

2.1.1. Placental vessel segmentation

Since the abnormal distribution of the anastomoses on the placenta is responsible for TTTS, exploration of its vascular network is crucial during the photocoagulation procedure. The work presented by Almoussa et al. (2011) is among the first in the field. The work, developed and tested with ex-vivo images, combined Hessian-based filtering and a custom neural network trained on handcrafted features. The approach was improved by Chang et al. (2013), which introduced a vessel enhancement filter that combined multi-scale and curvilinear filter matching. The multi-scale filter extends the Hessian filter, introducing

two scaling parameters to tune vesselness sensitivity. The curvilinear filter matches refined vessel segmentation, preserving all the structures that fit in the vessel shape template defined by a curvilinear function. The main limitation of both methods (Almoussa et al., 2011; Chang et al., 2013) lies in the analysis of ex-vivo images, which present different characteristics than in-vivo ones. More importantly, Hessian-based methods have been proven to perform poorly in the case of tortuous and irregular vessels (Moccia et al., 2018).

More recently, researchers have focused their attention on Convolutional Neural Networks (CNNs) to tackle the variability of intra-operative TTTS frames. Sadda et al. (2019) used U-Net, achieving segmentation performance in terms of Dice Similarity Coefficient (DSC) on a dataset of 345 in-vivo fetoscopic frames of 0.55 ± 0.22. U-Net is further explored by Bano et al. (2020a), which used segmented vessels as a prior for fetoscopic mosaicking (Section 2.2.3). The authors tested several versions of U-Net, including the original version by Ronneberger et al. (2015), and U-Net with different backbones (i.e. VGG16, ResNet50 and ResNet101). The segmentation performance was evaluated on a dataset of 483 in-vivo images from six TTTS surgery, the first publicly available fetoscopy placenta dataset released in Bano et al. (2020a).

Despite the advances introduced by CNNs, the state-of-the-art methods cannot tackle the high variability of intraoperative images. From one side, encoder–decoder architectures trained to minimize cross-entropy and DSC loss fail in segmenting poorly contrasted vessels and vessels with uneven margins. Furthermore, the datasets used to train these algorithms are small and the challenges of intra-operative images, as listed in Section 1, are not always represented.

Research in this field is strongly limited by the low availability of comprehensive expert-annotated datasets collected in different surgical settings that could encode such variability. This is mainly due to the low incidence of TTTS, which makes systematic data collection difficult, and the lack of annotators with sufficient domain expertise to ensure clinically correct groundtruth.

2.1.2. Inter-fetus membrane segmentation

At the beginning of the surgical treatment, due to the very limited FoV and poor image quality, the surgeon finds a reference for orientation within the amniotic cavity. The structure identified for this purpose is the inter-fetus membrane. The visibility of this membrane can be very variable, depending on the chorion characteristics, in addition to the challenges described so far in fetoscopic images. Once located, the surgeon refers to the inter-fetus membrane as a navigation reference during placental vascular network exploration.

Automatic inter-fetus membrane segmentation has been introduced by Casella et al. (2020) where an adversarial segmentation network based on ResNet was proposed to enforce placenta-shape constraining. The method was tested on a dataset of 900 intraoperative frames from 6 TTTS patients with an average DSC of 91.91%. Despite the promising results, this method suffered when illumination was too high or low, so the membrane was barely visible in such conditions.

The work by Casella et al. (2020) was extended (Casella et al., 2021) by exploiting dense connectivity and spatio-temporal information to improve membrane segmentation accuracy and tackle high illumination variability. The inter-fetus membrane segmentation performance outperformed the method previously proposed when tested on the first publicly available dataset (released in Casella et al. (2020)) of 2000 in-vivo images from 20 TTTS surgeries.

Despite the promising results achieved in the literature, the task of inter-fetus membrane segmentation remains poorly explored and requires further research for performance improvement and generalization.

2.2. Fetoscopic mosaicking and navigation

Video mosaicking aims at generating an expanded FoV image of the scene by registering and stitching overlapping video frames. Video mosaicking of high-resolution images has been extensively used as navigation guidance in the context of aerial, underwater, and street view imaging and also in consumer photography to build panorama shots. However, the outputs from off-the-shelf mosaicking methods have significantly poorer quality or fail completely when applied to fetoscopy videos due to the added visibility challenges of intra-operative images. Nevertheless, fetoscopy video mosaicking remains an active research topic within the context of computer-assisted intervention. Such a technique can facilitate the surgeon during the procedure in better localization of the anastomotic sites, which can improve the procedural outcomes.

Mosaicking for FoV expansion in fetoscopy has been explored using handcrafted feature-based and hybrid methods (Section 2.2.1), intensity-based (Section 2.2.2), and deep learning-based (Section 2.2.3) methods. These methods are either devised for synthetic placental images, ex-vivo placental images/videos or in-vivo videos.

2.2.1. Handcrafted feature-based and hybrid methods

Feature-based methods involve detecting and matching features across adjacent or overlapping frames, followed by estimating the transformation between the image pairs. On the other hand, hybrid methods utilize multimodal data (a combination of image and electromagnetic tracking data) or a combination of feature-based and intensity-based methods.

Early approaches focused on accomplishing fetoscopic mosaicking from videos or overlapping a pair of images only for image registration and mosaicking. Reeff et al. (2006) proposed a hybrid method that used classical feature detection and matching approach for first estimating the transformation of each image with respect to a reference frame, followed by global optimization by minimizing the sum of the squared differences of pixel intensities between two images. Multi-band blending was applied for seamless stitching. For testing the hybrid method, the authors recorded one ex-vivo placenta fixed in a hemispherical receptacle submerged in water to mimic an in-vivo imaging scenario. Such an experiment also allowed capturing camera calibration to remove lens distortion. A short sequence of 40 frames sampled at 3 frames per second was used for the evaluation. The matched feature correspondences were visually analyzed to mark them as correct or incorrect, which is a labor-intensive task. The generated mosaic with and without global optimization was shown for qualitative comparison.

Handcrafted feature-based methods, similar to what is commonly used in high-resolution image stitching in computer vision, were also explored for fetoscopic mosaicking. Daga et al. (2016) presented the first approach towards generating real-time mosaics. The approach considered using SIFT for feature detection and matching. For real-time computation, texture memory was used on GPU for computing extremes of the difference of Gaussian (DoG) that describes SIFT features. Planar images of ex-vivo phantom placenta recorded by mounting a fetoscope to a KUKA robotic arm were used for validating the approach. The robot was programmed to follow a spiral path that facilitated qualitative evaluation. Yang et al. (2016) proposed a SURF feature detection and matching based approach for generating mosaics from 100 frames long sequences that captured ex-vivo phantom and monkey placentas. Additionally, pair of images correspondence failure approach was proposed based on the statistical attributes of the feature distribution and an adaptive updating mechanism for parameter tuning to recover registration failures. Gaisser et al. (2017) used different keypoint descriptors (SIFT, SURF, ORB) along with the Least Median of Squares (LMedS) for estimating the transformation between overlapping pairs of images.

Through experiments on both ex-vivo and in-water phantom sequences, the authors showed that handcrafted features return either no features or low confidence features due to texture paucity and

dynamically changing visual conditions. This leads to inaccurate or poor transformation estimation.

Sadda et al. (2018) proposed a feature-based method that relied on extracting AGAST corner detector (Mair et al., 2010), SIFT as descriptor and grid-based motion statistics (GMS) (Bian et al., 2017) for refining feature matching for homography estimation. The validation was performed on 22 in-vivo fetoscopic image pairs. Additionally, in a hybrid approach by Sadda et al. (2019), vessel segmentation masks were also used for selecting AGAST features only around the vessel regions. However, the reported error was large mainly because of linear and single vessels in the 22 image pairs under analysis. Using handcrafted feature descriptors such as SIFT shows poor performance in the case of in-vivo placental videos due to the added challenges introduced by poor visibility, texture paucity and low-resolution imaging.

A few approaches used an additional electromagnetic tracker in an ex-vivo setting to design a feature-based method for improved mosaicking. Tella et al. (2016) and Tella-Amo et al. (2018) assumed the placenta to be planar and static and used a combination of visual and electromagnetic tracker information for generating robust and drift-free mosaics. Mosaicking performance was increased by Tella-Amo et al. (2019), where the pruning of overlapping frames and generation of a superframe for reducing computational time was proposed. An Aurora electromagnetic tracker (EMT) was mounted on the tip of a laparoscope to obtain camera pose measurements. Using this setup, a data sequence of 701 frames was captured from a phantom (i.e., a printed image of a placenta). Additionally, a synthetic sequence of 273 frames following only planar motion was also generated for quantitative evaluation. The camera pose measurements from the EMT were incorporated with frame-based visual information using a probabilistic model to obtain globally consistent sequential mosaics. It is worth mentioning that laparoscopic cameras used are considerably better than fetoscopic cameras. However, current clinical regulations and the limited form factor of the fetoscope hinder the use of such a tracker in intraoperative settings.

2.2.2. Intensity-based methods

Intensity-based image registration is an iterative process that uses raw pixel values for direct registration by first selecting features, such as edges, and contours, followed by a metric, such as mutual information, cross-correlation, the sum of squared difference, absolute difference, for describing how similar two overlapping input images are and an optimizer for obtaining the best alignment through fitting a spatial transformation model.

The use of direct pixel-wise alignment of oriented image gradients for creating a mosaic was proposed by Peter et al. (2018) that was validated on only one in-vivo fetoscopic sequence of 600 frames. An offline bag of words was used to improve the global consistency of the generated mosaic.

Bano et al. (2020a) proposed a placental vessel-based direct registration approach. A U-Net model was trained on a dataset of 483 vessel annotated images from 6 in-vivo fetoscopy for segmenting vessels. The vessel maps from consecutive frames were registered, estimating the affine transformation between the frames. Testing was performed on 6 additional in-vivo fetoscopy video clips. The approach facilitated overcoming visibility challenges, such as floating particles and varying illumination. However, the method failed when the predicted segmentation map was inaccurate or in views with thin or no vessels. Li et al. (2021) further extended this approach to propose a graph-based globally optimal image mosaicking method. The method detected loop closures with a bad-of-words scheme followed by direct image registration. Only 3 out of 6 in-invivo videos had loop closures present in them. Global refinement in alignment is then performed through G2O framwork (Kümmerle et al., 2011).

2.2.3. Deep learning-based methods

Existing deep learning-based methods for fetoscopic mosaicking mainly focused on training a CNN network (Bano et al., 2019, 2020b)

for directly estimating homography between adjacent frames, extracting stable regions (Gaisser et al., 2016) in a view, or relying on flow fields (Alabi et al., 2022) for robust pair-wise images registration.

A deep learning-based feature extractor was proposed by Gaisser et al. (2016) that used similarity learning using contrastive loss when training a Siamese convolutional neural network (CNN) architecture between pairs of similar and dissimilar small patches extracted from ex-vivo placental images. The learned feature extractor was used for extracting features from pairs of overlapping images, followed by using LMedS for the transformation estimation. Due to motion blur and texture paucity that affected the feature extractor performance, the method was validated only on a short sequence (26 frames) that captured an ex-vivo phantom placenta. Gaisser et al. (2018) extended their similarity learning approach (Gaisser et al., 2016) for detecting stable regions on the vessels of the placenta. These stable regions' representation is used as features for placental image registration in an in-water phantom setting. The obtained homography estimation did not result in highly accurate registration, as the learned regions were not robust to visual variability in underwater placental scenes.

Methods for estimating 4-point homography using direct registration with deep learning exist in computer vision literature (DeTone et al., 2016; Nguyen et al., 2018). (Bano et al., 2019, 2020b) extended (DeTone et al., 2016) to propose one of the first homography-based methods for fetoscopic mosaicking, which was tested on 5 diverse placental sequences, namely, synthetic sequence of 811 frames, ex-vivo placenta planar sequence of 404 frames, ex-vivo phantom placenta sequence of 681 frames, in-vivo phantom placenta sequence of 350 frames and in-vivo TTTS fetoscopic video of 150 frames. In Bano et al. (2019, 2020b), a VGG-like model was trained to estimate 4-point homography between two patches extracted from the same image with known transformation. Controlled data augmentation was applied to the two patches for network training. Filtering is then applied during testing to obtain the most consistent homography estimation. The proposed approach led to advancing the literature on fetoscopic mosaicking, although the network mainly focused on estimating rigid transformation (rotation and translation) between adjacent frames due to controlled data augmentation. As a result, the generated mosaics in non-planar sequences accumulated drift over time.

More recently, deep learning-based optical flow combined with inconsistent motion filtering for robust fetoscopy mosaicking has been proposed (Alabi et al., 2022). Their method relied on FlowNet-v2 (Ilg et al., 2017) for obtaining dense correspondence between adjacent frames, robust estimation using RANSAC and local refinement for removing the effect of floating particles and specularities for improved registration. Unlike (Bano et al., 2020a) which used placental vessel prediction to drive mosaicking, (Alabi et al., 2022) did not rely on vessels, as a result, it managed to generate robust and consistent mosaic for a longer duration of fetoscopic videos. Their approach was tested on the extended fetoscopy placenta dataset from Bano et al. (2020a).

Recent computer vision literature has also introduced deep learning-based interest point descriptors (DeTone et al., 2018; Sarlin et al., 2020) and detector-free dense feature matching (Sun et al., 2021) techniques. These techniques have shown robustness in multiview feature matching. Inspired from DeTone et al. (2018), Casella et al. (2022) proposed a learning-based keypoint proposal network and an encoding strategy for filtering irrelevant keypoints based on fetoscopic image segmentation and inconsistent homographies for producing robust and drift-free fetoscopic mosaics. Bano et al. (2022) proposed a placental vessel-guided hybrid framework for mosaicking that relies on best of Bano et al. (2020a), Sun et al. (2021). The framework combines these two methods through a selection mechanism based on the appearance consistency of placental vessels and photometric error minimization for choosing the best homography estimation between consecutive frames. Casella et al. (2022) and Bano et al. (2022) methods have been validated using the extended fetoscopy placenta dataset from Bano et al. (2020a).

While these approaches significantly improved fetoscopic mosaicking, further analysis is needed to investigate its performance in low-textured and highly non-planar placental regions.

Table 2

Summary of the *EndoVis FetReg 2021* training and testing dataset. For each video, center ID (I - UCLH, II - IGG), image resolution, the number of annotated frames (for the segmentation task), the occurrence of each class per frame and the average number of pixels per class per frame are presented. For the registration task, the number of unlabeled frames in each video clip is provided. Key: BG - background.

