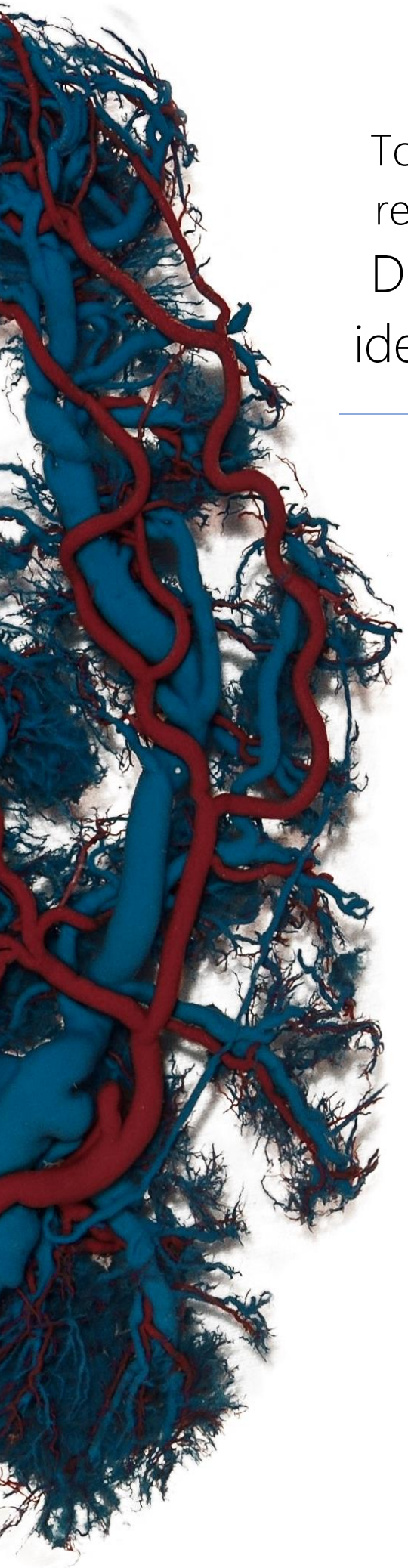


Towards real-time placental surface reconstruction during fetal surgery: Deep-learned placental vessel identification and segmentation

April 16, 2021

Marèll Niekolaas

Radboudumc
UNIVERSITY OF TWENTE.

Towards real-time placental surface reconstruction during fetal surgery: deep-learned placental vessel identification and segmentation

A thesis submitted for the degree of Master of Science
Faculty of Science and Technology
University of Twente, Enschede, The Netherlands

April 16, 2021

Marèll Niekolaas
Master Thesis in Technical Medicine
Department of Gynaecology and Obstetrics, Radboudumc, Nijmegen

Chairman & Technical supervisor
dr. ir. F. van der Heijden
Associate Professor
Department of Robotics and Mechatronics
University of Twente

Medical supervisor
dr. E. Sikkel
Gynecologist
Department of Obstetrics and Gynaecology
Radboudumc

Professional behavior supervisor
drs. R.M. Krol
Department of Science and Technology
University of Twente

Extra member & daily supervisor
drs. A.M. van der Schot
Technical Physician
Department of Obstetrics and Gynaecology
Radboudumc

External Member UT
dr. M. Heijblom
Technical Physician
TechMed Simulation Centre
University of Twente

Photo cover page: J. Bouten en U. Broers, de Levensboom

UNIVERSITY OF TWENTE. Radboudumc

“Practice makes progress.”

– Vince Lombardi

Preface

Alweer een jaar geleden begon ik aan mijn afstudeerstage op de afdeling Verloskunde en Gynaecologie in het Radboudumc, Nijmegen. Nadat ik eerder al met veel plezier een korte stage had gelopen op dezelfde afdeling, begon ik in april 2020 met veel enthousiasme, gedrevenheid en een gezonde dosis leergierigheid aan mijn afstudeerstage.

Terugkijkend op het afgelopen jaar kan ik niet anders dan zeggen dat het een bijzonder jaar is geweest. Naast dat afstuderen *an sich* al vele uitdagingen met zich mee kan brengen, zaten we midden een pandemie. Desondanks kijk ik met veel positiviteit terug op een jaar dat ik vooral als heel leerzaam en waardevol heb ervaren. Ik ben er dan ook van overtuigd dat ik het afgelopen jaar op technisch, medisch én persoonlijk vlak echt sprongen heb gemaakt. En omdat ik dit zeker niet alleen had gekund, wil ik bij deze graag een moment nemen om mijn dank uit te spreken aan iedereen die hierin een bijdrage heeft gehad.

Om te beginnen wil ik graag Anouk bedanken voor de ontzettend fijne begeleiding. Ik vond het erg fijn dat ik bij je terecht kon voor alles, zowel op professioneel als op persoonlijk vlak. Bovendien gaf je mij de vrijheid om mezelf te ontwikkelen op de gebieden waarop ik dat zelf graag wilde. Hierdoor had ik de ruimte om mij te ontwikkelen tot de Technische Geneeskundige die ik ambieer te zijn.

Daarnaast zou ik graag Esther en de afdeling Verloskunde en Gynaecologie in het algemeen willen bedanken voor alle klinische ervaringen die ik op heb mogen doen. Ik voelde me welkom en kreeg de vrijheid om te leren wat ik wilde leren. Niks moest, alles mocht. Tevens wil ik graag Ferdi bedanken voor de begeleiding op technologisch en wetenschappelijk gebied. Je hield me scherp wanneer nodig en hielp mij *the big picture* te zien.

Ruby en Paul, jullie wil ik graag bedanken voor de begeleiding met betrekking tot mijn persoonlijke ontwikkeling. Over je afstuderen wordt vaak gezegd dat dit één van de momenten is waarop je jezelf écht tegen komt. Dit was zeker ook het geval en ik durf te wedden dat dit effect –in positieve zin- nóg sterker is geworden dankzij alle intervisies, reflectiemomenten en de goede begeleiding. Mijn zelfvertrouwen als medisch-technisch professional heeft in een lift gezeten de afgelopen twee jaar en dat heb ik zéker mede te danken aan jullie. Ruby, bedankt dat je me hebt leren voelen, waardoor ik mijn authenticiteit weer terug heb gevonden.

En *last, but certainly not least*, ben ik de mensen in mijn directe omgeving erg dankbaar voor al jullie steun, hulp en luisterende oren gedurende het afgelopen jaar. Vera, Celine en Iris, bedankt voor alle gezelligheid en de fijne sfeer thuis. Bij jullie kon (en kan) ik altijd terecht, of het nou is om een biertje te drinken of om even uit te huilen, de deur staat altijd open.

Sammel, jullie zijn simpelweg geweldig. Anne, bedankt voor alle wandel-belafspraken die we hebben gehad gedurende het jaar. Je hebt me er echt doorheen gesleept. De meiden van mijn jaarclub, en in het bijzonder Rianne, Meike en Diantha, ontzettend bedankt voor het meedenken, de openheid en gezelligheid. Pandemie of niet, met jullie is het altijd genieten.

Tot slot wil ik graag mijn ouders en Ruben bedanken. Jullie staan altijd achter mij, wat ik ook doe of welke keuze ik ook maak. Bedank voor jullie onvoorwaardelijke steun en vertrouwen in mij.

