Inter-Foetus Membrane Segmentation for TTTS using Adversarial Networks

Alessandro Casella[1,2,*] Sara Moccia[2,3,*] Emanuele Frontoni[3]
Dario Paladini[4] Elena De Momi[1] Leonardo S. Mattos[2]

affiliations: [1]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

[2]Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy

[3]Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy

[4]Department of Foetal and Perinatal Medicine, Istituto Giannina Gaslini, Genoa, Italy

[*]These authors equally contributed to the work

abbreviated title: Adversarial Networks for Membrane Segmentation in TTTS

correspondence: Alessandro Casella Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy. e-mail: `alessandro.casella@polimi.it`

Abstract

Twin-to-Twin Transfusion Syndrome (TTTS) is commonly treated with minimally invasive laser surgery in fetoscopy. The inter-foetal membrane is used as a reference to find abnormal anastomoses. Membrane identification is a challenging task due to small field of view of the camera, presence of amniotic liquid, foetus movement, illumination changes and noise. This paper aims at providing automatic and fast membrane segmentation in fetoscopic images. We implemented an adversarial network consisting of two Fully-Convolutional Neural Networks (FCNNs). The former (the *segmentor*) is a segmentation network inspired by U-Net and integrated with residual blocks, whereas the latter acts as *critic* and is made only of the encoding path of the *segmentor*.

A dataset of 900 images acquired in 6 surgical cases was collected and labelled to validate the proposed approach.

The adversarial networks achieved a median Dice similarity coefficient of 91.91% with Inter-Quartile Range (IQR) of 4.63%, overcoming approaches based on U-Net (82.98% - IQR : 14.41%) and U-Net with residual blocks (86.13% - IQR : 13.63%). Results proved that the proposed architecture could be a valuable and robust solution to assist surgeons in providing membrane identification while performing fetoscopic surgery.

keywords: Deep learning adversarial networks fetoscopy intraoperative-image segmentation

1 Introduction

Twin-to-Twin Transfusion Syndrome (TTTS) is a pathology with deadly consequences that occurs in the 15% of monochorionic pregnancies (75% of twin homozygous pregnancies)[3]. The aetiology of TTTS is correlated to the anomalous presence of unidirectional inter-placental anastomoses, which cause an imbalance in the blood flow between the foetuses. The risk of perinatal mortality of one or both foetuses can exceed 90% without any treatment, with an incidence of physical or neurological complications in the 50% of the surviving foetuses[31,32]. Fetoscopic Minimally Invasive Surgery (MIS) has largely decreased maternal and foetal morbidity or mortality[35], becoming the recommended technique for the first-line treatment of TTTS. The surgery consists of a direct interruption of anastomoses that are responsible for TTTS via laser photo-coagulation. The procedure is performed using a fetoscope and a fibre laser for ablation, which is driven through a small working channel in the fetoscope. Fetoscopic MIS is performed in a selective way (*i.e.*, only communicating vessels among the foetuses should be coagulated, preserving all the others).

The selection of the vessels to be treated relies on the location of abnormal vascular formations, at the small branches of normal blood vessels. The first step for the surgeon, to find these abnormal vessels, is a visual inspection of the entire foetal environment (most of the time randomly moving the fetoscope) until he/she locates the inter-foetus membrane, which is used as a reference for the navigation of the vascular network. However, as described in the clinical literature[29,37,25], the identification of the membrane is a challenging task since the surgeon's ability to maintain orientation is hampered by several factors: (i) there is a limited Field of View (FoV) on the surgical scene, constraining the surgeon to view only a small portion of the placental surface[17,8]; (ii) the fetoscope often goes out of focus due to dynamic changes in the foetal environment; (iii) foetuses can unpredictably move and often occlude the camera FoV, hiding the membrane; (iv) the surgical environment is immersed to the amniotic fluid, which is turbid. Additional challenges include large variability in the illumination level, which ranges from intense illumination (causing specular reflections) to dim lighting conditions. Some visual examples to highlight the challenges in identifying the membrane are shown in Fig. 1.

[Figure 1 about here.]

Computer-assisted solutions may be used to identify and segment the membrane in order to support surgeons during TTTS surgery. Such solutions may tackle the complexity of intraoperative images through learning-based segmentation, as highlighted by a review on medical-image segmentation[20]. Recently, researchers in other medical fields (*e.g.*, skin lesion segmentation) have shown the potentiality of adversarial training for further increasing segmentation performance[42].

Inspired by such considerations, this paper proposes a framework based on adversarial networks for the segmentation of inter-foetal membrane from in-vivo fetoscopy images acquired during TTTS MIS. The additional L_1 loss term computed by the *critic* network realise a multi-scale features analysis during training preserving high-level features connected to macro appearance. The proposed framework aims at supporting clinicians by automatically detecting the membrane on the fetoscope video stream and highlighting its borders. The integration of this framework in a smart fetoscope system may lead to a decrease in surgeon mental workload during the surgery, possibly reducing the duration of the intervention.

The paper is organised as follows: Sec. 1.1 surveys intraoperative medical image segmentation strategies, with a focus on learning algorithms and adversarial training; Sec. 2 presents the proposed segmentation framework and describes the experimental protocol for validating it. The obtained results are presented in Sec. 4 and discussed in Sec. 5. Conclusive remarks are presented in Sec. 5.1.

1.1 Related work on intraoperative tissue segmentation

[Figure 2 about here.]

In the past, intraoperative tissue segmentation approaches mostly dealt with filtering or deformable models. For instance, in[24] steerable filters and textural descriptor were used for gastric-lesion segmentation in capsule endoscopy. To tackle some of the limitations of these approaches (*e.g.*, needs for parameter tuning and long processing time), supervised machine learning algorithms have been proposed to provide fast and accurate segmentation[36]. Supervised machine learning addresses the segmentation as a two-step problem: first, image features are extracted (*e.g.* intensity and textural features), then such features are classified (*e.g.* with support vector machines (SVMs) and decision trees). Applications include uterus segmentation from endoscopic images, where Gabor filtering is used for feature extraction[5]; other examples are segmentation of Fallopian tubes from endoscopic images, obtained using tube-specific geometrical features[30] and segmentation of abdominal organs with textural features[27]. More recently, Fully-Convolutional Neural Networks (FCNNs) have emerged as a powerful supervised-learning tool for many visual recognition tasks such as segmentation of complex scenes from in-vivo endoscopic images. FCNNs allow for accurate segmentation when the large annotated training datasets are available. FCNN first layers are responsible for automatic image-feature extraction, while the last layer classifies the features and provides the segmentation mask[14]. After their first implementation[21], FCNNs have been deployed in a variety of architectures, such as U-Net[33], SegNet[2] and residual architectures[9] (mainly based on the residual blocks proposed in ResNet architectures[16]). In[39,13], SegNet is used for polyp segmentation. In both cases, SegNet is pre-trained on the ImageNet dataset[7] and then fine-tuned to address the segmentation task. Similarly, in[4] several state-of-the-art FCNNs (*i.e.*, AlexNet, GoogleNet, VVG and residual network) are pre-trained on the PASCAL VOC[11] and fine tuned for polyp segmentation.