Sr.	Video name	Center ID	Image Resolution (pixels)	No. of labeled frames	Occurrence (frame)			Occurrence (Avg. pixels)				Unlabel-led clips # frames
					Vessel	Tool	Fetus	BG	Vessel	Tool	Fetus	
TRAINING DATASET												
1.	Video001	I	470 × 470	152	152	21	11	196 463	21 493	1462	1482	346
2.	Video002	I	540 × 540	153	153	35	1	271 564	16 989	3019	27	259
3.	Video003	I	550 × 550	117	117	52	32	260 909	27 962	3912	9716	541
4.	Video004	II	480 × 480	100	100	21	18	212 542	14 988	1063	1806	388
5.	Video005	II	500 × 500	100	100	35	30	203 372	34 350	2244	10 034	722
6.	Video006	II	450 × 450	100	100	49	4	171 684	28 384	1779	653	452
7.	Video007	I	640 × 640	140	140	30	3	366 177	37 703	4669	1052	316
8.	Video008	I	720 × 720	110	105	80	34	465 524	28 049	13 098	11 729	295
9.	Video009	I	660 × 660	105	104	40	14	353 721	68 621	7762	5496	265
10.	Video011	II	380 × 380	100	100	7	37	128 636	8959	184	6621	424
11.	Video013	I	680 × 680	124	124	54	21	411 713	36 907	8085	5695	247
12.	Video014	I	720 × 720	110	110	54	14	464 115	42 714	6223	5348	469
13.	Video016	II	380 × 380	100	100	16	20	129 888	11 331	448	2734	593
14.	Video017	II	400 × 400	100	97	20	3	151 143	7625	753	479	490
15.	Video018	I	400 × 400	100	100	26	11	139 530	15 935	1503	3032	352
16.	Video019	II	720 × 720	149	149	15	31	470 209	38 513	1676	8002	265
17.	Video022	II	400 × 400	100	100	12	1	138 097	21 000	650	253	348
18.	Video023	II	320 × 320	100	92	14	8	94 942	6256	375	828	639
All training videos				**2060**	2043	581	293	4 630 229	467 779	58 905	74 987	**7411**
TESTING DATASET												
19.	Video010	II	622 × 622	100	92	7	28	341 927	40 554	1726	19 410	320
20.	Video012	II	320 × 320	100	100	54	0	95 845	5132	1422	0	507
21.	Video015	I	720 × 720	125	124	83	28	452 552	47 221	12 082	6545	530
22.	Video020	I	720 × 720	123	100	15	1	436 842	59 884	15 259	6415	307
23.	Video024	II	320 × 320	100	110	72	13	203 372	34 350	2244	10 034	269
24.	Video025	I	720 × 720	110	648	320	83	459 947	43 189	9801	5464	272
All testing videos				**658**	648	320	83	1 880 090	205 009	40 638	37 879	**2205**

2.3. Surgical event recognition

TTTS laser therapy has a relatively simple workflow with an initial inspection of the vasculature and placenta surface to identify and visualize photocoagulation targets. Fetoscopic laser therapy is conducted by photocoagulation of each identified target in sequence. Automatic identification of these surgical phases and surgical events is an essential step towards general scene understanding and tracking of the photocoagulation targets. This identification can provide temporal context for tasks such as segmentation and mosaicking. It could also provide prior to finding the most reliable images for registration (before ablation) or identify changes in the appearance of the scene (after ablation).

The CAI literature has hardly explored event detection or workflow analysis methods. Vasconcelos et al. (2018) used a ResNet encoder to detect ablation in TTTS procedures, additionally indicating when the surgeon is ready for ablating the target vessel. The method was validated on 5 in-vivo fetoscopic videos. Bano et al. (2020c) combined CNNs and recurrent networks for the spatio-temporal identification of fetoscopic events, including clear view, occlusion (i.e., fetus or working channel port in the FoV), laser tool presence, and ablating laser tool present. The method was effective in identifying clear view segments (Bano et al., 2020c) suitable for mosaicking and was validated on 7 in-vivo fetoscopic videos. Due to inter- and intra-case variability present in fetosopic videos, evaluation on a larger dataset is needed to validate the generalization capabilities of the current surgical event recognition methods.

3. The FetReg challenge: Dataset, submission, evaluation

In this section, we present the dataset of the *EndoVis FetReg 2021* challenge and its tasks (Section 3.1), the evaluation protocol designed to assess the performance of the participating methods (Section 3.2) and an overview of the challenge setup and submission protocol (Section 3.3).

3.1. Dataset and challenge tasks

The *EndoVis FetReg 2021* challenge aims at advancing the current state-of-the-art in placental vessel segmentation and mosaicking (Bano et al., 2020a) by providing a benchmark multi-center large-scale dataset that captured variability across different patients and different clinical institutions. We also aimed to perform out-of-sample testing to validate the generalization capabilities of trained models. The participants were required to complete two sub-tasks which are critical in fetoscopy, namely:

- **Task 1: Placental semantic segmentation**: The participants were required to segment four classes, namely, background, vessels, tool (ablation instrument, i.e. the tip of the laser probe) and fetus, on the provided dataset. Fetoscopic frames from 24 TTTS procedures collected in two different centers were annotated for the four classes that commonly occur during the procedure. This task was evaluated on unseen test data (6 videos) independent of the training data (18 videos). The segmentation task aimed to assess the generalization capability of segmentation models on unseen fetoscopic video frames.

- **Task 2: Registration for Mosaicking**: The participants were required to perform the registration of consecutive frames to create an expanded FoV image of the fetoscopic environment. Fetoscopic video clips from 18 multi-center fetoscopic procedures were provided as the training data. No registration annotations were provided, as it is not possible to get the groundtruth registration during the in-vivo clinical fetoscopy. The task was evaluated on 6 unseen video clips extracted from fetoscopic procedure videos, which were not part of the training data. The registration task aimed to assess the robustness and performance of registration methods for creating a drift-free mosaic from unseen data.

The *EndoVis FetReg 2021* dataset is unique as it is the first large-scale fetoscopic video dataset of 24 different TTTS fetoscopic procedures. The videos contained in this dataset are collected from two fetal surgery centers across Europe, namely,

- Center I: Fetal Medicine Unit, University College London Hospital (UCLH), London, UK,
- Center II: Department of Fetal and Perinatal Medicine, Istituto "Giannina Gaslini" (IGG), Genoa, Italy,

Both centers contributed with 12 TTTS fetoscopic laser photocoagulation videos each. A total of 9 videos from each center (18 videos in total) form the training set, while 3 videos from each center (6 videos in total) form the test set. Alongside capturing the intra-case and inter-case variability, the multi-center data collection allowed capturing the variability that arises due to different clinical settings and imaging equipment at different clinical sites. At UCLH, the data collection was carried out as part of the GIFT-Surg[6] project. The requirement for formal ethical approval was waived, as the data were fully anonymized in the corresponding clinical centers before being transferred to the organizers of the *EndoVis FetReg 2021* challenge.

3.1.1. Multi-center data comparison

Table 2 summarizes *EndoVis FetReg 2021* dataset characteristics and also indicates the center from which it is acquired. Videos from the two centers varied in terms of the resolution, imaging device and light source. The videos from UCLH are of higher resolution (minimum resolution: 470×470, maximum resolution: 720×720) with majority videos having 720p resolution compared to IGG (minimum resolution: 320×320, maximum resolution: 622×622) videos with a majority having 400p or lower resolution. From Figs. 4 and 5, we can observe that most of the IGG center videos have a dominant red spotlight light visible with most views appearing to be very close to the placental surface. On the other hand, no domain light reflection is visible in any of the UCLH center videos and the imaging device captured a relatively wider view compared to the IGG videos. Additionally, the frame appearance and quality changes in each video due to the large variation in intra-operative environment among different cases. Amniotic fluid turbidity resulting in poor visibility, artefacts introduced due to spotlight light source, low resolution, texture paucity, and non-planar views due to anterior placenta imaging, are some of the major factors that contribute to increasing the variability in the data from both centers. Large intra-case variations can also be observed from Figs. 4 and 5. All these factors contribute towards limiting the performance of the existing placental image segmentation and registration methods (Bano et al., 2020a, 2019, 2020b). The *EndoVis FetReg 2021* challenge provided an opportunity to make advancements in the current literature by designing and contributing novel segmentation and registration methods that are robust even in the presence of the above-mentioned challenges. Further details about the segmentation and registration datasets are provided in the following sections.

3.1.2. Dataset for placental semantic segmentation

Fetoscopy videos acquired from the two different fetal medicine centers were first decomposed into frames, and the excess black background was cropped to obtain squared images capturing mainly the fetoscope FoV. From each video, a subset of non-overlapping informative frames (in the range 100–150) is selected and manually annotated. All pixels in each image are labeled with background (class 0), placental vessel (1), ablation tool (2) or fetus class (3). Labels are mutually exclusive.

The annotation of 7 out of 24 videos was performed by four academic researchers and staff members with a solid background in fetoscopic imaging. Additionally, annotation services are obtained from
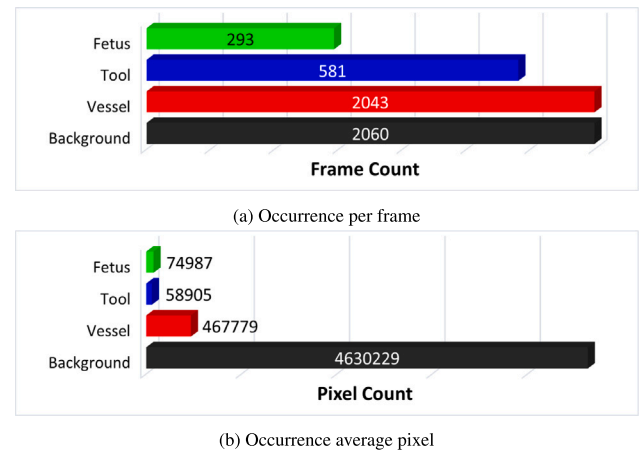

Fig. 2. Training dataset distribution: (a) and (b) segmentation classes and their overall distribution in the segmentation data.
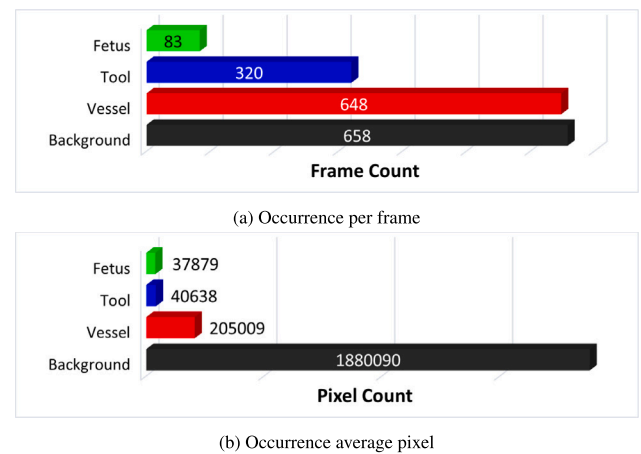

Fig. 3. Testing dataset distribution: (a) and (b) segmentation classes and their overall distribution n the segmentation data.

Humans in the Loop (HITL)[7] for a subset of videos 17 out of 24 videos), who provided annotators with clinical background. Each image was annotated once following a defined annotation protocol. All annotations were then verified by two academic researchers for their correctness and consistency. Finally, two fetal medicine specialists verified all the annotations to confirm the correctness and consistency of the labels. The publicly available Supervisely[8] platform was used for annotating the dataset.

The *FetReg* train and test dataset for the segmentation task contains 2060 and 658 annotated images from 18 and 6 different in-vivo TTTS fetoscopic procedures, respectively. Fig. 2(a) and Fig. 2(b) show the overall class occurrence per frame and class occurrence in average pixels per frame on the training dataset. The same for test dataset is shown in Fig. 3(a) and Fig. 3(b). Note that the frames present different resolutions as the fetoscopic videos are captured at different centers with different facilities (e.g., device, light scope). The dataset is highly unbalanced: *Vessel* is the most frequent class while *Tool* and *Fetus* are presented only in a small subset of images corresponding to 28% and 14%, respectively, of the training dataset and 48% and 13% of the test dataset. When observing the class occurrence in average pixels per image, the *Background* class is the most dominant, with *Vessel, Tool* and
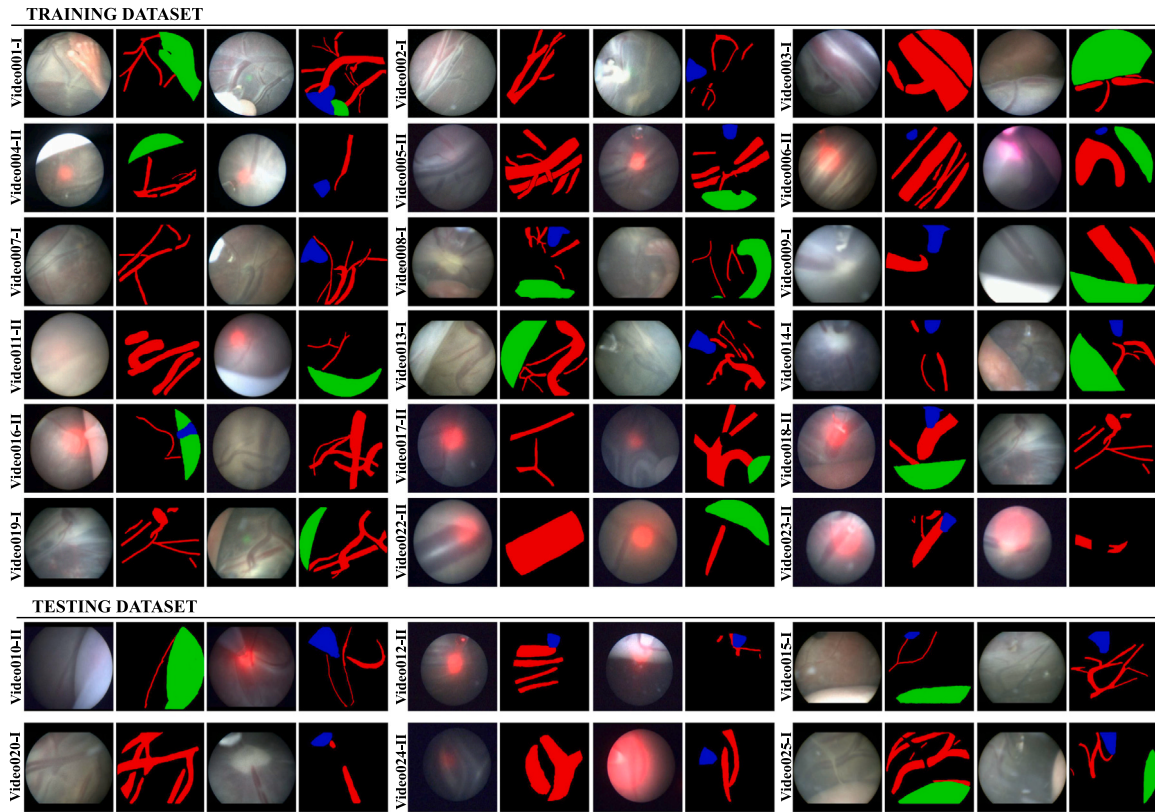

Fig. 4. Representative images from training and test datasets along with the segmentation annotations (groundtruth). Each center ID is also indicated next to video name (I - UCLH, II - IGG) for visual comparison of variabilities between the two centers.