Marèll Niekolaas
16 april 2021

Summary

Introduction – Twin-To-Twin Transfusion Syndrome (TTTS) is a condition that occurs in monochorionic twin pregnancies and is characterized by a disbalanced blood supply between the two fetuses. When left untreated, TTTS is associated with a mortality rate of 90-95%. To date, Fetoscopic Lasercoagulation Of Vascular Anastomoses (FLOVA) is the only treatment option for TTTS that addresses the underlying pathology.

One of the main drawbacks of FLOVA is the limited Field of View (FOV). Therefore, providing the surgeon with an overview of the placental vasculature is thought to increase the success rates of FLOVA. Through the recent years, more and more research has been conducted on reconstruction of the placental surface using image stitching algorithms. However, to date, none of these approaches were proven successful when applied to (longer) *in-vivo* video sequences.

The *in-vivo* fetoscopic videos contain numerous frames that are either irrelevant or disruptive for image stitching. Therefore, the first part of this thesis focused on automatic classification of frames that are suitable for image stitching using a deep learning approach. The second part focused on the training and potential use of a vessel segmentation network. We hypothesized that the resulting segmentation maps can be used for 1) improved inlier feature detection by using selective regional image enhancement and 2) intensity-based image stitching.

Methods – Ten *in-vivo* fetoscopic videos from FLOVA procedures were included in this thesis. First, the effect of the frame content on the number and quality of the detected inlier feature matches is evaluated. Thereafter, a total of 62,422 labeled frames were extracted and labeled. A pre-trained CNN with a VGG-16 architecture was trained for binary classification of the *in-vivo* video frames. For the vessel segmentation network (VesSeg), a U-Net was trained using 729 *in-vivo* frames and ground truth vessel segmentations. Lastly, the potential use of the vessel segmentations for both feature-based and intensity-based image stitching was briefly explored.

Results & Discussion – Frames in which the vessels are visible without any occlusions were most suitable for image stitching, followed by frames that show vessels that are partly occluded. These frames were therefore labeled as *suitable* for the vessel identification network. The trained vessel identification network (VesDet) generated predictions with an ROC-AUC of 0.95 when tested using an unseen video. A prediction rate of 714 fps was reported when using Google Colab's GPU. Based on these results, the network is considered useful and applicable for future clinical implementation.

Our best performing U-Net generated vessel segmentations with a Dice Score of 0.80 (± 0.13) and ROC-AUC of 0.98. In literature, two studies proposed similar networks and reported Dice Scores of 0.55 (± 0.22) and 0.78 (± 0.13) for their best networks. Therefore, our network significantly outperformed the network from the first study and slightly outperformed the network from the second study described in literature. Additional qualitative analysis supported these findings. Moreover, an average prediction rate of 7 fps was measured, which is considered sufficient for future clinical applications of the network.

Lastly, experimentations with VesSeg for multiple different feature-based and intensity-based image stitching approaches showed an increase in the number of frames that were stitched together successfully. However, systematic research on the different image stitching approaches is highly recommended.

Conclusion – Based on the studies and experimentations performed in this thesis, we conclude that the vessel identification and segmentation deep learning networks are of added value for image stitching of *in-vivo* fetoscopic video frames. Moreover, the networks are considered suitable for clinical applications based on their high performance when tested using unseen *in-vivo* data and fast prediction rates. However, the image stitching algorithm requires further development before it can be used in clinical settings.

List of Abbreviations

AI	Artificial Intelligence
AUC	Area Under Curve
BRIEF	Binary Robust Independent Elementary Feature
CLAHE	Contrast-Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
DL	Deep Learning
EM	Electromagnetic
FAST	Features from Accelerated and Segments Test
FCN	Fully Convolutional Network
FCNN	Fully-Connected Neural Network
FLOVA	Fetoscopic Laser Occlusion of Vascular Anastomoses
FOV	Field of View
GA	Gestational Age
GPU	Graphics Processing Unit
GT	Ground Truth
IFM	Inlier Feature Matches
LSTM-RNN	Long Short-Term Memory Recurrent Neural Network
MDCNN	Multiple Deep Convolutional Neural Networks
MSAC	The M-estimator SAmple Consensus
ORB	Oriented FAST and rotated BRIEF
PPROM	Preterm Prelabor Rupture Of Membranes
ROC	Receiver Operating Characteristics
ROI	Region of Interest
RSGD	Regular Step Gradient Descent
SLAM	Simultaneous Localization and Mapping
SRIE	Selective Regional Image Enhancement
TTTS	Twin-to-Twin Transfusion Syndrom
VesDet	Vessel Detection (classification) network
VesSeg	Vessel Segmentation network

Table of Contents

1. General introduction

1.1 Twin-to-Twin Transfusion syndrome

Twin-To-Twin Transfusion Syndrome (TTTS) is a condition that occurs in monochorionic twin pregnancies and is characterized by a disbalanced blood supply between the two fetuses. The blood is disproportionately redirected from one fetus (the donor) to the other fetus (the recipient) through vascular anastomoses.[1,2] Monochorionic twin pregnancies account for 3 in 1000 of all deliveries worldwide.[3] It is estimated that TTTS occurs in 10-15% of these pregnancies.[1,4]

TTTS can have serious consequences for both twins, including 1) severe hypotension and heart failure in the recipient twin, 2) pulmonary hypoplasia and permanent brain damage in the donor twin and 3) eventually death in one or both twins.[2,5] When left untreated, TTTS is associated with a mortality rate of approximately 90-95%.[6,7]

1.2 Fetoscopic Laser Coagulation

To date, Fetoscopic Lasercoagulation Of Vascular Anastomoses (FLOVA) is the only definitive treatment option for TTTS. This treatment aims to interrupt the undesired blood transfusion between both fetuses.[8,9] During FLOVA, the placental surface is visually inspected to identify the vascular anastomoses.[10,11] This is done using a fetoscope, which is a specialized endoscope with a relatively small outer diameter (1.0-3.8 mm).[12] Thereafter, the anastomoses are coagulated with a built-in laser (Figure 1).

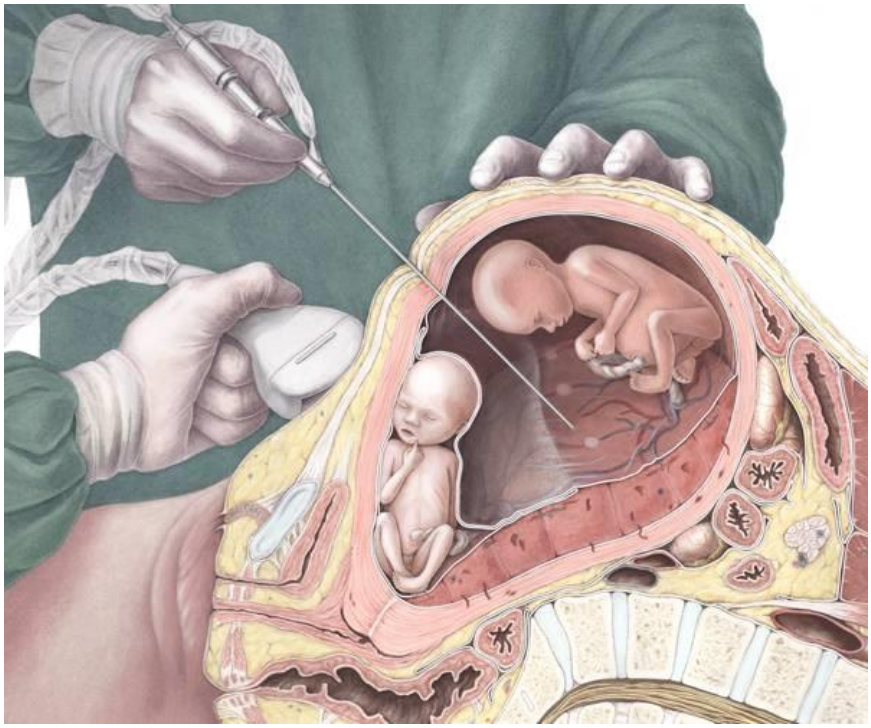


Figure 1. Illustration of the Fetoscopic Lasercoagulation of placental Vascular Anastomoses (FLOVA) procedure in a TTTS pregnancy.[13]

The small diameter of the fetoscope is necessary to reduce the risk of preterm prelabor rupture of membranes (PPROM) and thus reduce the risk of preterm birth.[14] However, smaller diameter instruments demand technical compromises, which affects the overall image quality and the size of the field of view (FOV). As a result, inspection of the placental surface and identification of the vascular anastomoses can be challenging.

