HOG Injection into ANN for Fetoscopic Multi-Class Segmentation
FetReg Challenge: Placental Vessel Segmentation and Registration in Fetoscopy

BioPolimi Team: Jessica Biagioli[1] Ilaria Anita Cintorrino[2] Gaia Romana De Paolis[3] Chiara Lena[4]

Abstract— **Twin-to-Twin Transfusion Syndrome (TTTS) is a rare pregnancy condition affecting identical twins that share one placenta and a network of blood vessels that supply unbalanced oxygen and nutrients to donor and recipient twins. Fetoscopic Minimally Invasive Surgery (MIS) is the best first line treatment for this pathology. It consists of a direct interruption of pathological anastomoses characterizing TTTS. In Fetoscopic Placental Vessel Segmentation and Registration (FetReg) challenge a large-scale multi-centre dataset for semantic segmentation and video mosaicking algorithms in fetal environment was presented. In this paper we provide a new framework for the first task of this challenge, which is based on the combination of features extracted with *Histogram of Oriented Gradients* (HOG) and deep learning methods. HOG is particularly efficient in recognizing Tool and Fetus classes that are the less present in the dataset. Moreover we compare the traditional Convolutional Neural Networks (CNNs) based on ResNet50 with a recent Transformer-based architecture, TransUnet, in multi-class segmentation task. The injection of HOG in the features of ResNet50 latent space (ResNet50-HOG) achieves a significant improvement in the segmentation of Tool and Fetus objects and a robust segmentation of vessels.**

I. INTRODUCTION

The Twin-to-Twin Transfusion Syndrome (TTTS) occurs in the 10 - 15% of monochorionic pregnancies (i.e. twin pregnancies with shared placenta). The aetiology of TTTS is the unequal blood flow along placental blood vessels known as anastomoses. Due to the formation of abnormal vascular placental connections, blood flow becomes unbalanced between the fetuses. The pathology can have serious consequences for both twins and the risk of perinatal mortality of one or both fetuses can exceed 90% without any treatment [4].

The definitive treatment is Fetoscopic Minimally Invasive Surgery (MIS), that consists of a direct interruption of pathological anastomoses, via laser photo-coagulation. The surgeon must identify the vessels during surgery due to the absence of imaging techniques to preoperatively visualize them.

Therefore, the identification of the inter-twin anastomoses is a challenging task due to different factors:

- limited Field of View (FoV)
- possible out of focus of the fetoscope due to dynamic changes in the fetal environment
- turbid surgical environment (amniotic fluid)
- large variability in the illumination level
- poor video quality and resolution

All these factors impair the surgeon ability to remain oriented during the procedure. The result is an increased procedural time and incomplete ablation of anastomoses leading to persistent TTTS. One solution is to provide Computer-Assisted Interventions (CAI) support to the surgeon by the automatic segmentation of the main landmarks in the surgical scenario as the laser, the fetus and especially placental vessels. Recent studies on placental vessel segmentation [14], [3] focused on U-Net architecture [13], which is a milestone for image segmentation. However, the reduced dimension of the datasets limits the generalizability and performance of these methods. In **FetReg: Placental Vessel Segmentation and Registration in Fetoscopy**[1] organized inside the EndoVis MICCAI Grand Challenge[2], a multi-centre large-scale dataset is realised, providing the opportunity to improve the current state-of-the-art. In this paper we describe the proposed model for placental vessel segmentation, that is based on the integration of traditional feature engineering and deep learning techniques. Feature Engineering is the process to extract features from images by means of *transformation* functions [15]. Due to its high performance and robustness in visual object recognition, we used Histogram of Oriented Gradients (HOG) [12] as attribute extraction method, to obtain information about the distribution of the target edges. To our knowledge, in literature it was already introduced the association of HOG features and neural networks but it was mainly related to classification [11] or detection problems [1], [10].