Fetus occur 10%, 0.13% and 0.16% in train dataset and 11%, 0.22%, and 0.20% in the test dataset, respectively.

Fig. 4 shows some representative annotated frames from each video. Note that the frame appearance and quality change in each video due to the large variation in the intra-operative environment among different cases. Amniotic fluid turbidity resulting in poor visibility, artifacts introduced due to spotlight light source and reddish reflection introduced by the laser tool, low resolution, texture paucity, and non-planar views due to anterior placenta imaging are some of the major factors that contribute to increasing the variability in the data. Large intra-case variations can also be observed from these representative images. All these factors contribute towards limiting the performance of the existing placental image segmentation and registration methods (Bano et al., 2020a, 2019, 2020b). The *EndoVis FetReg 2021* challenge provided an opportunity to make advancements in the current literature by designing and contributing novel segmentation and registration methods that are robust even in the presence of the above-mentioned challenges.

3.1.3. Dataset for registration for mosaicking

A typical TTTS fetoscopy surgery takes approximately 30 min. Only a sub-set of fetoscopic frames is suitable for frame registration and mosaicking because fetuses, laser ablation fiber, and working channel port can occlude the field-of-view of the fetoscope. Mosaicking is mainly required in occlusion-free video segments that capture the surface of the placenta (Bano et al., 2020c) as these are the segments in which the surgeon is exploring the intraoperative environment to identify abnormal vascular connections. Expanding the FoV through mosaicking in these video segments can facilitate the procedure by providing better visualization of the environment.

For the registration for the mosaicking task, we have provided one video clip per video for all 18 procedures in the training dataset. Likewise, one clip per video from all 6 procedures in the test dataset is

selected for testing and validation. These frames are neither annotated with segmentation labels nor have registration groundtruth. The number of frames in each video clip is reported in Table 2 for training and test dataset. Representative frames from each clip are shown in 5.

Representative frames every 2 s from some video clips are shown in Fig. 5. Observe the variability in the appearance, lighting conditions and image quality in all video clips. Even though there is no noticeable deformation in fetoscopic videos, which is usually thought to occur due to breathing motion, the views can be non-planar as the placenta can be anterior or posterior. Moreover, there is no groundtruth camera motion and scene geometry that can be used to evaluate video registration approaches for in-vivo fetoscopy. In Section 3.2.2, we detail how this challenge is addressed with an evaluation metric that is correlated with good quality, consistent, and complete mosaics (Bano et al., 2020a).

3.2. Evaluation protocol

3.2.1. Segmentation evaluation

Intersection over union (IoU) is another most commonly used metric for evaluating segmentation algorithms which measure the spatial overlap between the predicted and groundtruth segmentation masks as:

$$IoU = \frac{TP}{TP + FP + FN} \tag{1}$$

where TP are the correctly classified pixels belonging to a class, FP are the pixels incorrectly predicted in a specific class, and FN are the pixels in a class incorrectly classified as not belonging to it. For evaluating the performance of segmentation models (Task 1), we compute for each frame provided in the test set the mean Intersection over Union (mIoU) per class between the prediction and the manually annotated segmentation masks. Overall mean mIoU over all three classes and all test samples are also computed and used for ranking different methods under comparison.
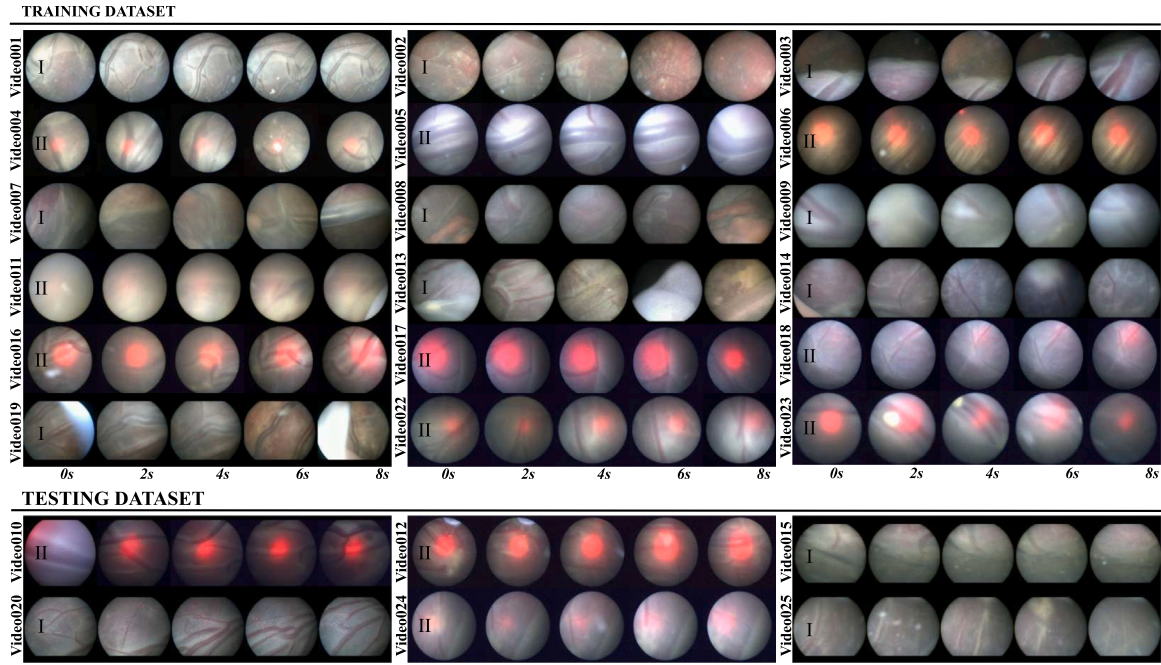
TRAINING DATASET

TESTING DATASET

Fig. 5. Representative frames from training and test datasets at every 2 *seconds*. These clips are unannotated and the length of each clip mentioned in Table 2. Center ID is also marked on each video sequence (I - UCLH, II - IGG) for visual comparison of the data from the two different centers.
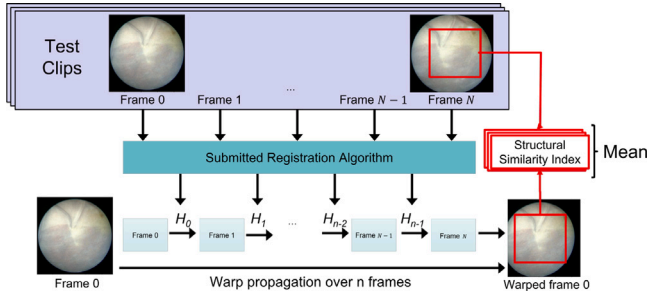

Fig. 6. Illustration of the N-frame SSIM evaluation metric from Bano et al. (2020a).

3.2.2. Frame registration and mosaicking evaluation

For evaluating homographies and mosaics (Task 2), we use the evaluation metric presented by Bano et al. (2020a) in the absence of groundtruth. The metric that we referred as N-frame structural similarity index measure (SSIM) aims to evaluate the consistency in the adjacent frames. A visual illustration of the N-frame SSIM metric is presented in Fig. 6. Given N consecutive frames and a set of $N - 1$ homographies $\{H_1, H_2, \ldots, H_{N-1}\}$, we evaluate the consistency between them. The ultimate clinical goal of fetoscopic registration is to generate consistent, comprehensible and complete mosaics that map the placental surface and guide the surgeon. Considering adjacent frames will have a large overlap along them, we evaluate the registration consistency between pairs of non-consecutive frames N frames apart that have a large overlap in the FoV and present a clear view of the placental surface. Consider a source image I_i, a target image I_{i+n}, and a homography transformation $H_{i \to i+n}$ between them, we define the consistency s between these two images as:

$$s_{i \to i+n} = \text{sim}(w(\tilde{I}_i, H_{i \to i+n}), \tilde{I}_{i+n}) \tag{2}$$

where sim is an image similarity metric that is computed based on the target image and warped source image, and $\tilde{I}$ is a smoothed version of the image I. *Smoothing* $\tilde{I}$ is obtained by applying a 9×9 Gaussian filter with a standard deviation of 2 to the original image I. This is fundamental to make the similarity metric robust to small outliers

(e.g., particles) and image discretization artifacts. For computing the *similarity*, we start by determining the overlap region between the target $\tilde{I}$ and the warped source $w(\tilde{I}_i, H_{i \to i+n})$, taking into account their circular edges. If the overlap contains less than 25% of $\tilde{I}$, we consider that the registration failed, as there will be no such cases in the evaluation pool. A rectangular crop fits the overlap, and the SSIM is calculated between the image pairs after being smoothed, warped, and cropped.

3.3. Challenge organization and timeline

The *FetReg 2021* challenge is a crowdsourcing initiative that was organized by Sophia Bano (University College London, London, UK), Alessandro Casella (Istituto Italiano di Tecnologia and Politecnico di Milano, Italy), Francisco Vasconcelos (University College London, London, UK), Sara Moccia (Scuola Superiore Sant'Anna, Italy) and Danail Stoyanov (University College London, London, UK). The *FetReg 2021* challenge was organized as part of the EndoVis challenge series, which is led by Stefanie Speidel (German Cancer Research Center, Heidelberg, Germany), Lena Maier-Hein (German Cancer Research Center, Heidelberg, Germany) and Danail Stoyanov (University College London, London, UK).

The FetReg challenge was organized according to The Biomedical Image Analysis Challenges (BIAS) (Maier-Hein et al., 2020) reporting guideline to enhance the quality and transparency of health research.

3.3.1. Challenge timeline and details release

The challenge timeline and submission statistics are presented in Fig. 7. The challenge was announced on April 1st 2021, through the FetReg2021 Synapse 3 website. The training dataset for task 1 and task 2 was released on May 1st and 29th, respectively. No restrictions were imposed on using additional publicly available datasets for training. A challenge description paper (Bano et al., 2021) that also included baseline method evaluation was also published on June 10th. All the details regarding the baseline methods (i.e., architecture, algorithms, and training settings) for segmentation and registration have been publicly disclosed along with its release. Additionally, a Slack support forum was launched for faster communication with the participants. Docker submission was opened on August 20th 2021, followed by the
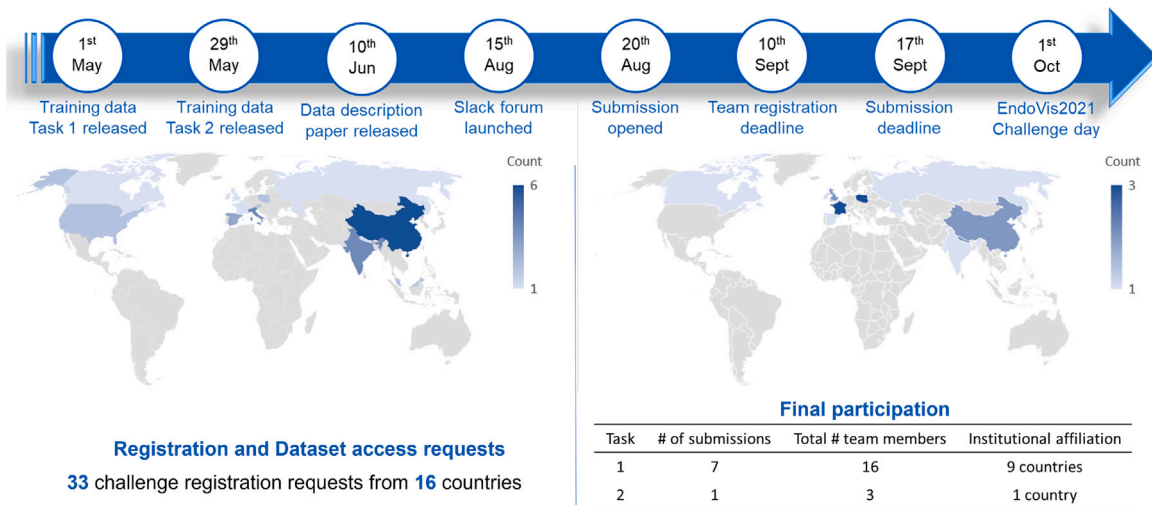
Fig. 7. FetReg2021 timeline and challenge participation statistics.

team registration deadline of September 10th, and the final submission deadline was set to September 17th. Members of the organizers' department may participate in the challenge but were not eligible for awards.

3.3.2. Submission protocol

The test dataset was not made available to the challenge participants to keep the comparison fair and avoid misuse of the test data during training. Each participating team was required to make submissions as a docker container that accepts a path to a folder containing video frames from a patient as input and outputs segmentation mask as an image (for task 1) or a text file with relative homography matrix (for task 2). Only fully automatic algorithms are allowed to participate in the challenge.

The teams could submit multiple docker dockers during the submission time (from August 20th to September 17th 2021) to check the validity of the docker. We provided the participants with docker examples for both tasks along with detailed submission guidelines through FetReg2021 GitHub repository.[9] The docker submission protocol is illustrated in Fig. 8. Each participating team submitted their docker through the Synapse platform. The submitted docker was verified for the validity of their output structure, i.e., they follow the same output format as requested and needed for the evaluation. Each participating team was then informed whether their submission passed the validity test. Each team was allowed to submit multiple dockers. However, only the last valid docker submission was used in the final evaluation.