Akkersmans et al. published a systematic review on the outcomes of 3,868 FLOVA procedures performed between 1990 and 2014. Based on the most recent cases (2011-2014), they described a benchmark for perinatal survival rates after FLOVA of 52–54% for both twins and 81– 88% for at least one twin, with a mean gestational age (GA) of 32.4 weeks.[15] Although the survival rates of one or both twins after FLOVA has increased since its first introduction[15,16], there is still scope for improvement.

Providing the surgeon with an overview of the placenta during FLOVA is thought to be of great value for faster navigation and identification of the anastomoses. This reduces the operation duration, which is generally associated with a reduction of postoperative complications and mortality rates.[17,18] Moreover, an overview of the placental vasculature is thought to help determine the appropriate order in which the anastomoses should be coagulated. Unfortunately, to date, no imaging modality is able to generate an overview of the placental vasculature pre- or intraperative.[19] A potential solution could be to artificially increase the FOV during FLOVA, using computer vision techniques. To do so, the individual fetoscopic video frames are used to generate a two- or three-dimensional (2D or 3D) reconstruction of the placental surface.

1.3 Previous Research

In literature, various attempts have been reported on intraoperative, real-time mosaicking of fetoscopic video sequences in order to increase the FOV artificially.[20–28] Most of the techniques and algorithms proposed in these studies have shown promising results when tested on placenta phantoms or *ex-vivo* (dye-injected) placentas. However, none of these approaches were proven successful when applied to (longer) *in-vivo* video sequences.[25]

Other studies proposed the use of additional imaging modalities, such as MRI[29] or photoacoustic imaging[30], in an attempt to generate an overview of the placental surface preoperatively. Moreover, intraoperative use of external electromagnetic (EM) or visual tracking devices to improve the camera position estimation has been reported.[31–33] The main drawback of these approaches is that they require (significant) changes to the current FLOVA workflow, either pre- or intraoperatively. Since our aim is to generate a placental overview without impacting the current clinical workflow, these approaches were excluded in our study. Furthermore, most of these studies were limited to simulation settings and *ex-vivo* placentas, meaning that they were not proven effective in *in-vivo* settings as well.

1.3.1 Image Stitching

One popular approach for generating an overview of the placental surface is image stitching. This technique combines multiple images with overlapping content to form a reconstruction or map. In

the field of fetoscopic surgery, most research focusses on feature-based image stitching.[21–24,34–36] Features can be described as recognizable regions in an image, such as corners. A team from the Department of Obstetrics & Gynaecology at the Radboudumc has been working on the so-called 3D FLOVA-SLAM technique, which is a feature-based image stitching approach as well. This technique involves the ORB-SLAM2 algorithm, an open-source algorithm that uses ORB feature detection combined with Simultaneous Localization And Mapping (SLAM)[37].

Hitherto, the proposed feature-based image stitching approaches showed promising results when tested using *ex-vivo* placentas and phantoms. Unfortunately, their performance drastically decreased when applied to *in-vivo* data. As a result, the proposed feature-based image stitching techniques have not been able to perform image stitching of (longer) *in-vivo* fetoscopic video sequences. One of the complications is that these algorithms are sensitive to drift.[32]

Feature-based image stitching of *in-vivo* fetoscopic video frames is limited due to multiple reasons, including 1) the lack of texture on the placental surface, 2) the poor overall image quality and 3) the presence of unstable features. The latter is caused by floating amniotic fluid particles (fetal skin flakes), fetal parts and the umbilical cord.[26,38] These dynamic features are problematic since feature-based stitching relies on stable features. Moreover, the poor image quality is partly the result of the poor lighting conditions in the intrauterine environment.[25,35,39]

An alternative to feature-based image stitching is intensity-based image stitching, also known as direct image registration or alignment. In 2018, Peter et al.[26] proposed the first approach for intensity-based image stitching of *in-vivo* fetoscopic video sequences. Although they demonstrated promising results, no further research is published on this topic yet.

Lastly, another factor that makes the transition from *ex-vivo* to *in-vivo* challenging is the fact that most studies rely on videos from previously performed FLOVA procedures. These videos were recorded without the intention to generate an overview. As a result, the scope moves relatively fast or chaotic in some frames, making retrospective image stitching challenging. In contrast, when testing an image stitching algorithm using a test setup, the operator can adjust the scope's movements based on the progress of the real-time generated reconstruction. For example, when the signal gets lost, the surgeon will return to the previous position and reduce the speed of the scope's movements. Test setups are therefore more useful for development of real-time stitching algorithms. However, the video frames from test setups appear visually different from *in-vivo* fetoscopic video frames, making the translation from *ex-vivo* to *in-vivo* applications difficult.[25]

1.3.2 Deep Learning

Through the recent years, the number of studies that use artificial intelligence (AI) approaches to generate an overview of the placental vasculature is rising. This is thought to be related to the high variability of image appearances between different fetoscopic frames and videos[40], making it challenging to develop *one size fits all* algorithms. In previous studies, deep learning (DL) networks have been developed and tested for improved homography matrix estimation[41], placental pose estimation[28], and stable feature detection[25]. Additionally, DL networks have been trained for other

purposes in the field of fetoscopic surgery, including occlusion identification[38] and segmentation of the inter-fetus membrane[42] or placental vessels[19].

Bano et al. trained a DL network to identify four different fetoscopic events, including *clear view*.[38] We hypothesize that a similar approach can be used for binary identification (classification) of frames that are suitable for image stitching. This network can then be used to exclude frames that are less suitable or disruptive during image stitching, increasing the likelihood of successful image stitching of (longer) *in-vivo* video sequences.

1.4 Goals and Thesis Outline

Providing the surgeon with an overview of the placental vasculature during FLOVA is thought to increase the success rates of FLOVA, increasing the survival rates of one or both twins. Although more and more research has been conducted on the topic of real-time image stitching, no solution is described in literature for (longer) *in-vivo* fetoscopic video sequences yet. We hypothesize that DL networks can be trained to overcome some of the problems related to development of an image stitching algorithm, such as the presence of useless or disturbing frames. Therefore, the aim of this thesis is to evaluate the potential use of DL approaches for image stitching of fetoscopic video frames, focusing explicitly on *in-vivo* data.