FCNNs are trained by minimising an error metric between the ground-truth and the predicted segmentation. This error metric is commonly computed by measuring the overlap or by comparing the pixel-probability distributions between the ground-truth and the predicted segmentation[14]. Following a different perspective, researchers have recently investigated the use of adversarial training. Adversarial training was initially proposed by Goodfellow et al.[15] as a generative framework for natural images (*i.e.*, in the context of Generative Adversarial Networks (GANs)) made of a generator and a discriminator network[15]. This framework consists of two networks that are trained one against the other by letting the two networks to contribute to the same loss. The segmentation loss is made of two terms: a segmentation overlap measure and a features equality measure. The addition of this second term in the loss function, as shown in[42,41,22], is useful to reach an increasing training robustness in segmentation task. Adversarial training in the context of image segmentation has been tested for natural images (*e.g.*, Pascal VOC dataset)[22] and skin lesions from dermoscopy images[45]. In fetoscopic images tissues may look very different and partially visible. The high level of noise, the blurred vision due to amniotic fluid with suspended particulate matter, the wide range of illumination and the variation of the fetoscope pose to the recorded tissues further increase the complexity of the structures segmentation.

With the aim of dealing with these issues, the present work investigates the use of a new adversarial framework for the segmentation of inter-foetus membrane. It is also compared to the state-of-the-art architectures, introducing a feature-based adversarial loss function as measure of segmentation confidence.

2 Materials and Methods

This section describes the proposed adversarial framework for fetoscopic membrane segmentation (Sec. 2.1) and the strategy to train it (Sec. 2.2).

[Table 1 about here.]

[Table 2 about here.]

2.1 SAN architecture

Similarly to the original generative framework, the Segmentation Adversarial Network (SAN) implemented in this work consists of two networks, where the generator (which here acts as segmentation network (S)) and the discriminator (here, the *critic* network (C)) are alternately trained to minimise and maximise an objective function, respectively. The latter will be defined in Sec. 2.2. Figure 2 shows the overall diagram of the framework.

The architecture of the *segmentor* network S (Table 1) is based on the U-Net[33] encoder-decoder structure, a fully convolutional network that naturally performs overlap-tile extraction, preserving spatial connectivity between tiles while speeding up network training. Each encoding layer is composed by a convolutional layer with 2 x 2 stride, with batch normalization and Leaky Rectified Linear Unit (ReLU)[23] activation. Strided convolution layer is used to avoid checkerboard artefacts[28]. The batch normalisation (BN) layer is applied to normalise the output of each layer, allowing for a larger learning rate that accelerates the training procedure[18]. Considering improvements in network training speed and performances reported in the literature[40], Leaky ReLU is chosen over the standard one. The leaky ReLU is defined as:

$$l_{relu}(x) = \begin{cases} \alpha x & x < 0 \\ x & x \geq 0 \end{cases} \tag{1}$$

where x is the network-layer output to be activated, and the coefficient α is introduced to avoid the problem of "*dying-ReLUs*". The dying ReLU is a vanishing gradient problem that occurs when ReLU neurons become inactive and only output 0 for any input. The output of every layer in the encoder path is passed to a residual block, following the approach of ResNet architectures[16].

A residual block featuring short skip connections is implemented to reduce the memory load of the network and to deal with the vanishing gradient issue. It is made of: (i) a convolution layer with 1 x 1 kernel and 1 x 1 stride at the ends to double (halve) the number of image channels, (ii) a convolution layer with 3 x 3 kernel size and 1 x 1 stride process the input maintaining the original size.

The decoder path mirrors the encoder path. Each step of the decoder is made of a strided deconvolution layer with BN and a ReLU activation layer followed by a residual block. The last step of the encoding path is made of a convolution layer with ReLU activation. To obtain a probability map for pixel classification, the output of the last encoder step is activated by a sigmoid function.

The architecture of the C network (Table 2) contains the same encoding path of the *segmentor* network for feature extractions. It takes for an input that of the fetoscopic image, which is masked by a binary mask. In particular, the first mask is the S prediction and the second one is the ground-truth, so that two feature vectors are generated.

2.2 Training strategy

In the SAN framework there are two loss functions, one for the *segmentor* S and one for the *critic* C network. The *segmentor* loss ($L_{S_{SAN}}$) (Eq. 5) in our framework, consists of two terms: a common overlap metrics based on Dice similarity coefficient (L_{DSC}) and an additional term derived from the *critic* (L_{L1}). The L_{DSC} is defined as:

$$L_{DSC} = -DSC = -\frac{2TP}{2TP + FN + FP} \tag{2}$$

where TP is the number of membrane pixels correctly identified, whereas FP and FN are the background and membrane pixels that are misclassified.

L_{L1} considers differences between feature vectors extracted from the ground-truth and the predicted image:

$$L_{L1} = \frac{1}{N} \sum_{n=1}^{N} L_1[f(x_n \cdot S(x_n)), f(x_n \cdot y_n)] \tag{3}$$

where $(x_n \cdot S(x_n))$ is the input image masked (pixel-wise multiplication) by the S prediction, $(x_n \cdot y_n)$ is the input image masked by the ground-truth mask, f represents the feature vector extracted by the *critic* network and L_1 is the Mean Absolute Error (MAE) defined as:

$$L_1(f(x_s, x)) = \frac{\sum_{i=1}^{M} |f_i(x_s) - f_i(y)|}{M} \tag{4}$$

where M is the total number of convolutional layers in C, $f_i(x_s)$ is the feature map for the input image masked by the predicted segmentation mask and $f_i(x)$ for the input image masked by the ground-truth mask (x) at the i-th layer. The computation of the L_{L1} loss term is based on high-level features differences between the predicted and the true segmentation extracted from the *critic* network. L_{L1} loss function force the *segmentor* to learn both global and local features that capture long- and short-range spatial relationships between pixels.

Then $L_{S_{SAN}}$ is defined as

$$L_{S_{SAN}} = \min_{S}(L_{DSC}(S) + L_{L1}(S, C)) \tag{5}$$

3 Experimental protocol

3.1 Dataset

In order to train the proposed framework, we built a new dataset in collaboration with the Department of Foetal and Perinatal Medicine, *Istituto Giannina Gaslini*, Genoa (Italy). The dataset consisted of 900 frames (frame size: 720 x 576 pixels) extracted from 6 videos (150 frames per video) of patients acquired during the normal surgical practice. We randomly assembled a dataset acquired from patients who received TTTS laser treatments at the same hospital.

The followed procedures were in accordance with the image data collection and retrospective study protocol approved by *Istituto Giannina Gaslini* and with the Helsinki Declaration of 1975, and revised in 2000. Data collection did not interfere or alter the current clinical practice. All the subjects involved in this research were informed and agreed to data treatment before the intervention.

The 150 frames per video were manually extracted among the ones in which the membrane was present. For each frame, ground-truth segmentation was obtained by manually tracing the membrane contour under the supervision of an expert surgeon.

The black borders surrounding the FoV do not bring any additional information to segment the membrane but increase the GPU-memory and computational-cost requirements during training. Thus, images were trimmed to the centre of the FoV and resized to fit in 256x256 pixels.

The dataset was split in a training set, consisting of intraoperative frames extracted from 4 patients, a validation set and a testing set, each one consisting of 150 frames from one video.

The achievement of such a large dataset, as recommended to avoid overfitting, was difficult because: (i) data manual annotation is a complex and time-consuming task, (ii) the data availability is limited, since TTTS is a rare pathology.