The novelty of this study is the combination of the information carried by the HOG feature into the Artificial Neural Network (ANN) in order to perform multi-class segmentation, by concatenating the HOG vector and the latent space of the ANN. This allows to improve the identification of less present classes in the dataset since HOG easily recognizes Tool and Fetus boundaries. We implemented this technique on the latent space, but it could be applied also to the other layers of the network. Moreover, the recent introduction of Transformers [17], which emerged as an alternative to Convolutional Neural Networks (CNN), leads us also to investigate and to compare these methods. In particular Chen et. al in [5] proposed an hybrid CNN-Transformer

[1,2,3,4] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy. **We all agree to make our submission public as part of the challenge archive.**

[1] jessica.biagioli@mail.polimi.it

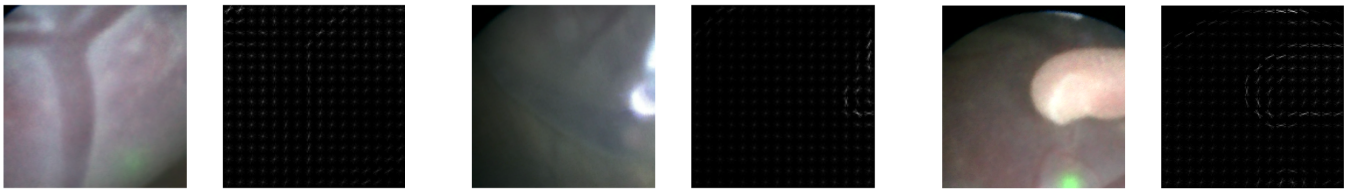
[2] ilariaanita.cintorrino@mail.polimi.it

[3] gaiaromana.depaolis@mail.polimi.it

[4] chiara.lena@mail.polimi.it

[1] FetReg2021 Challenge website: https://fetreg2021.grand-challenge.org/

[2] MICCAI EndoVis Challenge website: https://endovis.grand-challenge.org/



(a) *Vessel* (b) *Tool* (c) *Fetus*

Fig. 1: Examples of Histogram of Oriented Gradient features applied to fetoscopic images with presence of Vessel (a), Tool (b), Fetus (c). Images are obtained by randomly cropping the original frame, with patch dimension of 256x256 pixels, and by applying the transformations described in Sec.III-B.

architecture, *TransUnet*, for medical image segmentation, that enables both detailed high-resolution spatial information from CNN features and the global context encoded by Transformers. Our contributions can be summarized as follows:

- Hand-crafted HOG feature injection into ANN target layer in order to improve the performance of multi-class segmentation task.
- Comparisons between different architectures based on traditional CNN and Transformer-based models.

Our code is available at: `https://github.com/IlariaAnita/FetReg_Challenge.git`.

II. MATERIALS AND METHODS

This section describes the proposed model for vessels segmentation from intra-operative fetoscopic videos and the training strategy. The general framework is shown in Fig. 2. As far as the neural networks are concerned, we implemented the traditional ResNet50 [9] in its version for semantic segmentation (Fully Convolutional Neural Networks [16], DeepLabv3 [7]) and TransUnet [5]. As first, the raw images are preprocessed and cropped randomly into patches of 256x256. From each sample HOG is extracted and it is concatenated to the latent space of the ANN generating three new architectures: *ResNet50-HOG*, *DeepLabv3-HOG* and *TransUnet-HOG*. Some examples of HOG images are shown in Fig.1.

A. HOG Features Injection into ANN

HOG is a method to characterize object shape and appearance by the distribution of local intensity gradients or edge directions [12]. The image is divided into small cells and for each cell the histogram of gradient directions over its pixels is computed. We set cell dimension as 16x16. Each histogram divides the gradient angle range into a fixed number of predetermined bins (8 in our model). The gradient magnitudes of the pixels in the cell are used to vote into the orientation histogram. The histograms are then normalized for better illumination invariance. The complete representation is given from the combination of the local histograms and it is flattened to obtain a 1D vector, named here as *HOG vector*. The proposed HOG vector injection is applied in this study at latent space level, but it will be described in a general way, since it could be embedded at different layers of the network. Not considering the batch size, the feature of the target layer has dimension (*N x H x W*) which are respectively, number of channels for CNN or embedding space dimension in Transformer-based networks, height, width. The HOG vector length is L_{HOG} for each image (2048 in our model). Flatten operation is implemented on the target ANN feature in order to obtain a 1D vector to be concatenated with the HOG vector, resulting in a new 1D vector. To regain the original image size (WxH), the latter is reshaped in dimension (*N_new, H, W*) in which *N_new* is computed as follows:

$$N\_new = N + int(L_{HOG}/(H \times W)) \tag{1}$$

The new obtained feature is then given as input to the next layer.