One of the main challenges related to working with *in-vivo* data is the frame content: besides frames that show the placental vessels, a significant portion of the frames show other structures, including fetal body parts and the umbilical cord. Furthermore, the amniotic fluid contains fetal skin flakes, which can negatively affect the performance of image stitching techniques. Figure 2 shows a series of example frames from the orientation phase of a FLOVA procedure performed in the Radboudumc. Since not all frames are suitable for image stitching, we first focus on identifying non-occluded frames with vessels using a DL approach. Therefore, the first sub-goal is to train a DL network for binary classification of *in-vivo* fetoscopic video frames.

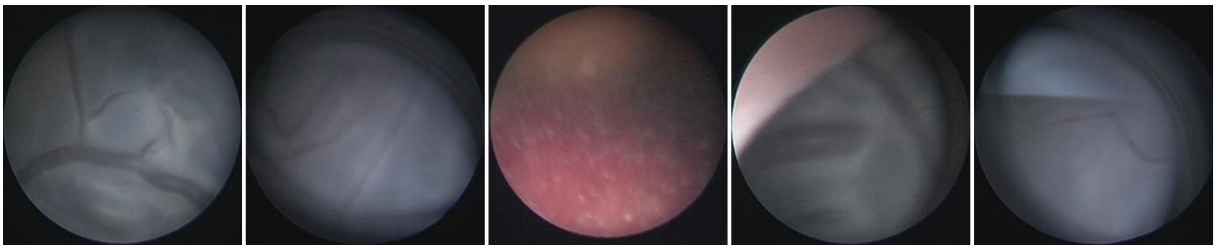


Figure 2. Example frames demonstrating the differences in frame content from a FLOVA procedure performed in the Radboudumc. All frames originate from the same video.

One of the main drawbacks of feature-based image stitching is the fact that some frames result in an inadequate number of matchable features. Although this is mainly seen in frames that show little to no vascular structures, it is also seen in frames that show 'clearly visible' vessels. We hypothesize that a solution can be found in semantic segmentation of the vessels through either 1) using the segmentation maps for selective regional image enhancement or 2) using the segmentations for intensity-based image stitching.

The added value of the first approach has been demonstrated in an earlier project (*M2-3*). This project demonstrated that segmentation maps of the vascular structures can be used to increase the number of inlier feature matches. This was done by using the segmentation maps to separate the vessels from the 'background' and performing different pre-processing steps to the separated parts. The report from this project is available on request. Hence, the second sub-goal of this thesis is to train a DL network for vessel segmentation of *in-vivo* fetoscopic video frames.

To summarize, this thesis first focuses on training a DL network for binary image classification. The aim of this network is to identify frames with vessels, since these are considered more suitable for image stitching. Thereafter, a second DL network will be trained for vessel segmentation. Lastly, the different uses of the segmentation maps for image stitching will be briefly explored. Figure 3 shows a schematic overview of the goals of the trained DL networks in the process of generating a reconstruction of the placental surface.

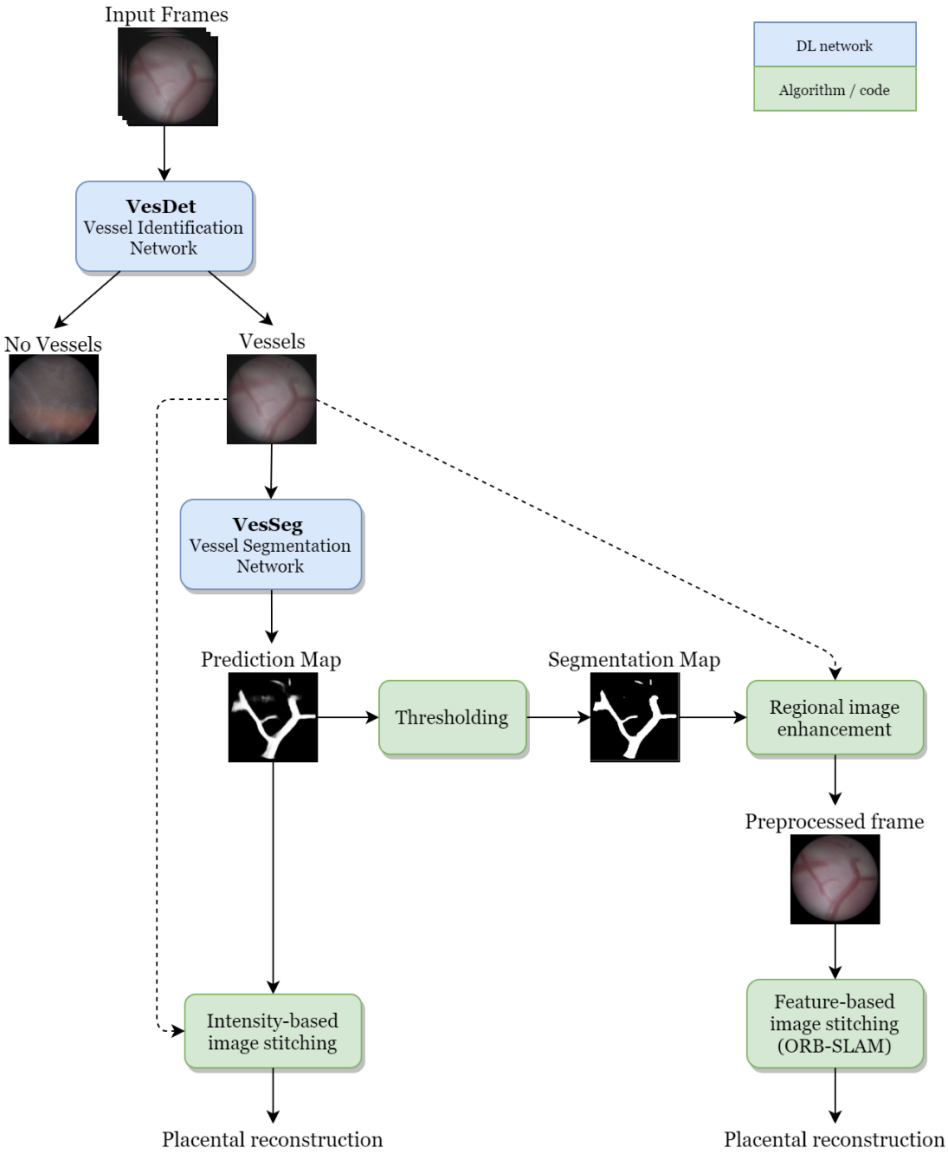


Figure 3. Overview of the steps taken in this thesis for generating a reconstruction of the placental surface of in-vivo fetoscopic video sequences, using two deep learning networks (blue).

1.4.1 Research Questions

The following research sub-questions were drafted:

1. What are the main hindrances when attempting image stitching of *in-vivo* fetoscopic video frames and what are potential solutions?
    a. What solutions are proposed in literature?
    b. What AI solutions are proposed in literature and what is their potential?
2. How does the frame content (e.g. vessels, fetal parts) affect the number and quality of inlier feature matches (IFM) found in two consecutive frames?
3. To what extent can a DL network be trained for binary classification of *in-vivo* video frames to identify frames with vessels / frames that are useful for image stitching?
    a. What is the performance of this network when tested using an unseen FLOVA video?
    b. How fast can the network generate its predictions?
4. To what extent can a DL network be trained for vessel segmentation of *in-vivo* fetoscopic video frames with vessels?
    a. What is the performance of this network (Dice Score) when tested using unseen frames?
    b. How fast can the network perform its predictions?
5. What application of the segmentation network has more potential?
    a. How many consecutive *in-vivo* frames can be successfully stitched by using the segmentation maps for selective regional image enhancement?
    b. How many consecutive *in-vivo* frames can be successfully stitched by using the prediction maps for intensity-based image stitching?
6. What is needed before the DL networks can be used in clinical applications and what are the potential uses of the networks?
7. *How can a priori knowledge on the geometrical shape of the placenta be used for improved image stitching?*

1.4.2 Thesis Outline

First, background information on feature detection and deep learning is provided (Chapter 2), followed by an overview of the characteristics of the data used in this thesis (Chapter 3). Thereafter, the effect of the video content on the feature detection is evaluated (Chapter 4). Based on the findings from this chapter, the data was labeled for the classification network. Then, the classification and segmentation DL networks were developed and their performance was evaluated (Chapter 5 & 6). The potential uses of the segmentation network for feature-based and intensity-based image stitching was briefly explored (Chapter 7).