Data augmentation was performed on the training set. In particular, image rotation (by 45, 90, 180 and 270 degrees) was applied to simulate different orientations of the fetoscope during real surgery.

3.2 Training setting and Ablation Study

To limit memory requirements in the training phase, still promoting the convergence of the gradient, SAN was trained with mini-batches (batch size = 30 frames) minimising L_{SAN} (Eq. 5) with Adam[19]. The adversarial framework was trained for 600 epochs. To initialise the weights, the *segmentor* was prior trained without the *critic* in the first 25 epochs. An initial learning rate of 0.002 was set with a decay of the learning rate of 5% every 25 epochs to adjust the gradient descent. The best model was selected as the one that minimised the DSC on the validation set.

Our framework was originally proposed for skin lesion segmentation for ISBI International Skin Imaging Collaboration 2017[41]. Network parameters and training phase were expanded to fit our purpose. An ablation study was performed in order to evaluate how the segmentation performance is affected by modifying the number of layers of the S network[10]. In particular, we evaluated the S configuration with 1 to 6 layers. The study was performed using both the RGB and grey-scale images of the dataset described in Sec. 3.1.

The proposed architecture was compared with two state-of-the-art segmentation FCNNs: U-Net[33] and a residual architecture, inspired by ResNet[9], consisting of U-Net with the introduction of residual blocks. These two networks were designed to have the same depth of the S network for fair comparison. The networks were trained with the same settings of SAN (Sec. 3.2).

To evaluate inter-annotator variability we asked a second expert to annotate the fetoscopic video used as test set. We asked the second expert to annotate only the test set (150 frames) due to the high time demand needed to perform manual annotation. The manual annotation by the second expert was compared in terms of DSC versus the first expert annotation. The segmentation performed by the two experts interviewed is comparable as evidenced by the median DSC equal to 98.09%. The results are reported in Fig. 4.

The SAN was implemented on *PyTorch*[1] library and trained on *Intel i5-8400* CPU with 16GB of RAM and *NVIDIA GeForce RTX 2080Ti* GPU.

3.3 Performance metrics

For performance evaluation, we computed the DSC, defined in Eq. 2, Precision ($Prec$) and Recall (Rec):

$$Prec = \frac{TP}{TP + FP} \tag{6}$$

$$Rec = \frac{TP}{TP + FN} \tag{7}$$

[Figure 3 about here.]

[Figure 4 about here.]

The Lilliefors test was used to assess population normality on DSC. The Kruskal-Wallis on DSC and Westenberg-Mood test on IQR, both imposing a significance level (p) equal to 0.05, were used to assess whether or not remarkable differences existed between the tested architectures.

4 Results

SAN training lasted ∼6 hours. The processing time of images in the test set was less than a millisecond, on average. This performance confirms the compatibility with real-time applications of this approach.

[1] https://pytorch.org/

The addition of encoding layers tended to enhance the network performances until 4 encoding-decoding layers are reached, as shown in Fig. 3. Further addition of encoding-decoding layers after two did not produce remarkable differences for the median DSC. The DSC increase from 76.49% with IQR of 41.48% to 92.35% with IQR of 20.60%. The best performing architecture was SAN, with five encoder-decoder layers which showed the highest $Prec$ (98.33%) and Rec (98.84%) with lower IQR of 4.19% and 1.44%, respectively. The median DSC (91.91%) was slightly lower than the other configurations, but guaranteed a significant ($p < 0.01$) lower IQR (4.63%).

The boxplots in Fig. 4 show the performance comparison between the state of art architectures (U-Net and residual network) and the adversarial framework in terms of DSC, for both the grey-scale and the RGB datasets. SAN achieved better results than the other tested networks in terms of DSC. When comparing with U-Net and the residual architecture on the grey-scale dataset, we observed an increase of 5.95% ($p < 0.01$) and 3.64% ($p < 0.01$) respectively. Instead, using the RGB dataset, the improvement compared to U-Net was of 8.93% ($p < 0.01$) and 5.78% ($p < 0.05$) more than the residual architecture. Comparing the results, the median DSC obtained for the grey-scale dataset for U-Net, the residual and the proposed adversarial network were 80.25% (16.92%), 82.56% (13.63%) and 86.20% (6.81%), respectively. Feeding the FCNNs with the RGB images allowed for the achievement of better results, as shown in Fig. 4. In this case, the median DSC (IQR in brackets) for U-Net, residual architecture and the proposed adversarial network of 82.98% (14.41%), 86.13% (13.63%) and 91.91% (4.63%).

Representative segmentation results are shown in Fig. 5, where grey, blue and green contours refer to the ground-truth, grey scale-based and RGB-based segmentation results, respectively. In (i) all the networks achieved good results despite the presence of spots and specularities; (ii) all networks achieved good results despite the fact that the U-Net and the residual architecture produced some spots in the lower area where the texture could suggest the presence of the membrane. The proposed framework achieves very good results; (iii) U-Net, and especially the residual architecture, produced some spots due to the high illumination of the environment caused by the closeness of the fetoscope to the membrane. Even though the scenario was more lightened up, this condition also increased the produced specularities. In (iv), the presence of low contrast and laser light compromises the detection of the membrane in U-Net and Residual networks while in our framework produces good segmentation. This suggests that the action of *critic* network provides the ability to the *segmentor* network to enhance the processing of poor quality images (*e.g.*, with laser pointer, light specularities, drop of light intensity, etc.). Furthermore, integrating our framework with frame selection strategies will allow the network to discard very low quality frames, avoid the need to process them.

U-Net and the residual architecture obtained unsatisfactory results, as confirmed by DSC values for RGB dataset of 41.06% for U-Net and 37.82% for the residual architecture. SAN model achieved better results with DSC of 91.42%; (v) In this scene there are some peculiarities of the foetal environment such the presence of vessels crossing the membrane and the suspended particulate, a common condition of advanced pregnancies. Here, the residual architecture achieved better results following the ground-truth in the upper part, even though it produces irregularities along the vessel, in particular near the membrane crossing area. SAN, achieved good results, despite the presence of minor inaccuracies where the vessel crosses the membrane; (vi) in this case U-Net and the residual architecture outperform the proposed adversarial network.

[Figure 5 about here.]

5 Discussion

During TTTS surgery, the identification of the inter-foetal membrane helps the surgeon to remain oriented in the surgical site. The complexity of the placental environment, especially in advanced pregnancies, makes this task very challenging also for expert clinicians when performing surgery.

In this paper, we propose and studied a novel framework based on adversarial training, inspired by[42], in order to address the problem of inter-foetus membrane segmentation. We also compare this framework with state-of-the-art FCNNs for medical-image segmentation. An ablation study was performed, showing that the S network with 5 encoding-decoding layers was the best combination between segmentation performance and robustness. We infer that one of the reasons is that the size of the training data is relatively small, which affects the generalisation capability of complex (deep) model.

The SAN framework showed encouraging improvement when compared to the tested state-of-the-art FCNNs, according to the DSC, $Prec$ and Rec. As shown in Fig. 5, the other tested networks predicted shapes that have never been seen during training, suggesting that macro-appearances were not considered. This may be related to the fact that the networks work with kernels with a small receptive field (kernel size = 5 x 5), as trade off between the number of parameters to be learnt during training and the segmentation accuracy. With the proposed architecture, the additional L_1 loss term computed by the adversarial C network realised a multi-scale features analysis during training. As a consequence, also the high-level features connected to macro appearance were preserved during prediction.