3.3.3. Participation policy and statistics

Through the FegReg website, it was announced since the start of the challenge that the top three performing methods would be announced publicly during the challenge day, and the top method for each task would be awarded a prize from the sponsors. The remaining teams could decide whether their identity should be publicly revealed or not (e.g., in the challenge publication). All participating team, whose method achieved an overall *mIoU* of over 0.25 were included in this joint publication. Only one team was excluded as their method resulted in an extremely low *mIoU* of 0.060 on the test set (see Section 4.)

We received 33 challenge registration requests from 16 different countries. A total of 13 team registration requests with a total number
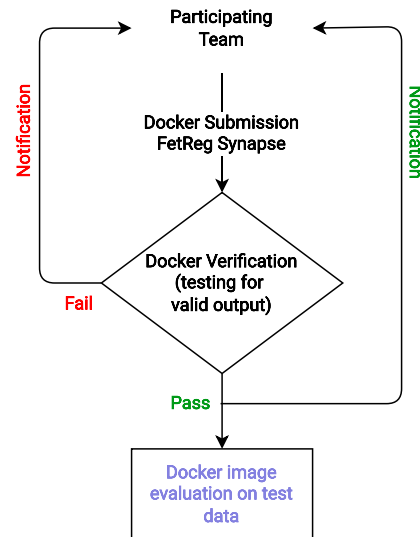
Fig. 8. FetReg2021 submission protocol illustrating the docker image verification protocol.

of 22 team members were received. For task 1, final submissions were received from 7 teams having 16 participants. For task 2, one submission was received, probably because of the challenging nature of this task.

We believe that the decrease in the number of teams participating in the challenge can be attributed to several factors. The difficulty of the tasks, particularly the mosaicking, was the primary reason why some participants were deterred. Some participants who had initially registered for the challenge may have opted to form groups with other participants to tackle both tasks but may have been dissuaded due to not seeing significant performance improvements in comparison to the baseline. Moreover, the high level of interest in our dataset may have encouraged some participants to join the challenge in order to gain early access to it. It is also worth highlighting that there is a general trend of receiving 3 times more registration requests than final submissions, particularly in EndoVis challenges which is generally due to research community interests in analyzing these unique datasets in the long run rather than participating in the challenges (Eisenmann et al., 2023).
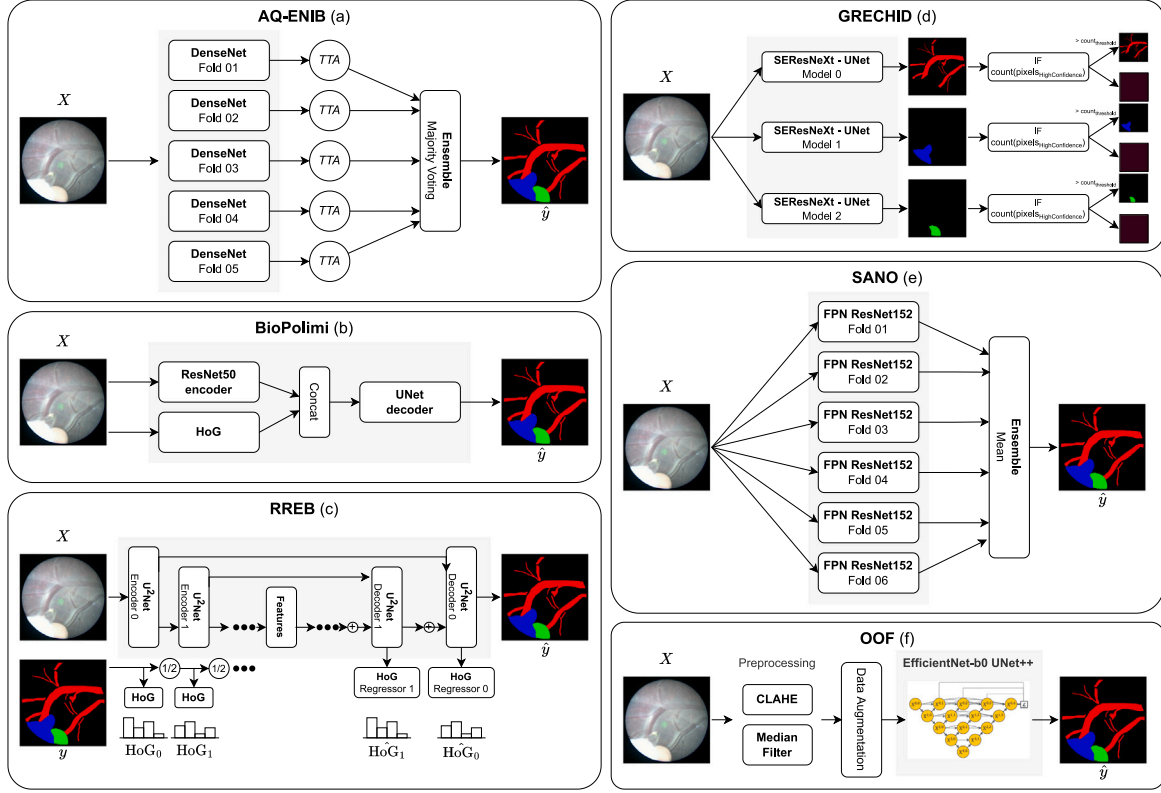
Fig. 9. Graphical overview of the participants' methodologies for Task 1 as described in Section 4 (Key: X - input frame; y - groundtruth; $\hat{y}$ - prediction). AQ-ENIB (a) proposed an ensemble of DenseNet models with Test Time Augment (TTA). BioPolimi (b) combined ResNet50 features with a Histogram of Oriented Gradients (HoG) computed on X. RREB (c) proposed a multi-task U^2Net for segmentation and multi-scale regression of HoG features ($\hat{\text{HoG}}_0$, $\hat{\text{HoG}}_1$, ...) computed on y (HoG_0, HoG_1, ...). GRECHID (d) used 3 SEResNeXt-UNet models individually trained on each class ensembled by thresholding, where $\text{pixels}_{HighConfidence}$ are pixels predicted with high confidence and $\text{count}_{threshold}$ is the empirical threshold. SANO (e) proposed a mean ensemble of Feature Pyramid Network (FPN) with ResNet152 backbone. OOF (f) used an EfficientNet UNet++, preprocessing images with contrast-limited adaptive histogram equalization (CLAHE) and median filter.

4. Summary of methods proposed by participating teams

In total, 7 teams participated in the challenge. Out of these, one team did not qualify to be included in this article as the achieved performance was extremely low with a *mIoU* of 0.060. In this section, we summarize the methodology proposed by each participating team.

4.1. AQ-ENIB

Team AQ-ENIB are Abdul Qayyum, Abdesslam Benzinou, Moona Mazher and Fabrice Meriaudeau from ENIB (France), University Rovira i Virgili (Spain) and University of Bourgogne (France). The method proposed by AQ-ENIB implemented a model made by a recursive dense encoder followed by a non-dense decoder. A dense encoder is chosen to enable efficient feature reuse, facilitating training convergence. The dense encoder consists of 5 dense blocks, each consisting of 6 dense layers followed by a transition layer. Each dense layer consists of 2 convolutional layers with batch normalization (BN) and ReLU activation functions. The first convolutional layer uses 1×1 kernels, while the second uses 3×3 kernels. The transition layers consist of a BN layer, a 1×1 convolutional layer, and a 2×2 average pooling layer. The transition layer helps to reduce feature-map size. The dense blocks in the encoder have an increasing number of feature maps at each encoder stage. The model is trained using 5-fold cross-validation. To compute the final prediction, test time augmentation (TTA) is performed. This means that the model is fed with raw images and their augmented versions (using flipping and rotation with different angles). The model predicts, for each input, a segmentation mask. All the segmentation masks are ensembled using maximum majority voting.

The recursive dense architecture proposed by AQ-ENIB enables improved feature learning on the small training dataset, attenuating the chance of overfitting. Test time augmentation allows the team to increase the variability of the test set. A graphical schema of the method has been provided in Fig. 9(a)

4.2. BioPolimi

The team BioPolmini from Politecnico di Milano (Italy) are Chiara Lena, Ilaria Anita Cintorrino, Gaia Romana De Paolis and Jessica Biagioli. The model proposed by BioPolimi has a *ResNet50* (He et al., 2016) backbone followed by the *U-Net* (Ronneberger et al., 2015) decoder for segmentation. The model is trained for 700 epochs with 6-fold cross-validation, using learning rate and batch size of 10^{-3} and 32, respectively. To be consistent with the FetReg Challenge baseline, training images are resized to 448×448 pixels. Data augmentation, consisting of random crop with size 256×256 pixels, random rotation (in range $(-45°, +45°)$), horizontal and vertical flip and random variation in brightness (in range $(-20\%, +20\%)$), is applied to the training data. During inference, testing images are cropped in patches of dimension 256×256 pixels. The final prediction is obtained by overlapping the prediction obtained for each patch with a stride equal to 8.

BioPolimi enhances the baseline architecture by incorporating handcrafted features to address the issue of low contrast. The Histogram of Oriented Gradients (HoG) is specifically combined with features from ResNet50 to strengthen the recognition of anatomical contours, thereby supplying the decoder with a spatial prior of the features. A graphical schema of the method has been provided in Fig. 9(b).

4.3. GRECHID

Team GRECHID is Daria Grechishnikova from Moscow State University (Russia). The method proposed by GRECHID consists of a *U-Net* model with *SEResNeXt50* backbone (Hu et al., 2018) trained sequentially for each class (i.e., vessels, fetus and surgical tools). The *SEResNeXt50* backbone contains *Squeeze-and-Excitation* (SE) blocks, which allow the model to weigh adaptively each channel of SE blocks. Before training, exact and near-duplicates were removed using an online software,[10] obtaining 783 unique images from the original training dataset. Multi-label stratification split is performed to allocate images into train, test, and validation sets. All the images are resized to 224×224 pixels. To improve model generalization, data augmentation is performed using horizontal and vertical flips, random rotation and flipping. The model is trained using Adam optimizer and cosine annealing with restart as a learning rate scheduler, with a loss that combines Dice and modified cross-entropy losses. The modified cross-entropy loss has additional parameters to penalize either false positives or false negatives. Training is carried out in two stages. During the first stage, the model is trained for 30 epochs with a higher learning rate of 10^{-3}, then the learning rate is lowered to 10^{-5}. Cosine annealing with restart scheduling is used until the best convergence.

A triple threshold-based post-processing is applied to the model output to remove spurious pixels.

GRECHID proposes the use of a ResNeXt encoder for feature extraction. This approach aims to address the challenges of large intra-class variability and poor image quality by providing a better representation of features. Additionally, the per-class model ensemble and triple threshold post-processing help manage the high data imbalance. A graphical schema of the method has been provided in Fig. 9(d).

4.4. OOF - overoverfitting

Team OOF are Jing Jiao, Bizhe Bai and Yanyan Qiao from Fudan University (China), University of Toronto (Canada) and MicroPort Robotics. Team OOF used *U-Net++* (Zhou et al., 2018) as the segmentation model. *EfficientNetb-0* (Tan and Le, 2019) pre-trained on the ImageNet dataset is used as *U-Net++* encoder. To tackle illumination variability, median blur and Contrast Limited Adaptive Histogram Equalization (CLAHE) are applied to the images before feeding them to the model. Data augmentation, including random rotation, flip, and elastic transform, is applied during training. Adam optimizer with an initial learning rate of 10^{-4} is used. The learning rate increases exponentially with 5 warm-up epochs.

OOF addresses the issue of low contrast in images by applying the Contrast Limited Adaptive Histogram Equalization (CLAHE) technique to enhance the visibility of vessel borders. Along with visual challenges, the team encountered moiré patterns in some images that could pose difficulties in identifying the vessels. To better learn features from a small and unbalanced dataset, various configurations of EfficientNet were used as feature extractors, combined with a U-Net++ architecture and trained using standard data augmentation techniques. After evaluating the results, the team determined that the EfficientNet-b0 configuration was the best option to submit, as deeper architectures did not result in improved performance during validation. A graphical schema of the method has been provided in Fig. 9(f).

4.5. RREB

Team RREB are Binod Bhattarai, Rebati Raman Gaire, Ronast Subedi and Eduard Vazquez from University College London (UK), NepAL Applied Mathematics and Informatics Institute for Research (Nepal)

and Redev Technology (UK). The model proposed by RREB uses U^2-Net (Qin et al., 2020) as the segmentation network. A regressor branch is added on top of each decoder layer to learn the Histogram of Oriented Gradients (HoG) at different scales. The loss L minimized during the training is defined as:

$$L = \alpha \text{CE}_{\text{seg}} + \beta \text{MSE}_{\text{HoG}} \tag{3}$$

where $\alpha = 1$, CE_{seg} is the cross-entropy loss for semantic segmentation, $\beta = 1$ and MSE_{HoG} is the mean-squared error of the HoG regressor.

All the images are resized to 448×448 pixels, and random crops of 256×256 are extracted. Random rotation between $(-45°, +45°)$, cropping at different corners and centers, and flipping are applied as data augmentation. The entire model is trained for 200000 iterations using Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a batch size of 16. The initial learning rate is set to 0.0002 and then is halved at 75000, 125000, 175000 iterations. The proposed model is validated through cross-validation.

RREB team proposes the use of U^2-Net to enhance the learning of multi-scale features in fetoscopic images. They believe that combining handcrafted features with semantic segmentation and detection can better represent the structure of interest without incurring extra costs. To achieve this, RREB's network learns HoG descriptors as an auxiliary task, by adding regression heads to U^2-Net at each scale. A graphical schema of the method has been provided in Fig. 9(c).

4.6. SANO

Team SANO from Sano Center for Computational Medicine (Poland) are Szymon Płotka, Aneta Lisowska and Arkadiusz Sitek. This is the only team that participated in both tasks.

Segmentation. The model proposed by SANO is a *Feature Pyramid Network* (FPN) (Lin et al., 2017) that uses *ResNet-152* (He et al., 2016) with pre-trained weights as the backbone. The first convolutional layer has a 3-input channel, $n = 64$ feature maps, 7×7 kernel with stride = 2, and padding = 3. The following three convolutional blocks have $2n$, $4n$ and $32n$ feature maps. Our bottleneck consists of three convolutional blocks with BN. During training, the images are resized to 448×448 pixels and the following augmentations are applied:

- Color jitter (brightness = [0.8, 1.2], contrast = [0.8, 1.2], saturation = [0.8, 1.2], and hue = [−0.1, 0.1])
- Random affine transformation (rotation = [−90, 90], translation = [0.2, 0.2], scale = [1, 2], shear = [−10, 10])
- horizontal and vertical flip.