2. Background Information

2.1 Feature Detection

An image feature can be described as a small patch of information composed of a feature keypoint, which is the 2D position of the patch, and a feature descriptor, which can be described as a visual description of the patch. One frequently used feature detector is ORB, since it is known for its speed and robustness.[43,44]

ORB stands for Oriented FAST and Rotated BRIEF feature detection.[45] FAST, or Features from Accelerated and Segments Test, is a method for real-time feature detection. It was developed for corner detection and is relatively computational efficient.[46] FAST works as follows: the pixels in a circle around a pixel *p* are sorted as lighter than, darker than or similar to *p*. If more than fifty percent of these pixels are darker or brighter than *p*, that pixel is selected as a keypoint (Figure 4).[45] In ORB, FAST is made partial scale invariant by augmenting it with multiscale image pyramids.[43]

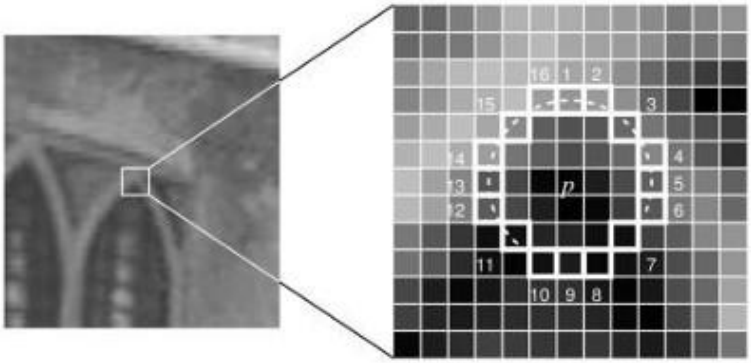


Figure 4. FAST keypoint detection (directly taken from Rosten and Drummond[46]).

After the feature points are detected using FAST, Binary Robust Independent Elementary Feature (BRIEF) calculates a binary descriptor for each feature point. BRIEF uses simple binary tests between pixels in a smoothed image region.[47]

2.2 Deep Learning

Deep learning (DL) is a sub-field of machine learning, which is in turn a form of Artificial Intelligence (AI). DL is inspired by the neural network of the human brain and by the way it learns, which is learning by experience. An artificial neural network exists of an input layer, multiple hidden layers and an output layer. The word *deep* refers to the multiple hidden layers that make up the structure of a DL neural network (Figure 5).

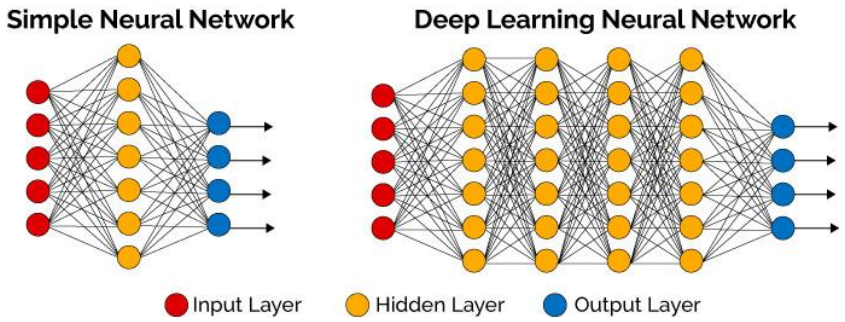


Figure 5. Illustration of a simple neural network versus a deep learning neural network.[48]

An example of a deep learning task could be classification of images of cats and dogs. When training a DL network, the network is *not* told what the distinguishable features are (such as the shape of the ears). Instead, it is rewarded or 'punished' when giving a correct or incorrect prediction, respectively. After multiple training sessions, the network learns what features are relevant and how they can lead to correct classification of the input images.

A DL neural network can be described as a large mathematical function with trainable parameters. At the start of training, these parameters are random numbers and the network generates random predictions. While training a network, these parameters are repeatedly altered to make the prediction match the desired output. While training a network, the performance of the network is measured by a loss function, also known as cost function. This function compares the prediction with the desired output and quantifies the error by a real number. When training a network, the goal is to minimize the loss function or *cost* of each repetition.

2.2.1 Supervised Learning & Transfer Learning

The abovementioned example of images classification is an example of supervised learning: the data is labeled prior to training and the network's performance is measured by comparing its prediction with the desired output. This is the first introduced and most widely used training method.[49] Other methods are reinforcement and unsupervised learning, which do not require labeled data. In reinforcement learning the data is not labeled, but instead feedback is given during training that a prediction was correct or incorrect without telling what the actual label is. Unsupervised learning refers to learning by clustering or grouping the data.

Transfer learning is a form of supervised learning that can be described as the reuse of a pre-trained network for a new task. The general idea is that the basic skills of the pre-trained, such as the ability to recognize patterns, are adopted. Thus, instead of training a network from scratch, the pre-trained weights of trained network are used as a starting point for training a new network. This leads to early convergence with better performance compared to training from scratch.[50] As a result, transfer learning approaches require comparatively little data.[51]

2.2.2 Convolutional Neural Network

A Convolutional Neural Network (CNN or ConvNet) is a class of DL networks that performs convolutional operations in some of the hidden layers. CNNs are commonly applied when working with images, such as classification or segmentation tasks.[49] The convolutional layers help to recognize spatially relevant features, such as shapes and textures.

In contrast to a fully-connected network (Figure 4), the convolutional layers are not fully-connected. In other words, the neurons in one layer are only connected to a few locally nearby neurons in the second layer (Figure 6). This reduces the cost in memory (weights) and computations (connections).

Figure 6 shows the general architecture of a CNN for image classification purposes. Here, the *feature learning* part, which contains convolutional layers, is followed by a *classification* part, which contains fully-connected layers.