Moreover, the presented framework achieved better results on the RGB dataset, showing that the networks can successfully exploit the additional information embedded in the colour channels. This limitation is probably due to the combination of multiple penalising factors like the very small size and unfavourable position of the membrane in the FoV and the low contrast of the membrane boundary due to the high level of illumination.

Despite our efforts, due to the limited amount of available videos and the complexity of the task, the dataset size remains a strong limitation of this study. For this reason, some kind of images (*e.g.*, with a small portion of the membrane) are less numerous than others, limiting the network learning capability.

This problem will be addressed in the future by investigating extensions of this framework supported by a broader dataset and more advanced data augmentation techniques.

Further improvements will deal with the exploitation of temporal features, as suggested in[43,6], considering that the temporal information is naturally encoded in the surgical videos.

Inclusion of temporal features could be implemented by considering the use of 3D convolutions leading to a complex network, in which training is not trivial due to the increasing number of parameters that should be learnt. In fact, temporal information was highly relevant during the followed manual tracing process to build the ground-truth dataset. It was clear that such information was essential since a common strategy used in difficult cases was that of scanning a certain number of consecutive frames to better identify the inter-foetal membrane. Frame selection strategies[26] could be exploited too, such as to avoid the processing of uninformative (*e.g.*, blurred) video portions.

The proposed approach may also be integrated with recent work, which deals with vessel segmentation from placenta images[1,34], stitching of fetoscopy images to build placental panoramic image[12,44] and classification of TTTS surgical phases[38].

5.1 Conclusion

In this paper, we proposed an adversarial framework for accurate and fast inter-foetal membrane segmentation in fetoscopic MIS images achieving a median DSC of 91.91% on a new dataset of 150 images from intraoperative TTTS surgery videos. This work is among the first attempts of surgical data science in TTTS surgery and has great potential to support surgeons during fetoscopic MIS surgery and enhance TTTS surgical outcomes, giving the encouraging results.

Ethical standards

This article followed the Ethics Guidelines for Trustworthy Artificial Intelligence, recently published by the European Commission[2].

[2] https://ec.europa.eu/futurium/en/ai-alliance-consultation

Conflict of Interest

No benefits in any form have been or will be received from a commercial party related directly or indirectly to the subjects of this manuscript.

References

1. Almoussa, N., B. Dutra, B. Lampe, P. Getreuer, T. Wittman, C. Salafia, and L. Vese. Automated vasculature extraction from placenta images. In: Medical Imaging 2011: Image Processing, volume 7962, p. 79621L, International Society for Optics and Photonics2011.

2. Badrinarayanan, V., A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39:2481–2495, 2017.

3. Bolch, C., M. Fahey, D. Reddihough, K. Williams, S. Reid, A. Guzys, S. Cole, A. Edwards, A. Fung, R. Hodges *et al.* Twin-to-twin transfusion syndrome neurodevelopmental follow-up study (neurodevelopmental outcomes for children whose twin-to-twin transfusion syndrome was treated with placental laser photocoagulation). BMC pediatrics 18:256, 2018.

4. Brandao, P., O. Zisimopoulos, E. Mazomenos, G. Ciuti, J. Bernal, M. Visentini-Scarzanella, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo *et al.* Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks. Journal of Medical Robotics Research 3:1840002, 2018.

5. Chhatkuli, A., A. Bartoli, A. Malti, and T. Collins. Live image parsing in uterine laparoscopy. In: IEEE International Symposium on Biomedical Imaging, pp. 1263–1266, IEEE2014.

6. Colleoni, E., S. Moccia, X. Du, E. De Momi, and D. Stoyanov. Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. IEEE Robotics and Automation Letters 4:2714–2721, 2019.

7. Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09. 2009.

8. Deprest, J. A., D. V. Schoubroeck, P. P. V. Ballaer, H. Flageole, F. A. V. Assche, and K. Vandenberghe. Alternative technique for nd : YAG laser coagulation in twin-to-twin transfusion syndrome with anterior placenta. Ultrasound in Obstetrics and Gynecology 11:347–352, 1998.

9. Drozdzal, M., E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The importance of skip connections in biomedical image segmentation. In: Deep Learning and Data Labeling for Medical Applications, pp. 179–187, Springer2016.

10. Du, X., T. Kurmann, P. Chang, M. Allan, S. Ourselin, R. Sznitman, J. D. Kelly, and D. Stoyanov. Articulated multi-instrument 2d pose estimation using fully convolutional networks. IEEE Transactions on Medical Imaging 37:1276–1287, 2018.

11. Everingham, M., L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88:303–338, 2010.

12. Gaisser, F., S. H. Peeters, B. Lenseigne, P. Jonker, and D. Oepkes. Stable image registration for in-vivo fetoscopic panorama reconstruction. Journal of Imaging 4, 2018.

13. Ghosh, T., L. Li, and J. Chakareski. Effective deep learning for semantic segmentation based bleeding zone detection in capsule endoscopy images. In: IEEE International Conference on Image Processing, pp. 3034–3038, IEEE2018.

14. Goodfellow, I., Y. Bengio, and A. Courville. Deep learning, MIT press2016.

15. Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In: Advances in Neural Information Processing Systems 27, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, pp. 2672–2680, Curran Associates, Inc.2014.

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

17. Huber, A., A. A. Baschat, T. Bregenzer, A. Diemert, M. Tchirikov, B. J. Hackelöer, and K. Hecher. Laser coagulation of placental anastomoses with a 30° fetoscope in severe mid-trimester twin–twin transfusion syndrome with anterior placenta. Ultrasound in Obstetrics and Gynecology 31:412–416, 2008.

18. Ioffe, S. and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456. 2015.

19. Kingma, D. P. and J. Ba. Adam: A method for stochastic optimization. In: International Conference on Learning Representation. 2015.

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

21. Long, J., E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440. 2015.

22. Luc, P., C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. In: Neural Information Processing Systems Workshop on Adversarial Training. 2016.

23. Maas, A. L., A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In: International Conference on Machine Learning Workshop on Deep Learning for Audio, Speech and Language Processing. 2013.

24. Mewes, P. W., D. Neumann, O. Licegevic, J. Simon, A. L. Juloski, and E. Angelopoulou. Automatic region-of-interest segmentation and pathology detection in magnetically guided capsule endoscopy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 141–148, Springer2011.

25. Miller, W., J. Novotny, D. Laidlaw, F. L. Luks, D. Merck, and S. Collins. Virtually visualizing vessels : A study of the annotation of placental vasculature from mri in large-scale virtual reality for surgical planning. In: Brown University, Providence. 2016.

26. Moccia, S., G. O. Vanone, E. De Momi, A. Laborai, L. Guastini, G. Peretti, and L. S. Mattos. Learning-based classification of informative laryngoscopic frames. Computer Methods and Programs in Biomedicine 158:21–30, 2018.

27. Moccia, S., S. J. Wirkert, H. Kenngott, A. S. Vemuri, M. Apitz, B. Mayer, E. De Momi, L. S. Mattos, and L. Maier-Hein. Uncertainty-aware organ classification for surgical data science applications in laparoscopy. IEEE Transactions on Biomedical Engineering 65:2649–2659, 2018.