The overall framework is trained with cross-entropy loss using a batch size of 4, Adam as optimizer with an initial learning rate of 10^{-4}, weight decay and step learning rate by 0.1, and cross-entropy loss. Validation is performed with 6-fold cross-validation.

SANO propose to use a deeper feature encoder *ResNet-152*, to increase the number of features extracted, on top of a FPN architecture to tackle image complexity and improve segmentation performance. A graphical schema of the strategy proposed by SANO team for Task 1 is shown in Fig. 9(e).

Registration. The algorithm uses the channel corresponding to the placental vessel (PV) from the segmentation network and the original RGB images. The algorithm only models translation with the precision of 1 pixel. If frames are indexed by $i = 1, \ldots, t, \ldots, T$, the algorithm finds $T - 1$ translations between neighboring frames. To compute the placenta vasculature (PV) image, softmax is applied to the raw output of the segmentation. The PV channel is extracted and multiplied by 255. A mask of non-zero pixels is computed from the raw image and applied to the PV image. The homography is then computed in two steps: The shift between PV images t and $t + 1$ is computed using masked Fast Fourier Transform. Then, the rotation matrix between t and the shifted $t + 1$ image $T + 1_s$ is computed by minimizing the mean square error.

---

[10] https://github.com/idealo/imagededup

Table 3
Results of segmentation on the test set for Task 1 by training the baseline on videos only from one center. Each center ID is also indicated (I - UCLH, II - IGG) for performance comparison between the two centers.

Train dataset	Video010	Video012	Video015	Video020	Video024	Video025	Overall mIoU
Center ID	II	II	I	I	II	I	
I+II	**0.5750**	**0.4122**	**0.6923**	**0.6757**	**0.5514**	**0.7045**	**0.6763**
I	0.0109	0.0092	0.1012	0.0754	0.0056	0.2180	0.1102
II	0.1968	0.2630	0.1525	0.1562	0.3545	0.1907	0.1761
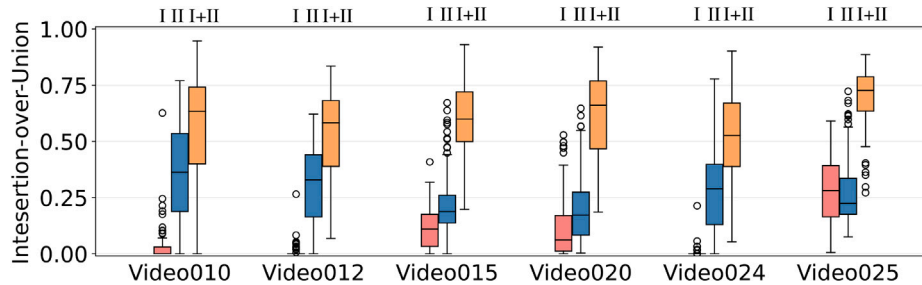
Fig. 10. Qualitative comparison showing results for baseline model when trained on single center data and multi-center data. *mIoU* over each test video for the baseline model trained with data from one center (I - UCLH, II - IGG). Bar colors from left to right indicate Centre I, II and I+II results.

4.7. Baseline

As the baseline model, we trained a *U-Net* (Ronneberger et al., 2015) with *ResNet50* (He et al., 2016) backbone as described in Bano et al. (2020a). Softmax activation is used at the final layer. Cross-entropy loss is computed and back-propagated during training. Before training, the images are first resized to 448×448 pixels. To perform data augmentation, at each iteration step, a patch of 256×256 pixels is extracted at a random position in the image. Each of the extracted patches is augmented by applying a random rotation in the range $(-45°, +45°)$, horizontal and vertical flip, scaling with a factor in the range of $(-20\%, +20\%)$ and random variation in brightness $(-20\%, +20\%)$ and contrast $(-10\%, +10\%)$. Segmentation results are obtained by inference using 448×446 pixels resized input image. The baseline model is trained for 300 epochs on the training dataset. We create 6 folds, where each fold contains 3 procedures, to preserve as much variability as possible while keeping the number of samples in each fold approximately balanced. The final model is trained on the entire dataset, splitting videos into 80% for training and 20% for validation. The data is distributed to represent the same amount of variability in both subsets. The baseline model was evaluated in Bano et al. (2021) on the training dataset before the release of this challenge. For completeness, the evaluation results from Bano et al. (2021) are presented in Table 6 and discussed in Section 5.2.1.

5. Quantitative and qualitative evaluation results

5.1. Data variability contribution

To assess data variability contribution from the multi-center dataset, we compute the performance of our baseline model when trained on data from one surgical center and tested on data from the other one. Table 3 shows the *mIoU* over each of the 6 test video samples and the overall *mIoU* over all videos with the baseline model trained on the dataset from a single center. Fig. 10 shows the qualitative comparison of mean performance over each test video for the baseline model trained with data from only one center. When training the model on data from Center I, the baseline performance on all test videos is generally lower compared to the one trained on data from Center II, except for Video025, which obtained an average *mIoU* of 0.1102 and 0.1761 respectively.

The difference in baseline model performance is mainly due to the variability and size of the dataset. In Center I, the images are

of higher quality and have well-visible structures. Although this is beneficial for clinicians, it needs to provide more information for the model learning process, which may lead to overfitting and poor segmentation performance. In contrast, data from Center II is more diverse, with various cases treated (e.g., different placenta positions and gestational weeks) and various imaging setups (e.g., straight or 30-degree fetoscope, brightness, FoV size). The increased image variability from these factors enables the model to generalize better to test images. Another crucial factor is that dividing the two datasets reduces the training set to about 900 images.

It can also be observed that when trained on individual center data, the model is not generalizable on the other center data due to data variability. However, combining the datasets (I+II) enhances the baseline model performance (average *mIoU* of 0.6763) and generalization capabilities, as it introduces a more extensive collection of images with higher variability.

5.2. Placental scene segmentation task

We perform both quantitative and qualitative comparisons to evaluate the performance of the submitted placental scene segmentation methods. Table 4 shows the *mIoU* for each team individually over each of the 6 test video samples, the overall *mIoU* overall videos and inference time (including preprocessing and postprocessing time) per frame. Table 5 presents the *mIoU* per class per frame where the overall *mIoU* computed per frame differs from the challenge metric that computes an aggregate *mIoU* for all frames. To test the rank stability, the total number of times a team is ranked 1st on a video is also reported. Fig. 11 shows the qualitative comparison of each team on each video. The qualitative results for the placental scene segmentation task are presented in Fig. 14.

5.2.1. Baseline k-fold cross validation

Prior to releasing the challenge dataset, we evaluated the Baseline (Section 4.7) on the training data, providing a benchmark for participants. Table 6 presents the cross-validation results, detailing both per-class and aggregated *mIoU* for each fold and individual videos on the training data. We employed a k-fold cross-validation approach with $k = 6$ to mitigate selection bias in model evaluation. The dataset composition was patient-centric, we aimed to maintain uniformity in the size of validation and training datasets, ensuring diverse data representation in each fold. Given the multicentric nature of our dataset, each fold included patients from all centers to ensure representation.

Table 4
Performance of participating methods for Task 1 (segmentation) on the test dataset reported using the aggregated IoU and overall *mIoU* metrics as utilized in the challenge evaluation. Inference time in milliseconds (ms) per frame for each team is also reported. Additionally, each center ID is indicated (I - UCLH, II - IGG) for performance comparison between the two centers.

Team name	Video010	Video012	Video015	Video020	Video024	Video025	Overall mIoU	# Video won	Inference time
Center ID	II	II	I	I	II	I			(ms)
AQ-ENIB	0.5611	0.2745	0.4855	0.4848	0.3342	0.6414	0.5503	0	77.67
Baseline	**0.5750**	**0.4122**	**0.6923**	**0.6757**	0.5514	0.7045	**0.6763**	4	34.41
BioPolimi	0.3891	0.2806	0.2718	0.2606	0.3666	0.3943	0.3443	0	132.22 s
GRECHID	0.4768	0.3792	0.5884	0.5744	0.3097	0.6534	0.5865	0	**33.39**
OOF	0.1874	0.1547	0.2745	0.2074	0.0872	0.3724	0.2526	0	52.40
RREB	0.5449	0.3765	0.6823	0.6191	**0.6443**	**0.7585**	0.6411	2	38.51
SANO	0.4682	0.3277	0.5201	0.5863	0.4132	0.6609	0.5741	0	131.53
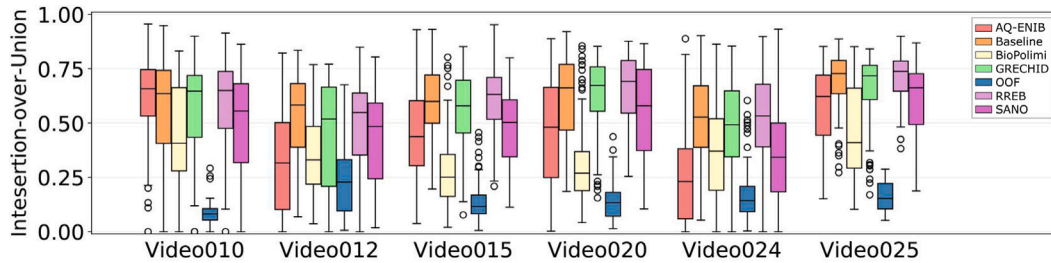
Fig. 11. Method comparison showing boxplot for frame-level *IoU* for each team on each video. Bar colors from left to right indicate teams in alphabetical order.

Table 5
Performance of participating methods for Task 1 (segmentation) per class and per frame. *mIoU* (frame-level) computes *mIoU* of each frame individually and averages all results. These results differ from the challenge metric that computes an aggregate *mIoU* for all frames.

Team name	IoU (vessel)	IoU (tool)	IoU (fetus)	mIoU (frame-level)
AQ-ENIB	0.4158	0.5038	0.2890	0.4508
Baseline	**0.5817**	0.5669	0.3824	**0.6019**
BioPolimi	0.4748	0.2612	0.1192	0.3757
GRECHID	0.5557	0.5093	0.3342	0.5617
OOF	0.2814	0.1979	0.0249	0.1833
RREB	0.5621	**0.6335**	**0.5178**	0.6005
SANO	0.4752	0.4561	0.3478	0.4915

For instance, Folds 1–3 comprised 8 patients from Center I and 7 from Center II for training, with validation on 1 patient from Center I and 2 from Center II. Folds 4–6 followed a similar pattern with reversed Center distributions.

From Fig. 12, it can be observed that overall vessel segmentation gave promising results. In challenging cases, such as when the laser glow was extremely strong (Video023), the vessels were not segmented properly. Another issue was found in the presence of vessels with different morphology and contrast with respect to the training set (e.g., Video003) that led to inaccurate vessel segmentation.

Dataset class imbalance, as discussed in Section 3.1, posed a significant challenge in identifying tools (*mIoU* 0.7637) and the fetus (*mIoU* 0.7522). In certain videos (Video001, Video003, Video005, Video008, Video023), the fetus was entirely missed in the scene. Variations in fetal shading also caused segmentation inconsistencies, as observed in Videos 011, 014, and 018. Although tools were consistently identified across all videos, their segmentation lacked precision, likely due to their regular structure.

Overall, the baseline demonstrated stable performance across all folds, albeit with noted limitations in specific challenging scenarios.

5.2.2. Challenge results using aggregated mIoU

The participating teams were evaluated using the aggregated *mIoU* for all test frames. Among the challenge participants, the best performing approach is that of RREB, which achieved an overall *mIoU* of 0.6411. RREB obtained the best performance for all videos, but Video010 and Video012, where AQ-ENIB and GRECHID are the best,

respectively. RREB performed the best among participants for all three classes, with median IoU for vessel and tool that overcame 60%. However, RREB obtained poor results for fetus class, with a median IoU lower than 40% with a large dispersion among images. As shown in Fig. 14(c) and (d), RREB meets challenges in the presence of fetus and tool. In the first case, RREB does not segment the fetus, while in the second the tool is segmented as fetus.

GRECHID scored second among all the participants, with a *mIoU* of 0.5865. As for RREB, GRECHID grants the best and lowest performance for tools and vessels, respectively. Fig. 14(b) and (f) show that GRECHID wrongly identifies and segments the fetus when it is not present in the FoV, while in Fig. 14(c), where the fetus is present, GRECHID does not segment it.

With an overall *mIoU* of 0.5741, SANO scored third, with the best performance achieved for vessels. SANO shows high variability in the IoU computed among frames for both fetus and tool. Despite the generalized good visual performance among videos, SANO tends to underestimate the areas.

AQ-ENIB obtained an overall *mIoU* of 0.5503 with the least performance obtained with fetus segmentation. Despite the good performance for vessel segmentation, vessel area is often underestimated as shown in Fig. 14(b), (e) and (f).

BioPolimi and OOF show the least performance with an *mIoU* of 0.3443 and 0.2526, respectively. OOF also faced challenges in images where one single vessel is present in the FoV, as shown in Fig. 14(b). Despite the low overall performance of BioPolimi, especially in tool and fetus segmentation, vessels are correctly segmented when visible and continuous (i.e., particles or specularities do not interrupt vessels surface), as shown in Fig. 14(d).

The baseline method is the best-performing method, achieving an overall *mIoU* of 0.6763, overcoming the performance of the challenge participants for all videos but Video024 and Video025 where RREB is the best-performing method.

When comparing the inference time that includes preprocessing and postprocessing, using A100 (40 GB) with EPYC 7452 (16 Cores) processor + 200 GB RAM workstation, GRECHID (33.39 ms) is slightly faster than the baseline (34.41 ms). BioPolimi is the slowest with an inference time of 132.22 *s* since they performed sliding window operation in preprocessing with an inefficient implementation.