B. ResNet50-HOG

This model is composed of ResNet50 as backbone, which is based on deep residual learning [9]. Shortcut connections are used to perform identity function and their outputs are added to the output of the next layers. We adopted the architecture with 50 convolution layers. The fully connected layers in the classifier are replaced by convolution layers in order to output a heatmap [16]. We will refer to this architecture as ResNet50-FCN where FCN stands for *Fully Convolutional Network*. As already mentioned, HOG vector injection is applied at latent space level. The feature dimension of the last layer in the backbone corresponds to (2048x32x32). As illustrated in Sec.II-A, HOG is embedded in this layer resulting in a new feature dimension of (2050x32x32) for this architecture. This new feature is now the input of the segmentation head.

C. DeepLabv3-HOG

DeepLabv3 [7] is an adaptation of ResNet for semantic segmentation by applying atrous convolution to extract dense features. Atrous convolution enables the adaptive modification of the filter's field-of-view by changing the dilation rate. In particular in this architecture it is used a revisited version of Atrous Spatial Pyramid Pooling (ASPP) [6] that with different atrous rates allows to capture multi-scale information. Also in this model we implement ResNet50 as backbone. The HOG injection is performed with the mechanism described in Sec.II-A and since the base network is the same, the references for features dimensions are the ones reported in Sec.II-B.

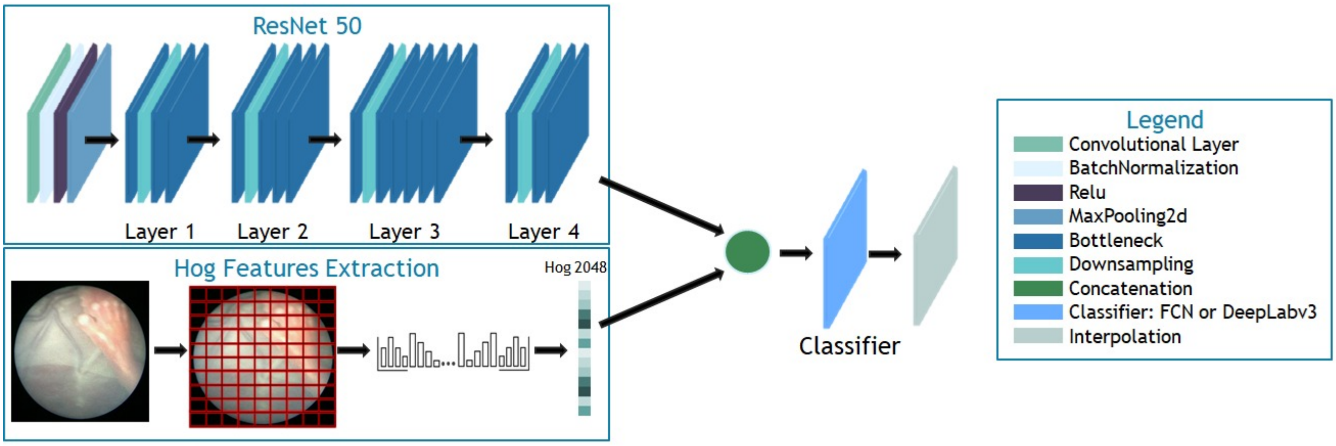


Fig. 2: Overview of the proposed framework. From the original image, HOG vector is extracted, capturing shapes and edge directions. HOG vector is concatenated with the features of a target layer of the network, performing HOG injection. The illustration refers to ResNet50 as architecture and to Latent Space as target layer, but it can be implemented to every ANN and layer. The combination of the two features (green circle) is then the input to the next layer of the network.

D. *TransUnet-HOG*

TransUnet [5] follows the traditional U-shape of U-Net [13]. Thus, it is composed of an encoder and a decoder. The encoder is a CNN-Transformer hybrid. Patches are extracted from the CNN features rather than from raw images. Then their linear projection becomes the input for the Transformer Layers. The decoder is implemented with a Cascade UPsampler (CUP) which is based on multiple upsampling blocks. Moreover the skip-connections allow feature addition at different resolution level. As far as the CNN is concerned, we adopted ResNet50 as in [5]. We used a patch size of 16x16 in all the experiments. We applied the HOG injection technique also in this architecture following the same steps described in Sec. II-A. In this model the output feature of the encoder, always excluding the batch size, is (*D x H/16 x W/16*), where D is the dimension of the embedding space used for linear projection, which was configured to 768. After HOG injection the new feature dimension is (*N_new x H/16 x W/16*) and *N_new* is equal to 776 in our model.