28. Odena, A., V. Dumoulin, and C. Olah. Deconvolution and checkerboard artifacts. Distill , 2016.

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

30. Prokopetc, K., T. Collins, and A. Bartoli. Automatic detection of the uterus and fallopian tube junctions in laparoscopic images. In: International Conference on Information Processing in Medical Imaging, pp. 552–563, Springer2015.

31. Quintero, R. A. Twin–twin transfusion syndrome. Clinics in Perinatology 30:591–600, 2003.

32. Roberts, D., J. P. Neilson, M. D. Kilby, and S. Gates. Interventions for the treatment of twin-twin transfusion syndrome. Cochrane Database of Systematic Reviews , 2014.

33. Ronneberger, O., P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention p. 234–241, 2015.

34. Sadda, P., M. Imamoglu, M. Dombrowski, X. Papademetris, M. O. Bahtiyar, and J. Onofrey. Deep-learned placental vessel segmentation for intraoperative video enhancement in fetoscopic surgery. International Journal of Computer Assisted Radiology and Surgery 14:227–235, 2019.

35. Senat, M.-V., J. Deprest, M. Boulvain, A. Paupe, N. Winer, and Y. Ville. Endoscopic laser surgery versus serial amnioreduction for severe twin-to-twin transfusion syndrome. New England Journal of Medicine 351:136–144, 2004.

36. Shen, D., G. Wu, and H.-I. Suk. Deep learning in medical image analysis. Annual Review of Biomedical Engineering 19:221, 2017.

37. Tella, M., P. Daga, F. Chadebecq, S. Thompson, D. I. Shakir, G. Dwyer, R. Wimalasundera, J. Deprest, D. Stoyanov, T. Vercauteren, and S. Ourselin. A combined em and visual tracking probabilistic model for robust mosaicking: Application to fetoscopy. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE2016.

38. Vasconcelos, F., P. Brandão, T. Vercauteren, S. Ourselin, J. Deprest, D. Peebles, and D. Stoyanov. Towards computer-assisted TTTS: Laser ablation detection for workflow segmentation from fetoscopic video. International Journal of Computer Assisted Radiology and Surgery 13:1661–1670, 2018.

39. Wickstrøm, K., M. Kampffmeyer, and R. Jenssen. Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. In: IEEE International Workshop on Machine Learning for Signal Processing, pp. 1–6, IEEE2018.

40. Xu, B., N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 , 2015.

41. Xue, Y., T. Xu, and X. Huang. Adversarial learning with multi-scale loss for skin lesion segmentation. In: IEEE International Symposium on Biomedical Imaging, IEEE2018.

42. Xue, Y., T. Xu, H. Zhang, L. R. Long, and X. Huang. Segan: Adversarial network with multi-scale l1 loss for medical image segmentation. Neuroinformatics 16:383–392, 2018.

43. Yu, L., H. Chen, Q. Dou, J. Qin, and P. A. Heng. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. IEEE Journal of Biomedical and Health Informatics 21:65–75, 2016.

44. Bano, S., F. Vasconcelos, M. Tella, G. Dwyer, C. Gruijthuijsen, J. Deprest, S. Ourselin, E. Vander Poorten, T. Vercauteren and D. Stoyanov. Deep Sequential Mosaicking of Fetoscopic Videos. Medical Image Computing and Computer Assisted Intervention, 2019

45. Zhu, X., X. Zhang, X.-Y. Zhang, Z. Xue, and L. Wang. A novel framework for semantic segmentation with generative adversarial network. Journal of Visual Communication and Image Representation 58:532 – 543, 2019.

List of Figures

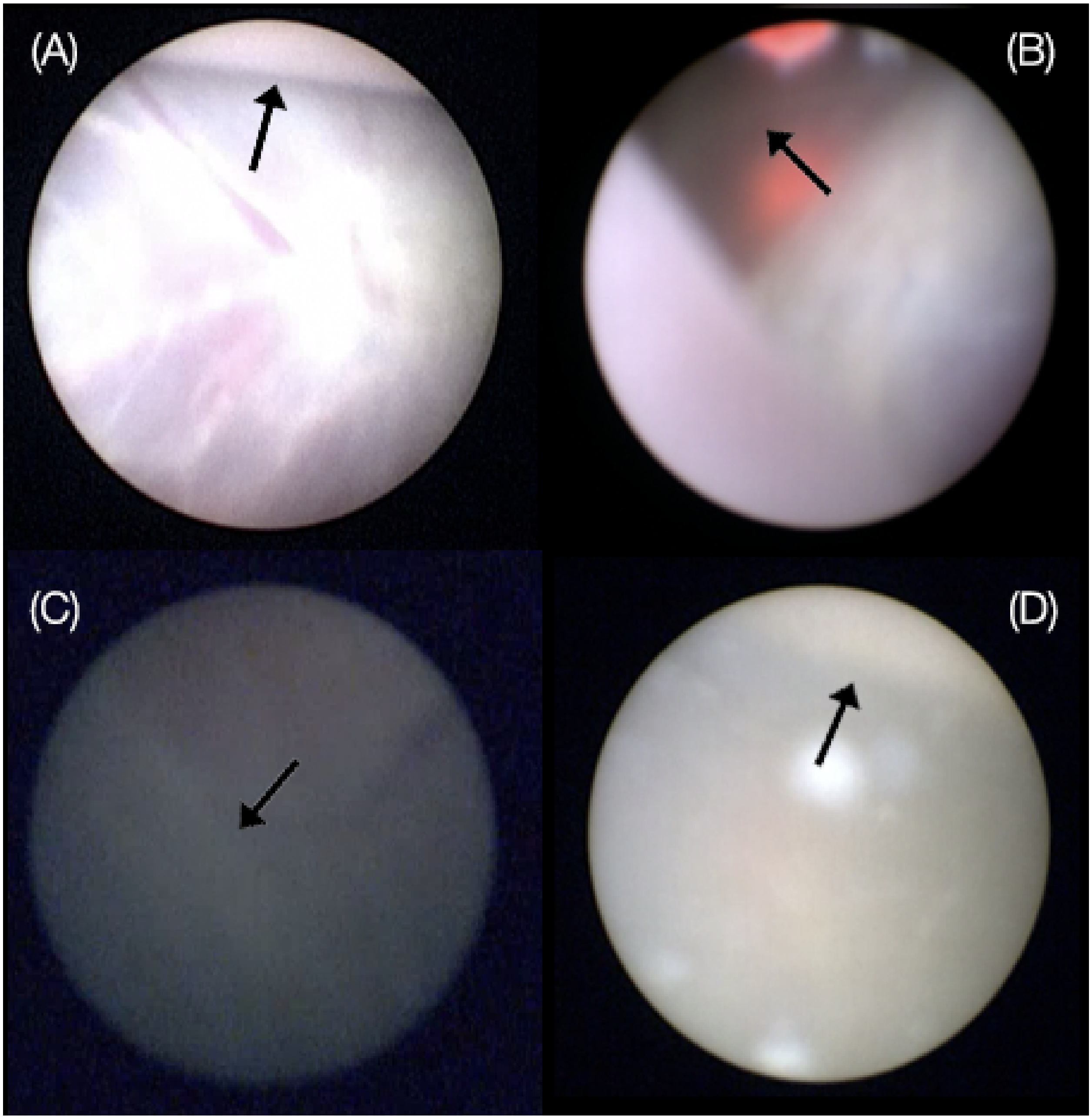

Figure 1: Examples of challenging cases for inter-foetus membrane identification. (A) The membrane covers a small portion of the field of view, (B) anterior placenta partially occludes the membrane, (C) image has low illumination level, (D) amniotic fluid turbidity makes the image blurred.