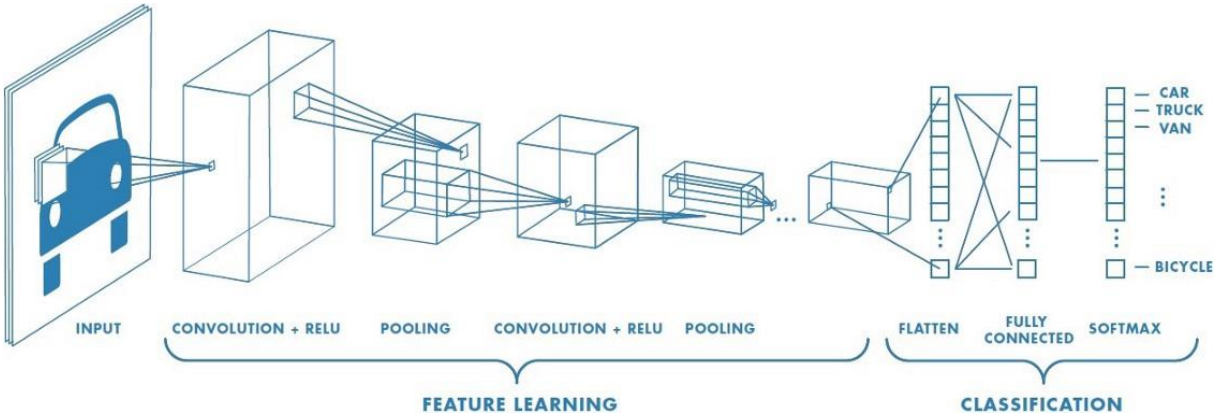


Figure 6. The general architecture of a convolutional neural network (CNN) for image classification.[52]

2.2.3 U-Net

U-Net is a DL network with a specific architecture, giving it a U-shape. It was first introduced by the computer science department of the University of Freiburg, Germany for a biomedical image segmentation task.[53]

U-Net is a fully convolutional network (FCN), meaning that the network only contains convolutional layers and thus no fully connected layers. U-Net consists of a contracting path with convolutional layers (common CNN) and an expansive path (Figure 7). In the expansive path, upsampling takes place, followed by deconvolution or "up-convolution". One of the key contributions of the U-Net architecture is the information transfer from the contracting path to the expansive path, which is illustrated by the gray arrows in Figure 7.[19] This helps to retain detailed information on the images that may have gotten lost during the max pooling operations.

Lastly, U-Net is known for its robustness and requiring relatively little training data.[53] According to Zhou et al.[54], U-Net is the most widely used encoder-decoder network for segmentation tasks in the field of medical imaging.

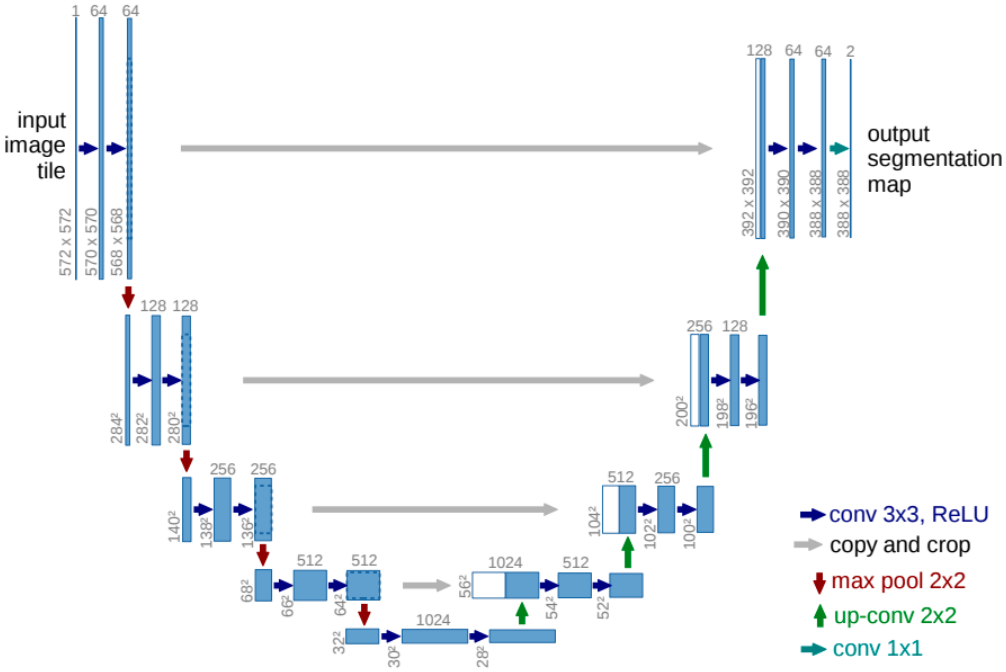


Figure 7. U-Net architecture as introduced by Ronneberger, Fischer and Brox.[53] *The blue boxes represent multi-channel feature maps. The value on top of each box corresponds to the number of channels. The values on the left of the boxes refer to the height and width of the matrices (images). The feature maps from the contracting path are cropped and copied to the expansive path (gray arrows; white boxes). Lastly, the remaining arrows represent different operations, including convolutional and max pooling operations.*

3. Data Acquisition

Ten *in-vivo* videos from nine FLOVA procedures were included in this thesis. The CMO has approved the use of these videos for research and development of the placental reconstruction algorithm. These FLOVA procedures were performed in the Radboudumc between 2018 and 2019. The GA at time of FLOVA varied from 15.2 to 27.1 weeks, with a mean GA of 21.4 weeks.

The characteristics of the videos can be found in Table I. All videos were recorded using Straight Forward Telescopes (Karl Storz, Tuttlingen, German). More specifically, a 2.0 mm HOPKINS® II Straight Forward Telescope 0° (26008AA) and a 2.9 mm HOPKINS® II Straight Forward Telescope 30° (26120BA) were used for posterior and anterior placentas, respectively. Video 8 and 9 were recorded during the same FLOVA procedure, using different fetoscopes. Video 9 was recorded using a 1.3 mm Miniature Straight Forward Telescope 0° (11540AA).

Only the frames from the so-called orientation phase were extracted from the videos, which is the first phase of the FLOVA procedure. During the orientation phase, the placental surface and vascular structures are visually scanned to identify the anastomoses. The frames were extracted using MATLAB R2020b. Thereafter, the frames were cropped to make the circular FOV from the fetoscope fully fit the frame. The frames were stored twice: a full resolution version stored as .png file and a compromised version, meaning that the frame was downsized to 256 x 256 x 3 pixels and stored as .jpg file. The latter will be used for the majority of the studies performed in this thesis to reduce computational time. The .png files will be used for image stitching (Chapter 7).

Table I. Characteristics of the in-vivo fetoscopic videos used in this study

Fetoscopic video #	Placenta location	Frame dimensions	Frames per second	Number of frames
1	Posterior	576 x 720	25	3,511
2	Anterior	576 x 720	25	1,995
3	Posterior	1080 x 1920	25	1,938
4	Anterior	1080 x 1920	60	775
5	Anterior	1080 x 1920	50	12,678
6	Posterior	1080 x 1920	25	5,308
7	Posterior/ Anterior	480 x 640	30	23,866
8	Anterior	1080 x 1920	25	1,646
9	Anterior	1080 x 1920	25	5,331
10	Anterior	720 x 1280	24	5,374

4. The effect of fetoscopic video content on ORB feature detection

4.1 Introduction

Providing the surgeon with an overview of the placental vasculature is thought to improve the outcomes of the FLOVA procedure in TTTS pregnancies. Feature-based image stitching of the *in-vivo* fetoscopic video sequences is one of the potential solutions for generating an overview of the placenta during FLOVA.

When performing a feature-based stitching technique, the matched feature points of two consecutive frames are used to calculate the geometric transformation between those frames. The larger the number of matched feature points, the higher the accuracy of the geometric transformation estimation. Besides, in order to estimate the geometric transformation matrix for projective transformations, at least four matched feature pairs are required.

In case of *in-vivo* fetoscopic video frames, some frames lead to an inadequate number of stable feature matches. This is thought to be caused by 1) the lack of texture resulting in a lack of (unique) detectable features, 2) the presence of moving structures, such as fetal parts, the umbilical cord and fetal skin flakes in the amniotic fluid, leading to instable features and 3) the poor image quality related to the fetoscope's limited cannula diameter.