Table 6
Results of the K-Fold cross-validation for the baseline on Task 1 training dataset (as reported in our earlier data analysis paper (Bano et al., 2021), included here for completeness) Mean *IoU* for each class over each video and, in the last row, the average mean *IoU* per class are reported. Key: BG-background.

| Video | Class | | | | Overall | Fold | Images | Class | | | | Overall |
	BG	Vessel	Tool	Fetus	per video		per fold	BG	Vessel	Tool	Fetus	per fold
Video001	0.83	0.85	0.69	0.74	0.64							
Video006	0.67	0.67	0.74	0.76	0.58	1	352	0.80	0.83	0.64	0.74	0.61
Video016	0.80	0.83	0.64	0.74	0.60							
Video002	0.78	0.79	0.80	0.53	0.56							
Video011	0.75	0.72	0.73	0.83	0.64	2	353	0.80	0.81	0.83	0.78	0.69
Video018	0.80	0.81	0.83	0.78	0.71							
Video004	0.80	0.80	0.72	0.80	0.66							
Video019	0.81	0.81	0.64	0.85	0.65	3	349	0.76	0.78	0.79	0.55	0.65
Video023	0.76	0.78	0.79	0.55	0.56							
Video003	0.79	0.81	0.72	0.79	0.66							
Video005	0.71	0.77	0.79	0.56	0.56	4	327	0.82	0.82	0.80	0.93	0.66
Video014	0.82	0.82	0.80	0.93	0.78							
Video007	0.78	0.77	0.84	0.72	0.66							
Video008	0.78	0.76	0.75	0.85	0.68	5	350	0.78	0.81	0.85	0.54	0.67
Video022	0.78	0.81	0.85	0.54	0.60							
Video009	0.80	0.80	0.80	0.73	0.66							
Video013	0.72	0.77	0.75	0.50	0.50	6	329	0.66	0.66	0.73	0.57	0.58
Video017	0.66	0.66	0.73	0.57	0.48							
per class	0.78	0.79	0.76	0.75								
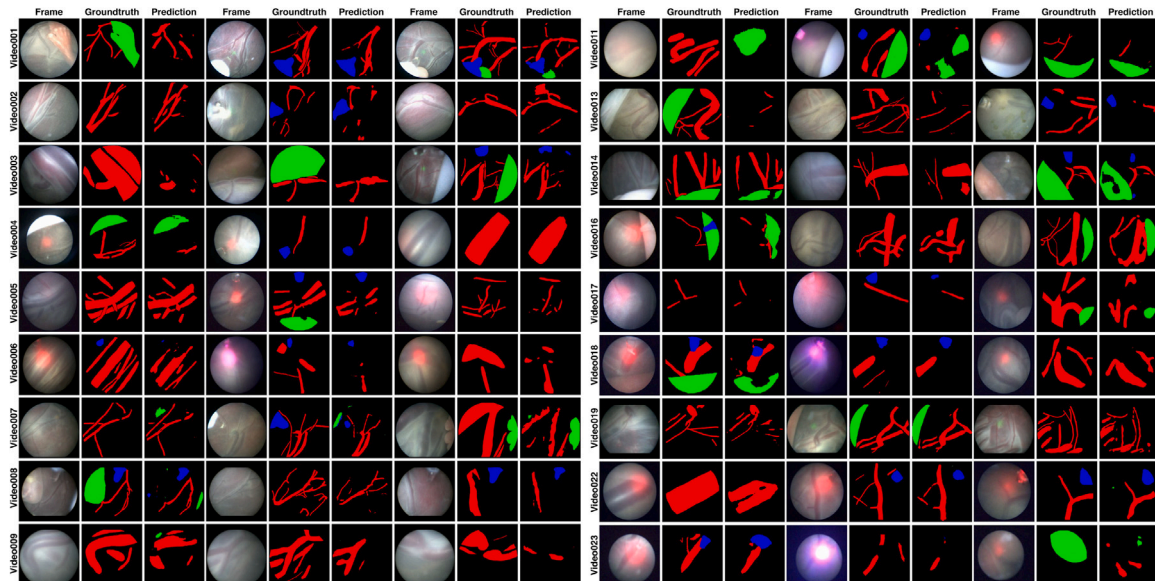
Fig. 12. Sample images from the K-Fold Cross-Validation (from Bano et al., 2021) along with the segmentation annotations (Groundtruth) and Baseline segmentation output (Prediction) for Video001, 002, 003, 004, 005, 006, 007, 008 and 009 videos. Background (black), vessel (red), tool (blue) and fetus (green) labels are shown.

5.2.3. Comparison using IoU per class per frame

When comparing IoU metrics per class or per frame, the baseline is comparable to the RREB (see Table 5). The baseline is the best at segmenting the dominant class (vessel), however, RREB is better at segmenting the less frequent classes (fetus, tool). From this perspective, the baseline is preferable as a means to achieve vessel-based registration/mosaicking (this is what the method was designed for in Bano et al., 2020a), however, RREB would overall be more reliable in aiding to detect relevant surgical events (tool usage, occlusions) which would be relevant for the event detection work in Bano et al. (2020c). If we compute *mIoU* per frame (Fig. 11), rather than aggregating all frames, we also see RREB outperforming the baseline, further supporting the idea that this method is better for event-related tasks. From this frame-level perspective (Fig. 11), we also highlight that the baseline only comes in 4th for the darkest light conditions (Video010), which may indicate some lack of robustness to significant shifts in light conditions.

AQ-ENIB, RREB, and GRECHID all outperform the baseline in these conditions.

5.2.4. Failure cases analysis

To get insight into different methods under comparison, we visualize samples where each of these methods failed, as shown in Fig. 13.

While the AQ-ENIB model shows robustness under varying light conditions, extreme low light scenarios present a significant challenge, as highlighted in Fig. 13(a) (*mIoU* 0.1896). In Fig. 13(b) (*mIoU* 0.1736), only the laser tip is correctly identified, while a substantial portion of the exposed placenta is erroneously classified as the fetus. This confusion can potentially be attributed to the inherent similarities between the placenta and the fetus under specific visual conditions, complicating their distinction. Furthermore, even when blood vessels are clearly visible, the model fails to correctly recognize them. These observations imply that the model may have struggled to learn critical
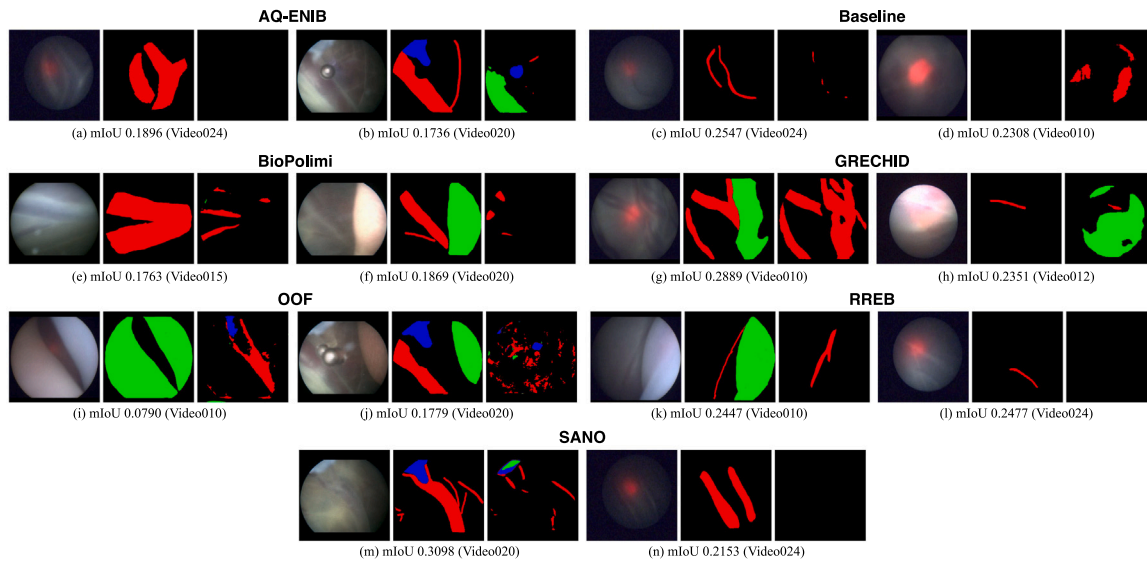
Fig. 13. Examples of failure cases from all methods. The image, the groundtruth, the video ID and the frame *mIoU* values (including background) for each sample are also reported.

features, such as identifying vessels in low-light conditions or recognizing classes that are bright but infrequent. This could likely stem from a limitation in the data augmentation strategy, which might not adequately simulate these specific conditions. In addition, as per the method description in Section 4.1, another potential issue might lie in the ensemble's majority voting mechanism. A probable disagreement between the models, potentially caused by an imbalance of examples across different folds, could adversely affect the consensus process, thereby impacting the overall performance.

As highlighted in Section 4.7, the baseline model is affected by extremely low light conditions and significant shifts in lighting within the same image. This is also reflected in Fig. 13(c) (*mIoU* 0.2547), where the model fails to detect the presence of vessels. In contrast, Fig. 13(d) (*mIoU* 0.2308) shows that shadows in the placental texture are misclassified as vessels. After examining these images, it becomes apparent that the model's performance is dependent on vessel thickness. Thin vessels, while having better contrast, are overlooked by the network, while shadows that look similar to wider vessels are misinterpreted and thus classified as vessels. This outcome suggests that the representation of vessels learned by the baseline may not adequately capture the variability in placental texture, impairing its ability to accurately discriminate between vessels and other structures. Moreover, the inherent loss of detail during the backbone feature extraction process could account for the baseline failure to identify very thin vessels. This observation underscores the need for an improved approach to preserve or enhance fine-grained features.

BioPolimi show to strongly rely on color cues to classify anatomical structures. In Fig. 13(e) (*mIoU* 0.1763) only darker gray regions, including similarly colored shadows, are segmented as vessels. In Fig. 13(f) (*mIoU* 0.1869), the fetus, though easily identifiable in this instance, is still undetected. This likely depends on the under representation of the fetus class, leading to the model's inability to effectively learn its features. This finding suggests the need for additional efforts to balance class distribution to enhance model performance. Nonetheless, a comparative analysis with the baseline on the same image suggests that the integration of HoG features might negatively impact the learning process.

The example of failure we sampled for GRECHID in Fig. 13(g) (*mIoU* 0.2889) appears visually accurate. However, GRECHID fails to correctly identify the fetus class. Upon closer inspection of the image, it is apparent that the misclassified area poses a challenging task for accurate classification, suggesting that the model cannot extract the

features required to correctly classify the fetus in this particular image. This shortfall could potentially be attributed to the limited numerosity of examples in the training dataset. Such paucity may lead to learning sparse and weak features, resulting in low confidence during per-class prediction and, consequently, erroneous classification due to the post-processing ensemble of the per-class models. Moreover, this behavior can also be responsible for unexpected results as observed in Fig. 13(h) (*mIoU* 0.2351).

Fig. 13(i) (*mIoU* 0.0790) and Fig. 13(j) (*mIoU* 0.1779) are examples of failures for OOF. In general, this method is the least performing on all the classes. We speculate that the underlying issue might reside within the preprocessing pipeline, or the sequence in which data augmentation and preprocessing are executed.

Observing Fig. 13(k) (*mIoU* 0.2447) and Fig. 13(l) (*mIoU* 0.2477) reveals that the performance of RREB is also affected by shadows, fetal parts, and particularly thin vessels. Although the regression of HoG within the training of the model aids in regularizing predictions and primarily impacts the final stage of the network, it appears that the feature extraction process remains susceptible to the same baseline limitations.

In line with other methods, except RREB, SANO fails to detect vessels under extremely low light conditions, as shown in Fig. 13(n) (*mIoU* 0.2153). Further, considering a similar behavior as AQ-ENIB, we postulate that the ensemble approach may lead the model to learn less discriminative features, as happened in Fig. 13(m) (*mIoU* 0.3098).

5.3. Registration for mosaicking task

Quantitative and qualitative results for the mosaicking task are presented in Table 7, Figs. 15 and 16.

The mosaics from the baseline and SANO methods and their 5-frame SSIM metric for every pair of images 5 frames apart in a sequence are shown in Fig. 15 for all 6 test video clips. Both methods utilized placental vessel maps for estimating the transformation between adjacent frames. From the mosaic of Video010, we observe that both methods followed different strategies for registration. SANO utilized translation registration having fewer degrees of freedom, while baseline performed affine registration of vessels having more degrees of freedom. Therefore, the baseline is able to deal with perspective warpings while SANO's approach is unable to deal with perspective changes and overestimates translation to compensate for such changes. As a result, the 5-frame SSIM for SANO is lower compared to the baseline
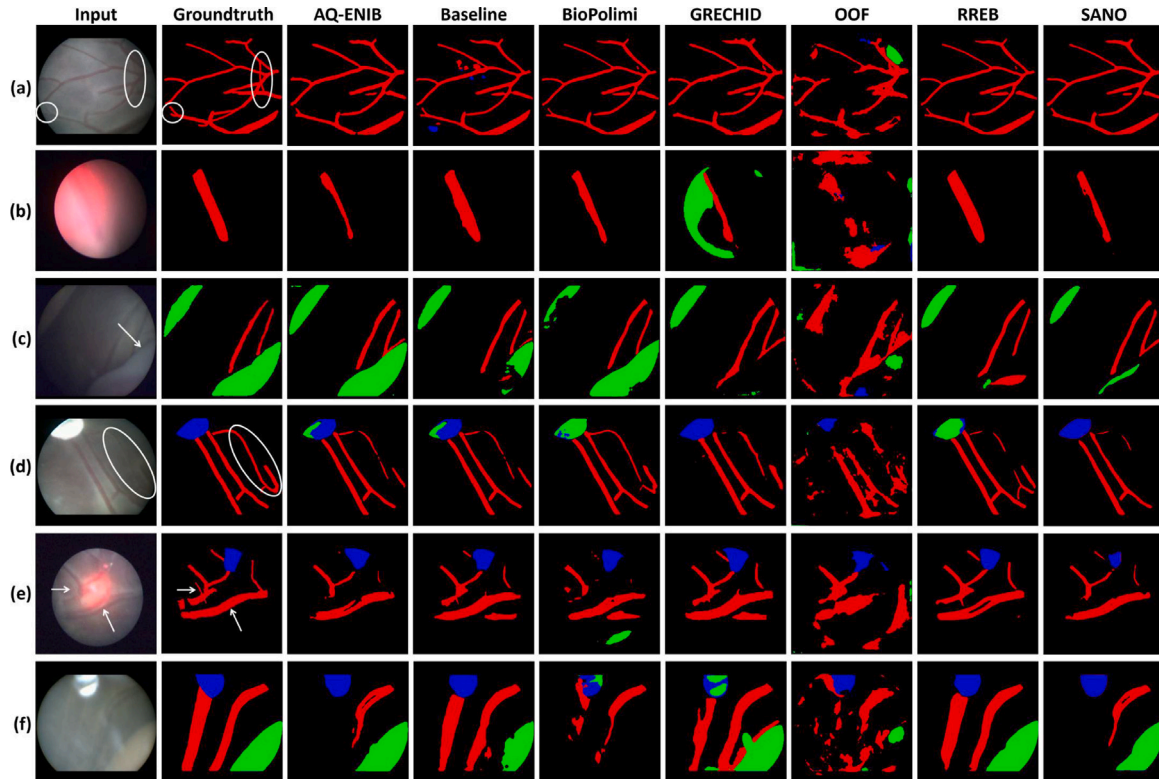
Fig. 14. Qualitative comparison of the 7 methods under analysis. Both baseline and RREB better generalize over the placental scene dataset. Baseline achieved better segmentation than RREB in (c), (d) and (e). OOF is the least performing as it failed to generalize, wrongly segmenting vessels and missing the fetus class. White markers on the input and groundtruth images indicate regions where observations can be drawn between the seven methods under comparison.