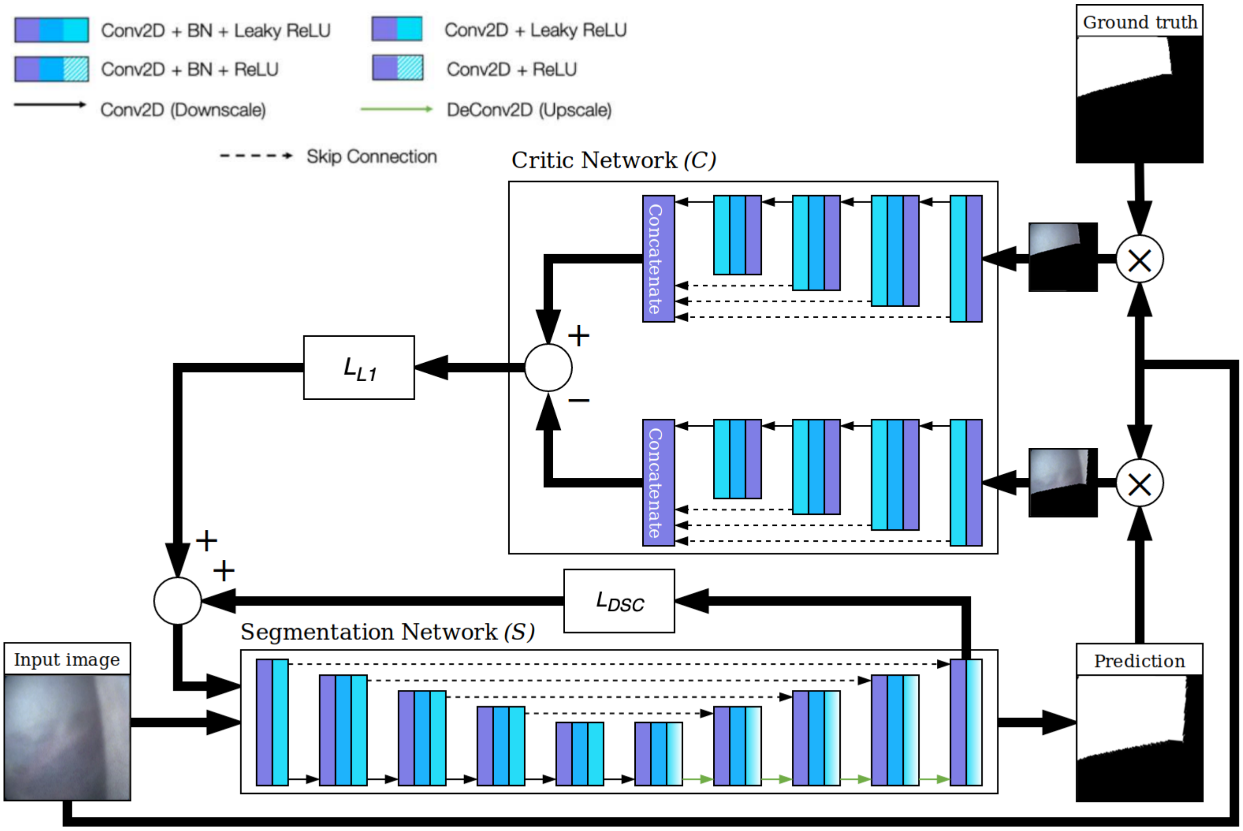

Figure 2: The proposed Segmentation Adversarial Network (SAN) architecture. Dashed arrows refer to skip connections. Black thin arrows refer to 2D strided convolution (downscale). Green thin arrow refers to 2D strided deconvolution. Conv2D-BN-ReLU module: 2D convolution followed by batch normalization (BN) and rectified linear unit (ReLU) activation. Conv2D-BN-Leaky ReLU module: 2D convolution followed by batch normalization (BN) and leaky rectified linear unit (ReLU) activation. Only the first downscale (last upscaling) block does not include a batch normalization (BN) layer. Concatenate: join the two feature vector with the same shape, from the *critic* network, to assemble a unique output. Masked images are calculated by pixel-wise multiplication (×) of the ground-truth (predicted) mask and the input image.