III. EXPERIMENTAL PROTOCOL

In this section we will outline the experimental setup of our work, illustrating the implemented experiments and the training setting.

A. *Experiments*

We developed different experiments which have mainly two aims: highlighting the improvement of the performance in multi-class segmentation task with HOG injection in the neural networks and searching for the best architecture for our purpose.

- ***Experiment 1*** *(E1): We trained ResNet50-HOG to compare the baseline with the injection technique.*
- ***Experiment 2*** *(E2): We developed DeepLabv3-HOG to evaluate the effect of the atrous convolutions together with feature embedding.*
- ***Experiment 3*** *(E3): In order to compare the difference between the CNN and the Transformer-based models and to assess the impact of injection strategy in this architecture, we trained TransUnet-HOG.*

B. *Training setting*

Each experiment was trained on 1708 images and evaluated on 352 images belonging to Video001, Video006, Video016 of *FetReg Dataset*, which corresponds to fold 1 of baseline results [2]. After assessing the best model, we performed 6-fold cross-validation to verify the robustness of the segmentation algorithm. To be consistent with the *FetReg Challenge* baseline, we assigned the same videos of [2] to each fold and we resized the training images to 448x448 pixels. We applied data augmentation consisting of: random crop of the image with dimension 256x256 pixels, random rotation in range $(-45°, +45°)$, horizontal and vertical flip and random variation in brightness $(-20\%, +20\%)$. The learning rate and batch size were set to 0.001 and 32 for the CNNs and 0.01 and 8 for TransUnet respectively. As common practice, we trained the ResnNet50-HOG and DeepLabv3-HOG with pre-trained ImageNet [8] weights initialization to boost the model performance. For the experiments, training was performed for 300 epochs.

The final weights submitted to the *FetReg Challenge* are obtained training the best model over all the 2060 annotated frames, following the same described setup, for 700 epochs. The networks were trained with two 32 GB of RAM and NVIDIA Tesla V100 GPU.

C. *Evaluation strategy*

In the evaluation phase, the images were cropped in patches of dimension 256x256. The entire frames were then reconstructed by overlapping the patches with stride equal to 8. To compare the results of this study with the baseline, we evaluated the performance of our models with the Mean Intersection Over Union (*IoU*), using the confusion matrix

TABLE I: Results of segmentation over Video001, Video006, Video 016 (fold 1) of *FetReg Task 1* dataset. The average Mean *IoU* per class for each architecture are reported. In bold the highest values are highlighted. Key: BG-background.

Model	**Class**				**Overall**
	BG	**Vessel**	**Tool**	**Fetus**	**fold 1**
Baseline	0.80	0.83	0.64	0.74	0.61
DeepLabv3-HOG	0.80	0.81	0.70	0.76	0.64
TransUnet-HOG	0.78	0.81	0.61	0.71	0.62
ResNet50-HOG	**0.82**	0.83	**0.74**	**0.80**	**0.69**

TABLE II: Results of segmentation obtained from 6-fold cross-validation on ResNet50-HOG model in *FetReg Task 1* dataset. The Mean *IoU* per class over each video are reported. The average values are shown in the right part of the table. In bold values better than baseline are highlighted. Key: BG-background.