Table 7

Results of Registration for Task 2 using test video clips. The mean and median of 5-frame-SSIM metric over individual video clips is reported.

Team name		Video010	Video012	Video015	Video020	Video024	Video025	Overall	# Video won
Center ID		II	II	I	I	II	I		
Baseline (Bano et al., 2020a)	Mean	0.9048	0.9204	0.9695	0.9169	0.9336	0.9558	0.9348	5
	Median	0.9303	0.9330	0.9767	0.9301	0.9478	0.9712	0.9524	
SANO	Mean	0.8231	0.9164	0.9588	0.8276	0.9420	0.9234	0.9019	1
	Median	0.8837	0.9289	0.9746	0.8825	0.9563	0.9608	0.9434	

in Video001. On Video012, both methods struggled to generate a meaningful mosaic, but overall the baseline resulted in better 5-frame SSIM metric compared to SANO (see Table 7). Video015 is an anterior placenta case in which the placental surface is not fronto-parallel to the camera. As a result, there is a large perspective warping across different frames. SANO's approach failed in Video015 as it estimated only translation transformation. On the other hand, the baseline successfully estimated the warping through affine transformation, resulting in better 5-frame SSIM metric. Qualitatively, SANO performed better on Video020 compared to the baseline, especially in regions where vessels are visible, and the mosaic remained bounded due to only translation transformation estimation. However, the error between 5 frames is particularly large for SANO as the warpings are not accurate. Video024 and Video025 show interesting cases where in some frames there are no distinguishable structures like vessels (frame 90 in Video024 and frame 148 in Video025), hence both methods lost tracking intermediately. Quantitatively, SANO's performance is slightly better than the baseline on Video024. Through the rank stability test, we found that baseline performance was better in 5 out of 6 videos (see Table 7).

Fig. 15 shows the qualitative comparison using 1 to 5 frame SSIM metric. We observe that with increasing frame distance, the error becomes large. In the case of SANO, Video010 and Video015 result

in large drift even at 2-frame distance. As SANO used a translation transformation estimation, its error becomes very large in all videos when observing from 1 to 5 frames SSIM. The baseline followed an affine transformation estimation, as a result, its errors appear to be relatively smaller than SANO, which mainly occurred when no visible vessels were present in the scene.

6. Discussion

An accurate placental semantic segmentation is necessary for better understanding and visualization of the fetoscopic environment; as a result, this may facilitate surgeons in improved localization of the anastomoses and better surgical outcomes. However, the high intra and inter-procedure variability remains a key challenge, as only a small subset of images from each procedure were manually annotated for model training. Additionally, datasets captured from different clinical centers vary in terms of the resolution, imaging device and light source, making model generalization even more challenging.

In light of this, we conducted k-fold validation, as detailed in Section 5.2.1 . Typically, k-fold validation is crucial for smaller datasets to mitigate the risk of biased results and it is not deemed necessary when the data scale is sufficiently large, however, we considered it
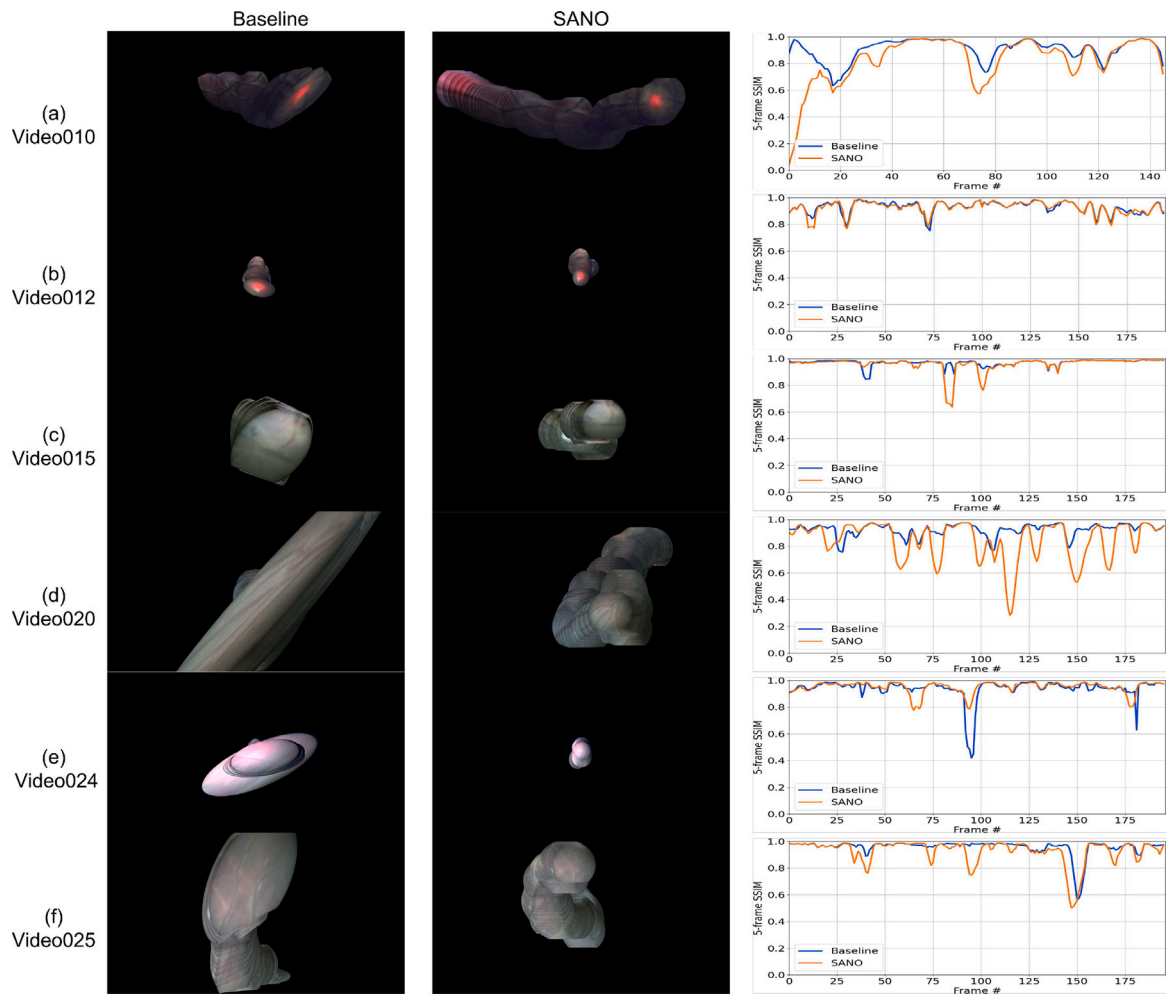
Fig. 15. Qualitative comparison of the Baseline (Bano et al., 2020a) and SANO methods showing (first column) generated mosaics from the Baseline method, (2nd column) generated mosaics from the SANO method, and (3rd column) 5-frame SSIM per frame for both methods. Baseline performance is better in all videos except Video020.
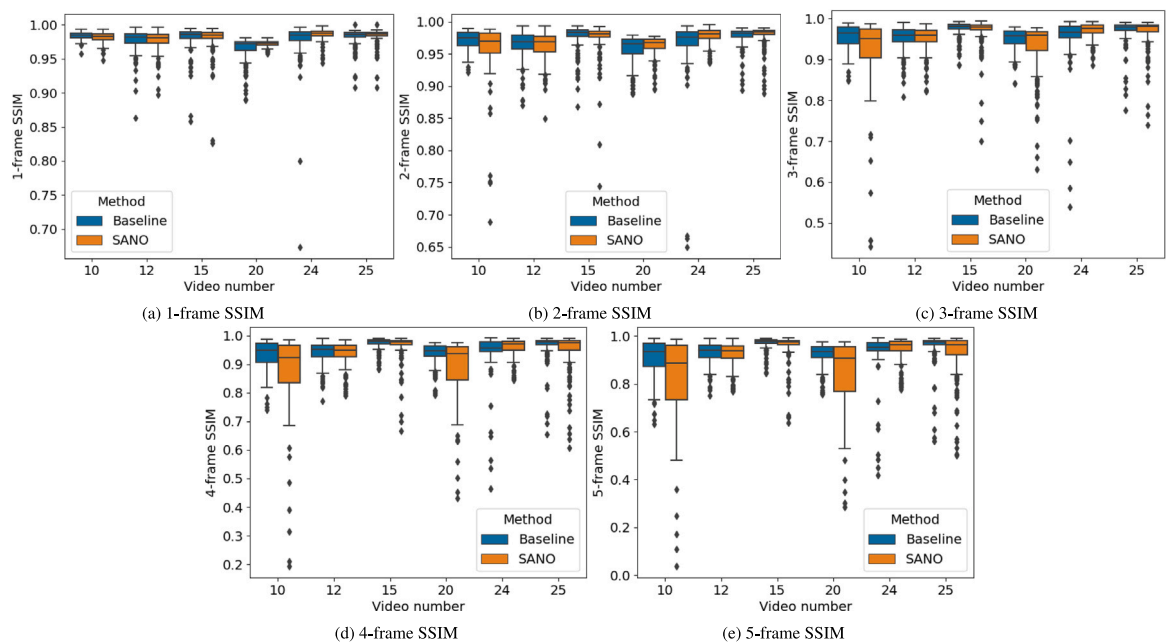

Fig. 16. Quantitative comparison of the Baseline (Bano et al., 2020a) and SANO methods using the N-frame SSIM metric.

fundamental to provide initial insights into the data and sensitize users about its challenges. Moreover, we carried out a variability analysis to assess the impact of single-center and multi-centric data on our baseline model performance in Section 5.1, highlighting the need for collecting more images from various centers. From the segmentation model results on individual 6 test videos, we observed large variability in the *mIoU* values of all methods (see Table 4). Note that Video010, 012 and 024 are from Center II and the remaining were acquired from Center I.

The performance of RREB, i.e., the winning team, may be explained by the use of a multi-task approach to segment anatomical structures while regressing the HoG. We hypothesize that training a CNN to regress multi-scale HoG from labels enhances borders and may help the network in segmenting poorly contrasted regions. RREB remains the best-performing team in the tool class. Despite HoG helping in better understanding the contours and thus producing smoother segmentation masks, this does not improve the performance with non-uniform texture, as for reflections on vessel surface (Fig. 14(e) which can cause holes in the final segmentation mask, and fetus.

The runner-up team, GRECHID, achieved the best performance in vessel segmentation close to RREB and baseline, with some issues in segmenting the fetus (Fig. 14(b,c,f)). GRECHID network architecture is rather similar to the baseline, but the adoption of per-class network configuration was chosen to achieve one-vs-all pixel classification and, thus eased data distribution learning for each class. While we cannot speculate whether this design actually improves the performance, it would be interesting to assess the data reduction impact on segmentation performance.

AQ-ENIB (average *mIoU* of 0.5503) and SANO (average *mIoU* of 0.5741) share the same segmentation strategy with only minor differences as also reflected from the comparable performance (*ΔmIoU* of 4.32%). Overall, both models perform well and have the same weakness producing no or under segmentation in case of reflections (Fig. 14(c)), small vessels (Fig. 14(d)) and poor contrast (Fig. 14(f)). Test-Time Augmentation in AQ-ENIB can provide some help in fetus segmentation but also cause false positives as in Fig. 14(d). Considering the low difference in performance, we can analyze the model's footprint and a positive aspect of AQ-ENIB is that DenseNet has lower parameter numbers (around 20 million) compared to SANO ResNet152 (around 60 million). This is also reflected in the inference time (Table 4 where AQ-ENIB took 77.67 ms and SANO took 131.53 ms to process a frame.

BioPolimi uses the same architecture of the baseline but achieves way lower performance (average *mIoU* of 0.3443). The integration of HoG features computed on the image seemed to have a negative impact on segmentation. We hypothesize that computing HoG features on the input frame does not provide a strong reference to help the network encoder to manage for low contrast, compared to HoG computed on groundtruth and multi-task as in RREB.

OOF method is the least performing (average *mIoU* of 0.2526) on all the test sets and produced several segmentation errors as shown in Fig. 14. We think that the additional preprocessing generates image with high contrast, thus polarizing the network in learning non-realistic features.

There was no single method that outperformed on all 6 test samples. This suggests that the proposed methods did not fully generalize to the dataset distributions from the two centers. Nonetheless, it is worth considering that some strategies presented by the participants are complementary and can be combined to effectively tackle some of the challenges and boost the segmentation performance.

To better model the variability in the dataset, more annotated images would be needed for supervised learning. Limited annotation problems can also be addressed through pseudo-labeling using semi-supervised learning techniques. A reliable and consistent mosaic is needed for visualizing an increased FoV image of the placental environment. The two methods under comparison relied on accurate placental vessel segmentation for mosaicking. However, during fetoscopy, the placenta regions might appear either with very thin and weak vessels or no vessels at all. A segmentation algorithm may fail in these scenarios, especially when no vessels are visible, leading to failure in consecutive frames' registration for mosaicking. This suggests that a registration algorithm should not solely rely on vessel segmentation predictions. More recent deep learning-based keypoint and matching approaches (DeTone et al., 2018; Sarlin et al., 2020; Sun et al., 2021) could be useful in improving placental frame registration for mosaicking. Some recent works (Casella et al., 2022; Bano et al., 2022) in mosaicking have already shifted interests towards exploiting learning-based keypoints and matching approaches.