Conversely, it is observed through experimentation that certain *in-vivo* fetoscopic video sequences lead to enough stable features, making image stitching feasible. To identify which frames are 'suitable' for image stitching, the effect of the frame content on the number of stable matched features will be analyzed. Therefore, this chapter aims to find out how the frame content affects the number and quality of the feature detection. Eventually, the results from this chapter will be used for appropriate data labeling for a vessel classification network.

We hypothesize that frames with clearly visible vessels are more suitable for image stitching, followed by frames that show vessels that are partly occluded by other structures. The frames with poor image quality or occlusions are thought to lead to the lowest number of stable feature matches.

4.2 Method

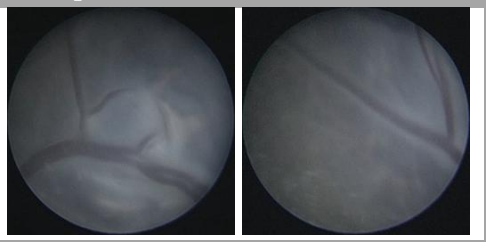
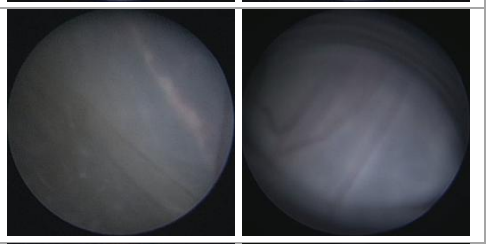
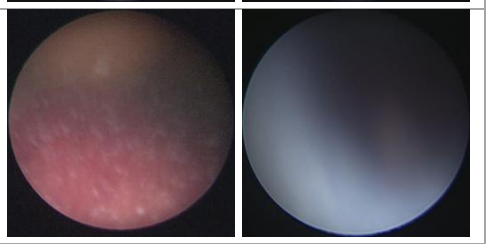
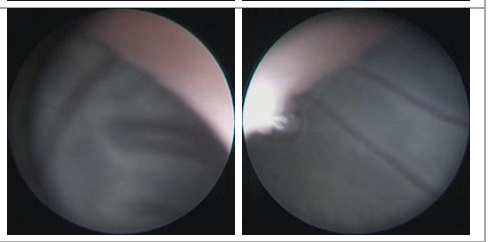
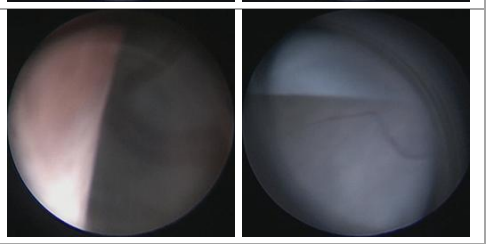
In order to assess the effects of the frame content on the number and quality on the detected features, the frames will be subdivided into five classes based on visual inspection. The number of inlier feature matches (IFM) will be calculated and compared between the classes. Moreover, a qualitative analysis will be executed to provide additional information on the quality or reliability of the found IFM.

In this chapter, ten *in-vivo* fetoscopic videos (Chapter 3, Table I) were used, resulting in a total of 62,422 frames. The frames that were previously downsized to 256 x 256 pixels and stored as .jpg files were used.

4.2.1 Data Labeling

The frames were divided into multiple classes based on visual inspection of the frame content of the *in-vivo* fetoscopic videos. The following five classes were defined: 1. vessels, 2. vessels + bad view, 3. no vessels, 4. partly vessels and 5. partly vessels + bad view. A description and example images for each class can be found in Table II. The frames were manually labeled by a technical medicine student and reviewed by a technical physician, specialized in fetal therapy. The number of frames assigned to each class can be seen in Table III.

Table II. The five classes used for labeling the in-vivo fetoscopic video frames.

Class	Class name	Class description	Example frames*
1	Vessels	Vessels are clearly visible with an acceptable image quality.	
2	Vessels *+ Bad View*	Vessels are visible, but with a 'bad view' (e.g. many fetal skin flakes, low image quality).	
3	No Vessels	No vessels are visible.	
4	Partly Vessels	Vessels are clearly visible, but partly occluded by other structures (e.g. fetal parts, umbilical cord).	
5	Partly Vessels *+ Bad View*	Vessels are visible, but partly occluded by other structures and the view is 'bad' (e.g. many fetal skin flakes, low image quality).	

**all example frames originate from fetoscopic video six.*

Table III. Number of frames per class in, per fetoscopic video.

Fetoscopic Video	Class 1 *Vessels*	Class 2 *Vessels + Bad View*	Class 3 *No Vessels*	Class 4 *Partly Vessels*	Class 5 *Partly Vessels + Bad View*	Total
1	1,104	848	973	284	302	3,511
2	691	184	998	80	42	1,995
3	191	131	995	251	370	1,938
4	105	40	602	12	16	775
5	6,789	2,881	2,095	502	411	12,678
6	1,437	926	2,116	462	367	5,308
7	15,209	836	3,477	3,608	736	23,866
8	155	514	966	3	8	1,646
9	996	1,969	2,141	108	117	5,331
10	1,729	997	2,299	232	117	5,374
All	*28,406*	*9,326*	*16,662*	*5,542*	*2,486*	*62,422*

4.2.2 Preprocessing

For each class, 1016 pairs of two consecutive frames were extracted from the total dataset. This was done using a step size of ~ $\frac{number\ of\ frames}{1016}$ per class to extract the frames evenly from the total dataset. The camera's intrinsic parameters were unknown for most videos. Therefore, the frames were not corrected for lens distortion.

Image enhancement was performed using the multiple steps. First, the *in-vivo* frames (RGB images) were converted to the HSV (Hue, Saturation, Value) color space. Thereafter, contrast enhancement was applied to the Value component to enhance the luminance.[55] This was done using the contrast-limited adaptive histogram equalization (CLAHE)[56] technique. The images were converted to grayscale images and a second contrast enhancement technique was applied: the intensity values of the grayscale images were adjusted to saturate the lowest and highest possible pixel values for each image. In other words: the grayscale values from a specific frame are stretched from the lowest possible value (0) to the highest (255). Image noise was reduced by applying pixelwise adaptive low-pass Wiener filtering with a neighborhood size of 3 x 3 pixels. Lastly, a circular mask was applied to the images to ensure no features are detected outside the FOV.

4.2.3 Feature Detection and Matching

ORB feature detection was performed using seven decomposition levels, each with a scale factor of 1.2. This resulted in an image height and width of $\frac{256\ pixels}{1.2^{(n-1)}}$ for each decomposition level (n). Thereafter, the features were extracted and matched for two consecutive frames. A geometric transformation matrix was estimated using a function that uses the M-estimator SAmple Consensus (MSAC) algorithm to exclude outliers. For the matrix estimation calculations, the following parameters were used: a confidence of 95%, a maximum of 1000 random trials and a maximum

distance of 3 pixels from a point to the projection of its corresponding point. These parameters were chosen based on *trial and error* and kept constant for all calculations.

4.2.4 Quantitative Analysis

For each class, the number of IFM were collected for all 1016 frame pairs plotted using boxplots. IBM SPSS Statistics 25 was used to determine whether the results were normally distributed. Thereafter, a Kruskal Wallis test was performed to determine whether a statistically significant difference was found between any of the classes. Finally, the Mann-Whitney *U* test was performed multiple times for pairwise comparisons between the classes.