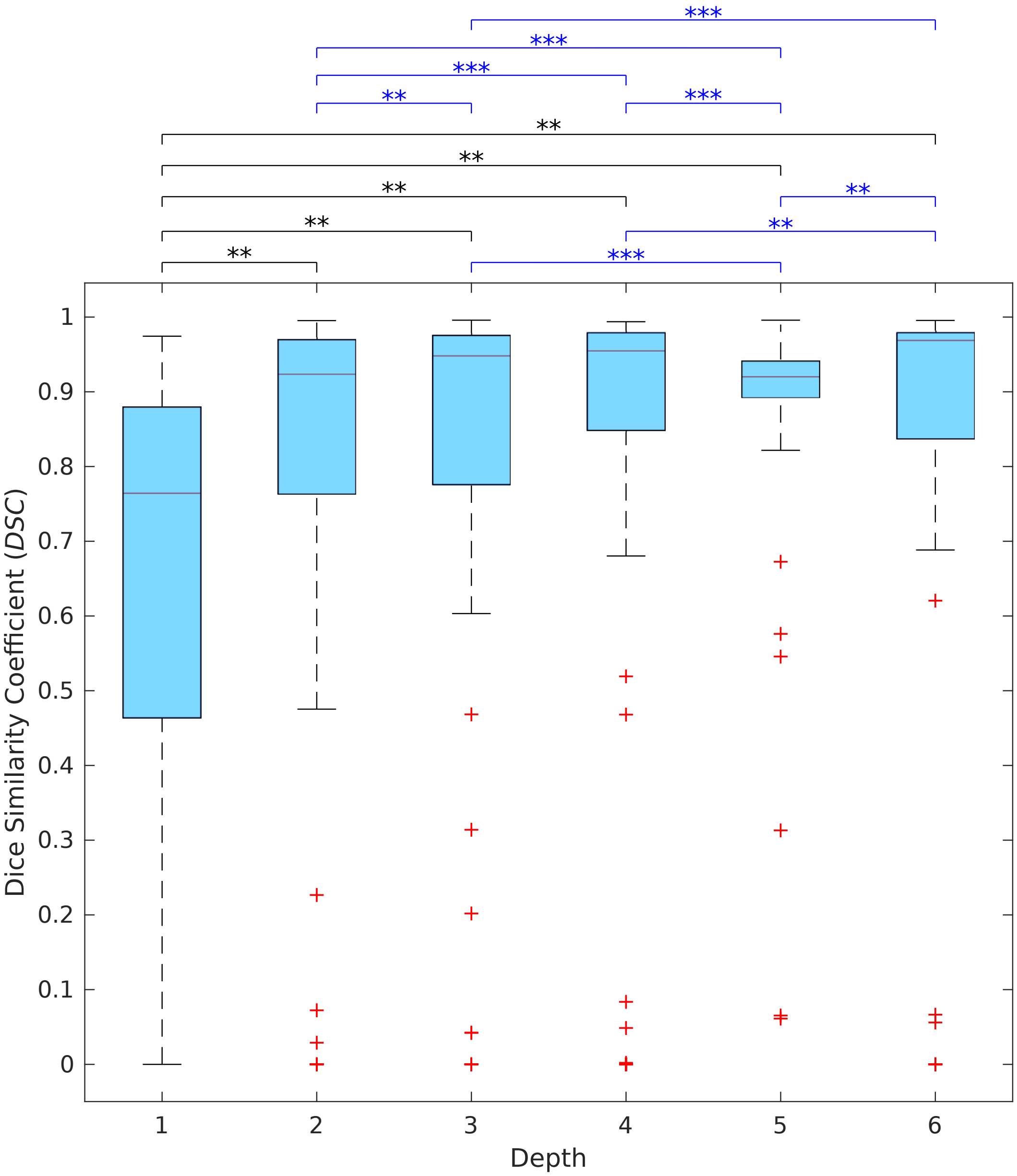

Figure 3: Results comparison using different depth of the *segmentor* Network for the RGB dataset. Blue and black asterisks highlight significant differences between the different architectures in terms of median DSC (Kruskal-Wallis) and inter-quartile range (Westenberg-Mood) ($*p < 0.05$, $**\,p < 0.01$, $***p < 0.001$), respectively.

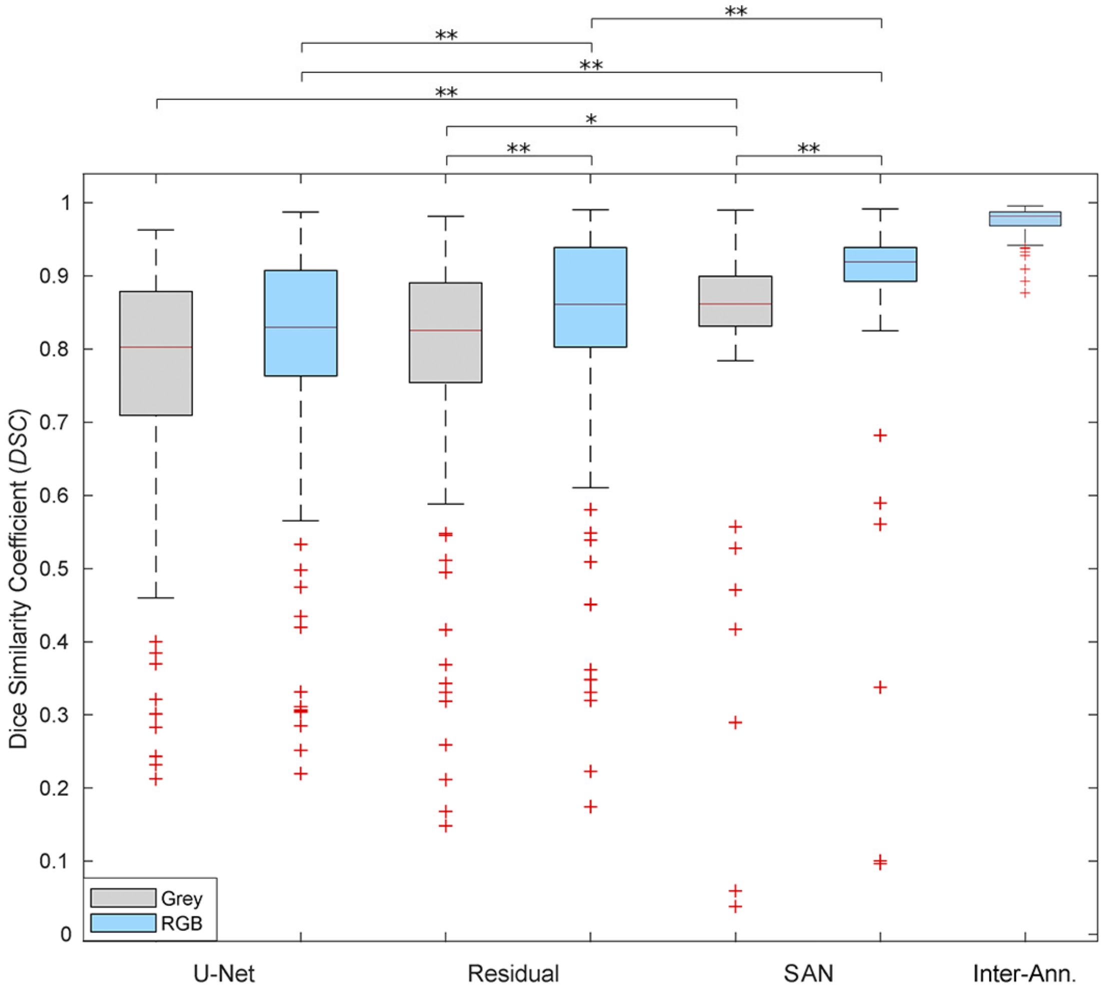

Figure 4: *Dice similarity coefficient* (DSC) obtained testing the state-of-the-art architectures (U-Net and residual architecture) and the SAN architecture. In the last boxplot, annotation performed by a second expert clinician is compared to consider inter-annotator variability (Inter-ann.). Performance metrics were calculated feeding the networks with the grey-scale dataset (in grey) and the RGB dataset (in blue). Asterisks indicate statistical difference in median DSC with Kruskal-Wallis test ($*p < 0.05$, $**\,p < 0.01$).