Video	**Class**				**Overall**	**Fold**	**Images**	**Class**				**Overall**
	BG	**Vessel**	**Tool**	**Fetus**	**per video**		**per fold**	**BG**	**Vessel**	**Tool**	**Fetus**	**per fold**
Video 001	**0.85**	**0.86**	**0.81**	**0.77**	**0.72**	1	352	**0.82**	0.83	**0.74**	**0.80**	**0.69**
Video 006	0.67	0.66	**0.83**	0.72	**0.61**							
Video 016	**0.82**	0.83	**0.74**	**0.80**	**0.69**							
Video 002	0.76	0.77	0.79	0.51	0.53	2	353	0.80	0.81	**0.84**	0.73	0.66
Video 011	0.73	**0.73**	0.73	0.75	0.61							
Video 018	0.80	0.81	**0.84**	0.73	0.69							
Video 004	**0.83**	**0.81**	**0.85**	**0.91**	**0.79**	3	349	0.76	**0.79**	0.71	0.54	0.62
Video 019	0.80	0.81	0.52	0.80	0.56							
Video 023	0.76	**0.79**	0.71	0.54	0.52							
Video 003	0.74	0.77	0.59	0.75	0.55	4	327	0.80	0.80	0.68	0.80	0.57
Video 005	0.70	0.76	**0.80**	**0.57**	0.56							
Video 014	0.80	0.80	0.68	0.80	0.64							
Video 007	0.70	0.71	0.78	0.59	0.54	5	350	**0.81**	**0.82**	**0.87**	0.51	0.61
Video 008	0.75	0.76	0.64	0.84	0.62							
Video 022	**0.81**	**0.82**	**0.87**	0.51	0.60							
Video 009	0.66	0.69	0.51	0.55	0.37	6	329	**0.71**	**0.71**	**0.85**	**0.58**	0.41
Video 013	0.67	0.73	0.55	0.49	0.39							
Video 017	**0.71**	**0.71**	**0.85**	**0.58**	**0.57**							
per class	0.68	0.71	0.57	0.53								

(CM) to evaluate the accuracy of each pixel. In particular, we used a CM with 2 classes for each category and a CM with 4 classes for the each overall video and fold. The IoU was computed as reported in equation 3, where the *IoUv* is a vector obtained by the equation 2 in which D is the CM's diagonal, r the number of rows and c the number of columns.

$$IoU_v = \frac{D}{\sum_{i=1}^{r}\sum_{j=1}^{c} CM_{ij} + \sum_{j=1}^{c}\sum_{i=1}^{r} CM_{ji} - D} \quad (2)$$

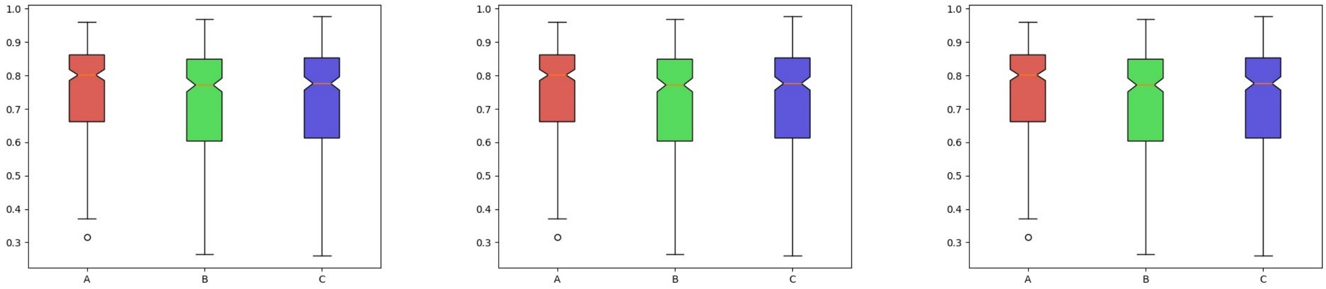
$$IoU = mean(IoU_v) \quad (3)$$

IV. RESULTS

For the experiments 1, 2, 3, we reported in Tab.I, the average values of the Mean *IoU* per class for each architecture. ResNet50-HOG is the most performing model, achieving a significant improvement with respect to the baseline of 10% and of 6% in Tool and Fetus class respectively and of 8% in the overall average of the metric. Also DeepLabv3-HOG shows a progress in these classes of 6% for Tool and of 2% for Fetus. On the contrary, TransUnet-HOG slightly decreases the performance of the baseline, in terms of Mean *IoU*. This is confirmed also by boxplots in Fig.3, in which it can be noticed also a lower dispersion in ResNet-HOG model. Since ResNet50-HOG reaches the highest values of the metric, we implemented the cross-fold validation for this model. Results are presented in Tab.II. In bold are marked the values which overcome the baseline [2]. The first fold reaches the highest performances. However, we can see an improvement also for fold 6 of 5% and 12% in Vessel and Tool class respectively. Vessel class achieves also better performance in fold 3 and 5.