7. Conclusion

Surgical data science has the potential to enhance intraoperative imaging by providing better visualization of the surgical environment with increased FoV to support the surgeon's decision during the procedure. Deep learning-based semantic segmentation algorithms can help in better understanding the fetoscopic placental scene during fetoscopy. However, large labeled datasets are required for training robust segmentation models. Through the FetReg2021 challenge, which was part of the MICCAI2021 Endoscopic vision challenge, we contributed a large-scale multi-center fetoscopy dataset containing data from 18 fetoscopy procedures for training and 6 fetoscopy procedures for testing. The test data was hidden from the challenge participants but followed a similar distribution to the training dataset. The challenge focused on solving the task of placental semantic segmentation and fetoscopy video frame registration for mosaicking. The segmentation solutions presented by the participating teams achieved promising results, though they were unable to beat the baseline method. Achieving generalizability remained an open question, and none of the methods outperformed in all test video samples. The contributed mosaicking approaches relied on accurate vessel segmentation and the presence of vessels in the fetoscopic placental view. Through the FetReg2021 challenge, we contributed a benchmark dataset for advancing the research in fetoscopic mosaicking.

CRediT authorship contribution statement

Sophia Bano: Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Alessandro Casella:** Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Francisco Vasconcelos:** Conceptualization, Formal analysis, Methodology, Writing – original draft. **Abdul Qayyum:** Methodology. **Abdesslam Benzinou:** Methodology. **Moona Mazher:** Methodology. **Fabrice Meriaudeau:** Methodology. **Chiara Lena:** Methodology. **Ilaria Anita Cintorrino:** Methodology. **Gaia Romana De Paolis:** Methodology. **Jessica Biagioli:** Methodology. **Daria Grechishnikova:** Methodology. **Jing Jiao:** Methodology. **Bizhe Bai:** Methodology. **Yanyan Qiao:** Methodology. **Binod Bhattarai:** Methodology. **Rebati Raman Gaire:** Methodology. **Ronast Subedi:** Methodology. **Eduard Vazquez:** Methodology. **Szymon Płotka:** Methodology. **Aneta Lisowska:** Methodology. **Arkadiusz Sitek:** Methodology. **George Attilakos:** Data curation. **Ruwan Wimalasundera:** Methodology. **Anna L. David:** Data curation. **Dario Paladini:** Data curation. **Jan Deprest:** Data curation. **Elena De Momi:** Supervision. **Leonardo S. Mattos:** Supervision. **Sara Moccia:** Formal analysis, Methodology, Supervision. **Danail Stoyanov:** Funding acquisition, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used has already been made publicly available.

References

Alabi, O., Bano, S., Vasconcelos, F., L. David, A., Deprest, J., Stoyanov, D., 2022. Robust fetoscopic mosaicking from deep learned flow fields. Int. J. Comput. Assist. Radiol. Surg..

Almoussa, N., Dutra, B., Lampe, B., Getreuer, P., Wittman, T., Salafia, C., Vese, L., 2011. Automated vasculature extraction from placenta images. In: Medical Imaging 2011: Image Processing, Vol. 7962. SPIE, p. 79621L.

Bano, S., Casella, A., Vasconcelos, F., Moccia, S., Attilakos, G., Wimalasundera, R., David, A.L., Paladini, D., Deprest, J., De Momi, E., et al., 2021. FetReg: Placental vessel segmentation and registration in fetoscopy challenge dataset. arXiv preprint arXiv:2106.05923.

Bano, S., Vasconcelos, F., David, A.L., Deprest, J., Stoyanov, D., 2022. Placental vessel-guided hybrid framework for fetoscopic mosaicking. Comput. Methods Biomech. Biomed. Eng.: Imaging Visual. 1–6.

Bano, S., Vasconcelos, F., Shepherd, L.M., Vander Poorten, E., Vercauteren, T., Ourselin, S., David, A.L., Deprest, J., Stoyanov, D., 2020a. Deep placental vessel segmentation for fetoscopic mosaicking. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 763–773.

Bano, S., Vasconcelos, F., Tella Amo, M., Dwyer, G., Gruijthuijsen, C., Deprest, J., Ourselin, S., Vander Poorten, E., Vercauteren, T., Stoyanov, D., 2019. Deep sequential mosaicking of fetoscopic videos. In: Lecture Notes in Computer Science. In: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 11764 LNCS, pp. 311–319.

Bano, S., Vasconcelos, F., Tella-Amo, M., Dwyer, G., Gruijthuijsen, C., Vander Poorten, E., Vercauteren, T., Ourselin, S., Deprest, J., Stoyanov, D., 2020b. Deep learning-based fetoscopic mosaicking for field-of-view expansion. Int. J. Comput. Assist. Radiol. Surg..

Bano, S., Vasconcelos, F., Vander Poorten, E., Vercauteren, T., Ourselin, S., Deprest, J., Stoyanov, D., 2020c. FetNet: A recurrent convolutional network for occlusion identification in fetoscopic videos. Int. J. Comput. Assist. Radiol. Surg. 15 (5), 791–801.

Baschat, A., Chmait, R.H., Deprest, J., Gratacós, E., Hecher, K., Kontopoulos, E., Quintero, R., Skupski, D.W., Valsky, D.V., Ville, Y., 2011. Twin-to-twin transfusion syndrome (TTTS). J. Perinat. Med. 39 (2), 107–112.

Baud, D., Windrim, R., Keunen, J., Kelly, E.N., Shah, P., Van Mieghem, T., Seaward, P.G.R., Ryan, G., 2013. Fetoscopic laser therapy for twin-twin transfusion syndrome before 17 and after 26 weeks' gestation. Am. J. Obstet. Gynecol. 208 (3), 1–197.

Bian, J., Lin, W.-Y., Matsushita, Y., Yeung, S.-K., Nguyen, T.-D., Cheng, M.-M., 2017. Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4181–4190.

Casella, A., Bano, S., Vasconcelos, F., David, A.L., Paladini, D., Deprest, J., De Momi, E., Mattos, L.S., Moccia, S., Stoyanov, D., 2022. Learning-based keypoint registration for fetoscopic mosaicking. http://dx.doi.org/10.48550/ARXIV.2207.13185, URL: https://arxiv.org/abs/2207.13185.

Casella, A., Moccia, S., Frontoni, E., Paladini, D., De Momi, E., Mattos, L.S., 2020. Inter-foetus Membrane Segmentation for TTTS Using Adversarial Networks. Ann. Biomed. Eng. 48 (2), 848–859.

Casella, A., Moccia, S., Paladini, D., Frontoni, E., Momi, E.D., Mattos, L.S., 2021. A shape-constraint adversarial framework with instance-normalized spatio-temporal features for inter-fetal membrane segmentation. Med. Image Anal. 70, 102008.

Chang, J.M., Huynh, N., Vazquez, M., Salafia, C., 2013. Vessel enhancement with multiscale and curvilinear filter matching for placenta images. In: International Conference on Systems, Signals, and Image Processing. IEEE Computer Society, pp. 125–128.

Cincotta, R., Kumar, S., 2016. Future directions in the management of twin-to-twin transfusion syndrome. Twin Res. Hum. Genet. 19 (3), 285–291.

Daga, P., Chadebecq, F., Shakir, D.I., Herrera, L.C.G., Tella, M., Dwyer, G., David, A.L., Deprest, J., Stoyanov, D., Vercauteren, T., Ourselin, S., 2016. Real-time mosaicing of fetoscopic videos using SIFT. In: Medical Imaging 2016: Image-Guided Procedures, Robotic Interventions, and Modeling, Vol. 9786. SPIE, p. 97861R.

Deprest, J.A., Flake, A.W., Gratacos, E., Ville, Y., Hecher, K., Nicolaides, K., Johnson, M.P., Luks, F.I., Adzick, N.S., Harrison, M.R., 2010. The making of fetal surgery. Prenat. Diagn. 30 (7), 653–667.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2016. Deep image homography estimation. arXiv preprint arXiv:1606.03798.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 224–236.

Eisenmann, M., Reinke, A., Weru, V., Tizabi, M.D., Isensee, F., Adler, T.J., Ali, S., Andrearczyk, V., Aubreville, M., Baid, U., et al., 2023. Why is the winner the best? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19955–19966.

Gaisser, F., Jonker, P.P., Chiba, T., 2016. Image registration for placenta reconstruction. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society, pp. 473–480.

Gaisser, F., Peeters, S.H., Lenseigne, B., Jonker, P.P., Oepkes, D., 2017. Fetoscopic panorama reconstruction: Moving from ex-vivo to in-vivo. In: Communications in Computer and Information Science. In: Communications in Computer and Information Science, vol. 723, Springer International Publishing, pp. 581–593.

Gaisser, F., Peeters, S.H., Lenseigne, B.A., Jonker, P.P., Oepkes, D., 2018. Stable image registration for in-vivo fetoscopic panorama reconstruction. J. Imaging 4 (1).

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7132–7141. http://dx.doi.org/10.1109/CVPR.2018.00745.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2462–2470.

Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., Burgard, W., 2011. G2O: A general framework for graph optimization. In: 2011 IEEE International Conference on Robotics and Automation. IEEE, pp. 3607–3613.

Lewi, L., Deprest, J., Hecher, K., 2013. The vascular anastomoses in monochorionic twin pregnancies and their clinical consequences. Am. J. Obstetrics Gynecol. 208 (1), 19–30.

Li, L., Bano, S., Deprest, J., David, A.L., Stoyanov, D., Vasconcelos, F., 2021. Globally optimal fetoscopic mosaicking based on pose graph optimisation with affine constraints. IEEE Robot. Autom. Lett. 6 (4), 7831–7838.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Lopriore, E., Middeldorp, J.M., Oepkes, D., Klumper, F.J., Walther, F.J., Vandenbussche, F.P., 2007. Residual anastomoses after fetoscopic laser surgery in twin-to-twin transfusion syndrome: Frequency, associated risks and outcome. Placenta 28 (2–3), 204–208.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., Saez-Rodriguez, J., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., 2020. BIAS: Transparent reporting of biomedical image analysis challenges. Med. Image Anal. 66, 101796. http://dx.doi.org/10.1016/j.media.2020.101796, URL: https://www.sciencedirect.com/science/article/pii/S1361841520301602.

Mair, E., Hager, G.D., Burschka, D., Suppa, M., Hirzinger, G., 2010. Adaptive and generic corner detection based on the accelerated segment test. In: European Conference on Computer Vision. Springer, pp. 183–196.

Maselli, K.M., Badillo, A., 2016. Advances in fetal surgery. Ann. Transl. Med. 4 (20).

Moccia, S., De Momi, E., El Hadji, S., Mattos, L.S., 2018. Blood vessel segmentation algorithms—review of methods, datasets and evaluation metrics. Comput. Methods Programs Biomed. 158, 71–91.

Nguyen, T., Chen, S.W., Shivakumar, S.S., Taylor, C.J., Kumar, V., 2018. Unsupervised deep homography: A fast and robust homography estimation model. IEEE Robot. Autom. Lett. 3 (3), 2346–2353.

Peter, L., Tella-Amo, M., Shakir, D.I., Attilakos, G., Wimalasundera, R., Deprest, J., Ourselin, S., Vercauteren, T., 2018. Retrieval and registration of long-range overlapping frames for scalable mosaicking of in vivo fetoscopy. Int. J. Comput. Assist. Radiol. Surg. 13 (5), 713–720.

Pratt, R., Deprest, J., Vercauteren, T., Ourselin, S., David, A.L., 2015. Computer-assisted surgical planning and intraoperative guidance in fetal surgery: A systematic review. Prenatal Diagn. 35 (12), 1159–1166.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M., 2020. U2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognit. 106, 107404.

Quintero, R.A., Ishii, K., Chmait, R.H., Bornick, P.W., Allen, M.H., Kontopoulos, E.V., 2007. Sequential selective laser photocoagulation of communicating vessels in twin-twin transfusion syndrome. J. Maternal-Fetal Neonatal Med. 20 (10), 763–768.

Reeff, M., Gerhard, F., Cattin, P., Székely, G., 2006. Mosaicing of endoscopic placenta images. In: INFORMATIK 2006 - Informatik fur Menschen, Beitrage der 36. Jahrestagung der Gesellschaft fur Informatik e.V. (GI), Vol. 1. pp. 467–474.

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.

Sadda, P., Imamoglu, M., Dombrowski, M., Papademetris, X., Bahtiyar, M.O., Onofrey, J., 2019. Deep-learned placental vessel segmentation for intraoperative video enhancement in fetoscopic surgery. Int. J. Comput. Assist. Radiol. Surg. 14 (2), 227–235.

Sadda, P., Onofrey, J.A., Bahtiyar, M.O., Papademetris, X., 2018. Better feature matching for placental panorama construction. In: Lecture Notes in Computer Science. In: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 11076 LNCS, Springer Verlag, pp. 128–137.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4938–4947.

Senat, M.-V., Deprest, J., Boulvain, M., Paupe, A., Winer, N., Ville, Y., 2004. Endoscopic laser surgery versus serial amnioreduction for severe twin-to-twin transfusion syndrome. N. Engl. J. Med. 351 (2), 136–144.

Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X., 2021. LoFTR: Detector-free local feature matching with transformers. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 8922–8931.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tella, M., Daga, P., Chadebecq, F., Thompson, S., Shakir, D.I., Dwyer, G., Wimalasundera, R., Deprest, J., Stoyanov, D., Vercauteren, T., Ourselin, S., 2016. A combined em and visual tracking probabilistic model for robust mosaicking: Application to fetoscopy. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society, pp. 524–532.

Tella-Amo, M., Peter, L., Shakir, D.I., Deprest, J., Stoyanov, D., Iglesias, J.E., Vercauteren, T., Ourselin, S., 2018. Probabilistic visual and electromagnetic data fusion for robust drift-free sequential mosaicking: Application to fetoscopy. J. Med. Imaging 5 (02), 1.

Tella-Amo, M., Peter, L., Shakir, D.I., Deprest, J., Stoyanov, D., Vercauteren, T., Ourselin, S., 2019. Pruning strategies for efficient online globally consistent mosaicking in fetoscopy. J. Med. Imaging 6 (03), 1.

Vasconcelos, F., Brandão, P., Vercauteren, T., Ourselin, S., Deprest, J., Peebles, D., Stoyanov, D., 2018. Towards computer-assisted TTTS: Laser ablation detection for workflow segmentation from fetoscopic video. Int. J. Comput. Assist. Radiol. Surg. 13 (10), 1661–1670.

Yang, L., Wang, J., Ando, T., Kubota, A., Yamashita, H., Sakuma, I., Chiba, T., Kobayashi, E., 2016. Towards scene adaptive image correspondence for placental vasculature mosaic in computer assisted fetoscopic procedures. Int. J. Med. Robot. Comput. Assist. Surg. 12, 375–386. http://dx.doi.org/10.1002/rcs.1700.

Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 3–11.