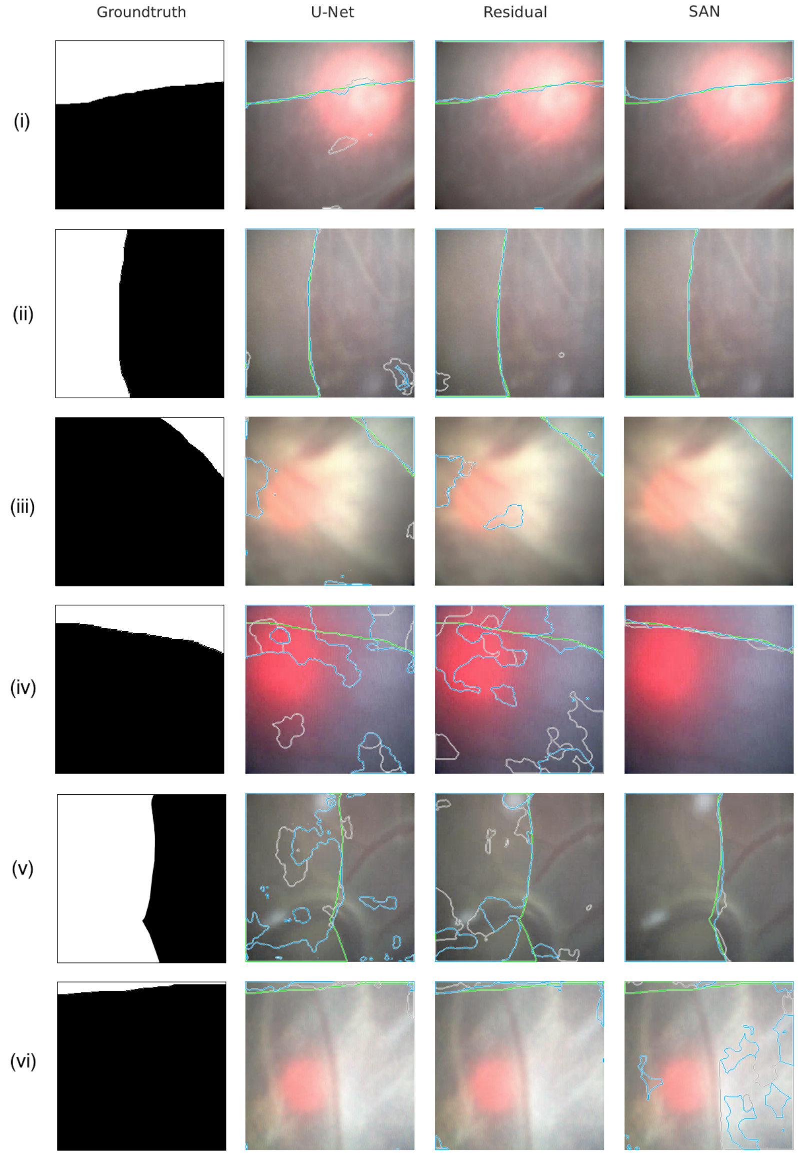

Figure 5: Sample segmentation results on the test set using (second column) U-Net, (third column) U-Net with the residual implementation and (last column) the proposed SAN along the manual expert clinician ground-truth (first column). Each network was trained both with grey-scale and RGB fetoscopic images. The green, grey and blue contours refers to the ground-truth, grey scale-based and RGB-based segmentation results, respectively.

List of Tables

Table 1: Specifications of the *segmentor* network (S) architecture. Kernel size and stride (kernel height x kernel width), as well as output dimensions (height (H) x width (W) x N. Channels) of each layer, are shown. The final output is a segmentation mask with the same dimension of the input.

Segmentor network (S)

	Encoder			*Decoder*	
	Kernel (Size / Stride)	**Output**		**Kernel (Size / Stride)**	**Output**
Strided Conv 0	7x7 / 2x2	$\frac{H}{2}$ x $\frac{W}{2}$ x 16	**DeConv 0**	3x3 / 1x1	$\frac{H}{16}$ x $\frac{W}{16}$ x 128
R0 Conv 0	1x1 / 1x1	$\frac{H}{2}$ x $\frac{W}{2}$ x 32	**R0 Conv 0**	1x1 / 1x1	$\frac{H}{16}$ x $\frac{W}{16}$ x 256
R1 Conv 0	3x3 / 1x1	$\frac{H}{2}$ x $\frac{W}{2}$ x 32	**R1 Conv 0**	3x3 / 1x1	$\frac{H}{16}$ x $\frac{W}{16}$ x 256
R2 Conv 0	1x1 / 1x1	$\frac{H}{2}$ x $\frac{W}{2}$ x 16	**R2 Conv 0**	1x1 / 1x1	$\frac{H}{16}$ x $\frac{W}{16}$ x 128
Strided Conv 1	5x5 / 2x2	$\frac{H}{4}$ x $\frac{W}{4}$ x 32	**DeConv 1**	3x3 / 1x1	$\frac{H}{8}$ x $\frac{W}{8}$ x 128
R0 Conv 1	1x1 / 1x1	$\frac{H}{4}$ x $\frac{W}{4}$ x 64	**R0 Conv 1**	1x1 / 1x1	$\frac{H}{8}$ x $\frac{W}{8}$ x 128
R1 Conv 1	3x3 / 1x1	$\frac{H}{4}$ x $\frac{W}{4}$ x 64	**R1 Conv 1**	3x3 / 1x1	$\frac{H}{8}$ x $\frac{W}{8}$ x 128
R2 Conv 1	1x1 / 1x1	$\frac{H}{4}$ x $\frac{W}{4}$ x 32	**R2 Conv 1**	1x1 / 1x1	$\frac{H}{8}$ x $\frac{W}{8}$ x 64
Strided Conv 2	5x5 / 2x2	$\frac{H}{8}$ x $\frac{W}{8}$ x 64	**DeConv 2**	3x3 / 1x1	$\frac{H}{4}$ x $\frac{W}{4}$ x 64
R0 Conv 2	1x1 / 1x1	$\frac{H}{8}$ x $\frac{W}{8}$ x 128	**R0 Conv 2**	1x1 / 1x1	$\frac{H}{4}$ x $\frac{W}{4}$ x 64
R1 Conv 2	3x3 / 1x1	$\frac{H}{8}$ x $\frac{W}{8}$ x 128	**R1 Conv 2**	3x3 / 1x1	$\frac{H}{4}$ x $\frac{W}{4}$ x 64
R2 Conv 2	1x1 / 1x1	$\frac{H}{8}$ x $\frac{W}{8}$ x 64	**R2 Conv 2**	1x1 / 1x1	$\frac{H}{4}$ x $\frac{W}{4}$ x 32
Strided Conv 3	5x5 / 2x2	$\frac{H}{16}$ x $\frac{W}{16}$ x 128	**DeConv 3**	3x3 / 1x1	$\frac{H}{2}$ x $\frac{W}{2}$ x 32
R0 Conv 3	1x1 / 1x1	$\frac{H}{16}$ x $\frac{W}{16}$ x 256	**R0 Conv 3**	1x1 / 1x1	$\frac{H}{2}$ x $\frac{W}{2}$ x 32
R1 Conv 3	3x3 / 1x1	$\frac{H}{16}$ x $\frac{W}{16}$ x 256	**R1 Conv 3**	3x3 / 1x1	$\frac{H}{2}$ x $\frac{W}{2}$ x 32
R2 Conv 3	1x1 / 1x1	$\frac{H}{16}$ x $\frac{W}{16}$ x 128	**R2 Conv 3**	1x1 / 1x1	$\frac{H}{2}$ x $\frac{W}{2}$ x 16
Strided Conv 4	5x5 / 2x2	$\frac{H}{32}$ x $\frac{W}{32}$ x 256	**DeConv 4**	3x3 / 1x1	H x W x 16
			R0 Conv 4	1x1 / 1x1	H x W x 16
			R1 Conv 4	3x3 / 1x1	H x W x 16
			R2 Conv 4	1x1 / 1x1	H x W x 8
			Conv 5	3x3 / 1x1	H x W x 1

Table 2: Specifications of the proposed *Critic* network architecture. Kernel size and stride (kernel height x kernel width), as well as output dimensions (height (H) x width (W) x N. Channels) of each layer, are shown. The final output is a segmentation mask with the same dimension of the input.

***Critic* network**

	Kernel (Size / Stride)	**Output**
Strided Conv 0	7x7 / 2x2	$\frac{H}{2}$ x $\frac{W}{2}$ x 16
Strided Conv 1	5x5 / 2x2	$\frac{H}{4}$ x $\frac{W}{4}$ x 32
Strided Conv 2	5x5 / 2x2	$\frac{H}{8}$ x $\frac{W}{8}$ x 64
Strided Conv 4	5x5 / 2x2	$\frac{H}{16}$ x $\frac{W}{16}$ x 128