V. DISCUSSIONS AND CONCLUSIONS

We proposed a novel approach for multi-class segmentation, combining an hand-crafted features extractor, as Histogram of Oriented Gradients, with two types of ANN: traditional CNN (ResNet50 and DeepLabv3) and Transformer-based network (TransUnet). The evaluation was implemented on a new dataset of 2060 manually annotated fetoscopic frames presented at *FetReg Challenge*. From Tab.I and Fig.3 we derive that ResNet50-HOG achieves the highest values of Mean *IoU* with respect to the other models. DeepLabv3-HOG shows comparable results to ResNet50-HOG but this is to be expected since the two architectures share the same backbone. The cross-fold validation on ResNet50-HOG model, which results are reported in Tab. II, shows that fold 1 reaches the best performance. The lowest values of the metrics are in fold 4, in which the images of Video005 present negative factors affecting the vessels identification: huge variability in the levels and colors of illumination, high presence of laser light and ablation white spots. However, there is an overall improvement in Tool and Fetus classes. This is the effect of HOG injection, which is particularly



(a) *Vessel* (b) *Tool* (c) *Fetus*

Fig. 3: Performance metric Mean *IoU* obtained testing ResNet50-HOG (1), TransUnet-HOG (2), DeepLabv3-HOG (3), over Video001, Video006, Video016 (fold 1), per Vessel (a), Tool (b) and Fetus (c) classes.

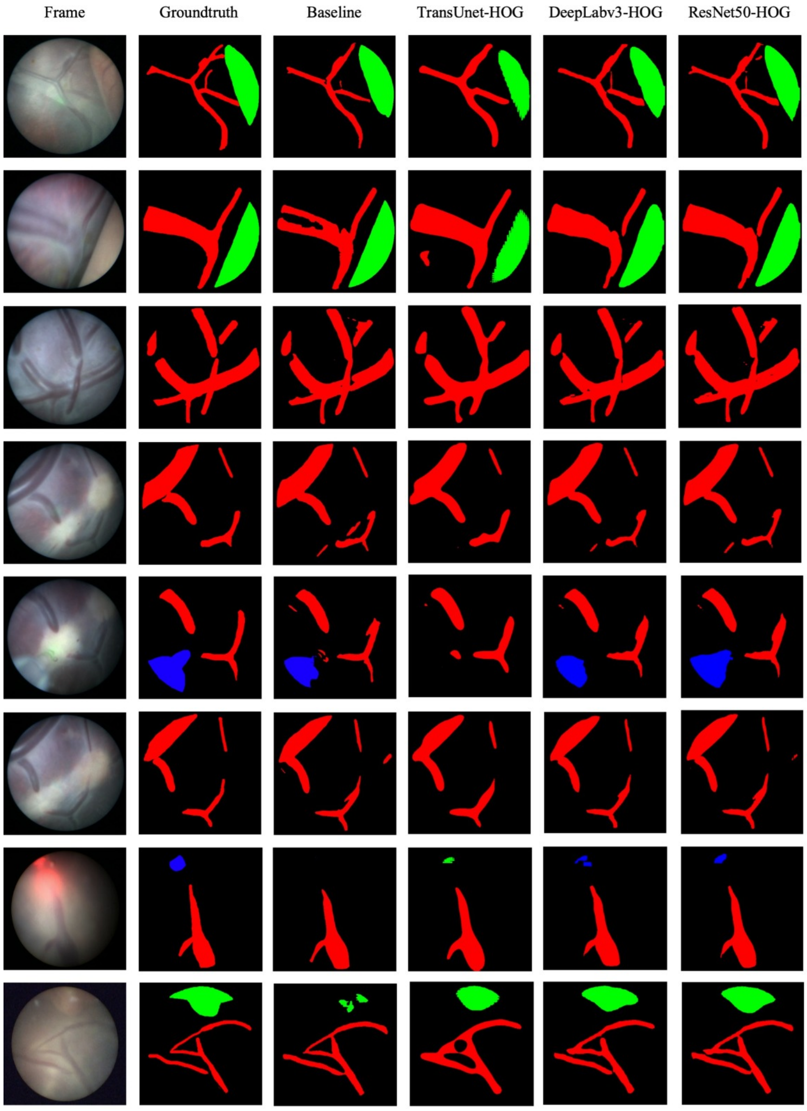


Fig. 4: Comparison segmentation results for Video001, Video006, Video016, with Baseline, TransUnet-HOG, DeepLabv3-HOG and ResNet50-HOG models.

efficient in detecting edges of these targets. In Fig.1 some examples of HOG application are illustrated. Vessels are not distinctly identified, while we can recognize the shape in the HOG representations of Tool and Fetus. Observing qualitatively the predictions in Fig.4, all the models identify vessels. TransUnet-HOG does not recognize thin vessels and it does not detect the Tool, as in the fifth and eighth prediction from top. This is coherent with the metric values. However the vessel edges are defined very sharply with respect to the other models. On the contrary, DeepLabv3-HOG and especially ResNet50-HOG provide quite accurate Tool and Fetus segmentation as in the fifth, seventh and last images from top. The proposed model ResNet50-HOG recognizes a lower number of false positives and it clearly discriminates thin vessels and branches as in the most of the reported frames in Fig.4. Therefore Transformer based architecture shows the lower performance both in terms of metric and qualitative evaluation, but we reserve to train it with more than 300 epochs. Moreover, in order to improve also ResNet-HOG results, we could try to inject the hand-crafted features in different layers of the network. To conclude, we introduce a new framework for multi-task segmentation of intra-operative fetoscopic frames, which results are robust and promising for further researches.

ACKNOWLEDGMENT

This study was supported by the Department of Advanced Robotics at Fondazione Istituto Italiano di Tecnologia, Genoa (Italy) and by the Department of Electronic and Bioengineering at Politecnico di Milano, Milan (Italy). We gratefully acknowledge the HPC infrastructure and the Support Team at Fondazione Istituto Italiano di Tecnologia for providing computing resources for this study.

REFERENCES

[1] M. R. Arefin, F. Makhmudkhujaev, O. Chae, and J. Kim. Aggregating CNN and HOG features for Real-Time Distracted Driver Detection. In *2019 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–3, 2019.

[2] S. Bano, A. Casella, F. Vasconcelos, S. Moccia, G. Attilakos, R. Wimalasundera, A. L. David, D. Paladini, J. Deprest, E. De Momi, L. S. Mattos, and D. Stoyanov. FetReg: Placental Vessel Segmentation and Registration in Fetoscopy Challenge Dataset. *ArXiv*, 2021.

[3] S. Bano, F. Vasconcelos, L. M. Shepherd, E. Vander Poorten, T. Vercauteren, S. Ourselin, A. L. David, J. Deprest, and D. Stoyanov. Deep Placental Vessel Segmentation for Fetoscopic Mosaicking. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12263 LNCS:763–773, 2020.

[4] C. Bolch, M. Fahey, D. Reddihough, K. Williams, S. Reid, A. Guzys, S. Cole, A. Edwards, A. Fung, R. Hodges, R. Palma-Dias, M. Teoh, and S. Walker. 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(1):1–11, 2018.

[5] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. pages 1–13, 2021.

[6] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 40(04):834–848, apr 2018.

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *ArXiv*, abs/1706.05587, 2017.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE Computer Society, dec 2015.

[10] T. Kobayashi, A. Hidaka, and T. Kurita. Selection of Histograms of Oriented Gradients. pages 598–607, 2008.

[11] S. A. Korkmaz, H. BINOL, A. Akçiçek, and M. F. Korkmaz. Artificial Neural Network by using HOG Features and Linear Discriminant Analysis: HOG_LDA_ANN. *IEEE 15th International Symposium on Intelligent Systems and Informatics*, pages 327–332, 2017.

[12] D. Navneet and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 1063–69 19/05, 2005.

[13] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. may 2015.

[14] P. Sadda, 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(2):227–235, 2019.

[15] H. Samulowitz, F. Nargesian, U. Khurana, E. B. Khalil, and D. Turaga. Learning Feature Engineering for Classification. *Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, 2017.

[16] E. Shelhamer, L. Long, and T. Darrel. Fully Convolutional Networks for Semantic Segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 3431–3440. IEEE Computer Society, 2015.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.