4.2.5 Qualitative Analysis

The matched feature points were plotted on the frame pairs for visual inspection. Additional qualitative analysis is relevant because a high number of IFM can be deceitful. It is observed through experimentation that frames that contain fetal skin flakes or fetal body parts can lead to a high number of IFM, which are considered unwanted. These IFM are unwanted because the corresponding transformation matrix is related to the movements of the skin flakes or fetus, instead of the camera movements. Figure 8 shows an example of a fetal body part (fingers) leading to a high number of IFM. Moreover, MSAC might fail to find the appropriate geometric transformation matrix, which leads to a wrong discrimination between inlier and outlier feature matches. Figure 9 demonstrates correct and incorrect selection of inliers by MSAC. The found IFM for all 1016 frame pairs were visually assessed to determine their reliability for each class.

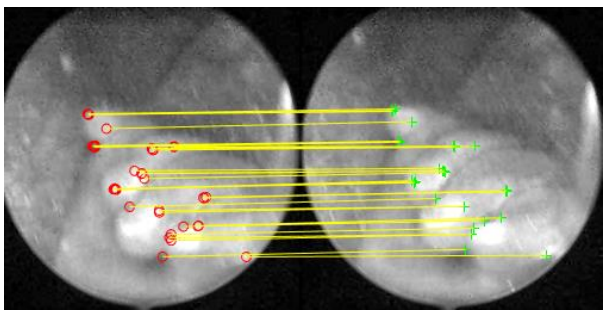


Figure 8. Examples of inlier ORB-feature matches found on a fetal structure (the fingers).

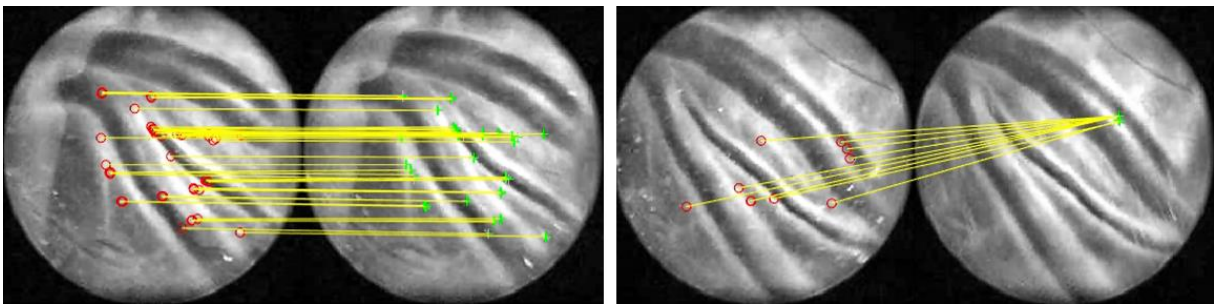


Figure 9. Examples of correct (left) and incorrect (right) selection of inlier ORB-feature matches, using MSAC.

4.3 Results

4.3.1 Quantitative Analysis

The number of inlier ORB-feature matches (IFM) for each class is shown in Table IV. Most IFM were found in the first class (*Vessels*), followed by the fourth class (*Partly Vessels*). Figure 10 shows a boxplot of the number of IFM found per class.

Table IV. Statistics on the number of inlier ORB-feature matches found for each class.

Class #	Class Name	Max.	Average	STD	Median
1	*Vessels*	262	18.42	±26.97	9
2	*Vessels + Bad View*	205	5.82	±10.12	5
3	*No Vessels*	220	7.61	±15.87	5
4	*Partly Vessels*	202	11.87	±14.72	8
5	*Partly Vessels + Bad View*	341	9.92	±24.89	5

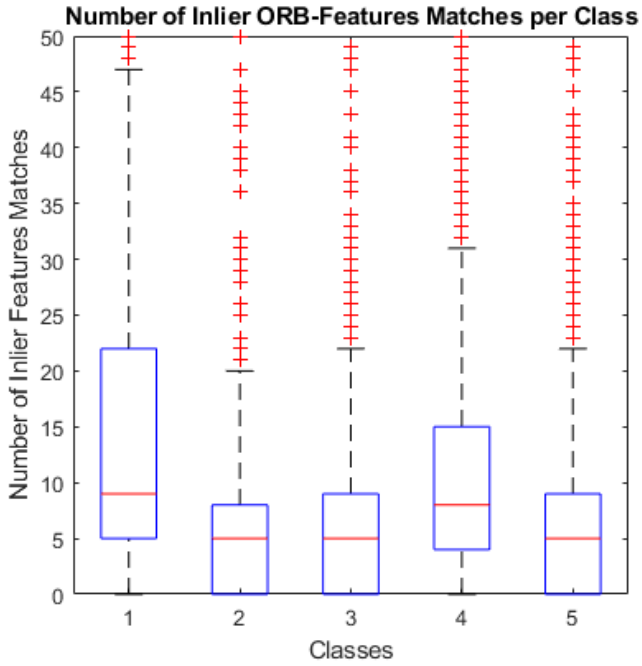


Figure 10. Boxplot of the number of inlier ORB-feature matches found for each class (red line is the median; bottom and top of blue boxes mark the 25th and 75 percentiles; + are outliers). Classes: 1. vessels, 2. vessels + bad view, 3. no vessels, 4. partly vessels and 5. partly vessels + bad view.

The Kruskal-Wallis H test showed that at least one of the classes resulted in a statistically significant different number of IFM compared to another class ($p < 0.001$). Multiple Mann-Whitney *U* tests showed a statistically significant difference when performing pairwise comparisons between all classes, except when comparing class 2 with 3, 2 with 5 and 3 with 5. An overview of the results can be found in Table V. The values correspond to the fraction of wins out of all pairwise comparisons. That is, *U*-value (number of wins) divided by the total number of comparisons (1016^2).

Table V. Results of multiple pairwise Mann-Whitney U tests with Asymp. Sig. (2-tailed). The values represent the fraction of wins out of all comparisons. Compared classes: 1. vessels, 2. vessels + bad view, 3. no vessels, 4. partly vessels and 5. partly vessels + bad view.

Class	2	3	4	5
1	0.29 ($p < 0.001$)	0.32 ($p < 0.001$)	0.44 ($p < 0.001$)	0.33 ($p < 0.001$)
2	-	0.48 (***p* = 0.090**)	0.33 ($p < 0.001$)	0.49 (***p* = 0.266**)
3	-	-	0.36 ($p < 0.001$)	0.50 (***p* = 0.695**)
4	-	-	-	0.37 ($p < 0.001$)

4.3.2 Qualitative Analysis

Figure 11 shows four examples of frame pairs that resulted in IFM that were considered true positives or *reliable*, based on visual inspection. It was noticed that in all classes, *unreliable* IFM were found on floating fetal skin flakes, fetal body parts (Figure 12, a-b) or image artefacts along the edge of the FOV (Figure 12, c). Furthermore, the IFM were considered false positives or *unreliable* in case the lines that connect the feature matches did not appear parallel, indicating incorrect transformation matrix estimation and thus incorrect discrimination between inlier and outlier feature matches (Figure 12, d-h).

False positive or *unreliable* IFM were mostly seen in frames from class 3, followed by frames from class 2 and 5. Class 1 and 4 showed the highest number of true positive IFM. However, in class 4, the edges of the structure that partly blocked the sight of the vessels led to some false positive IFM as well (Figure 12, a).

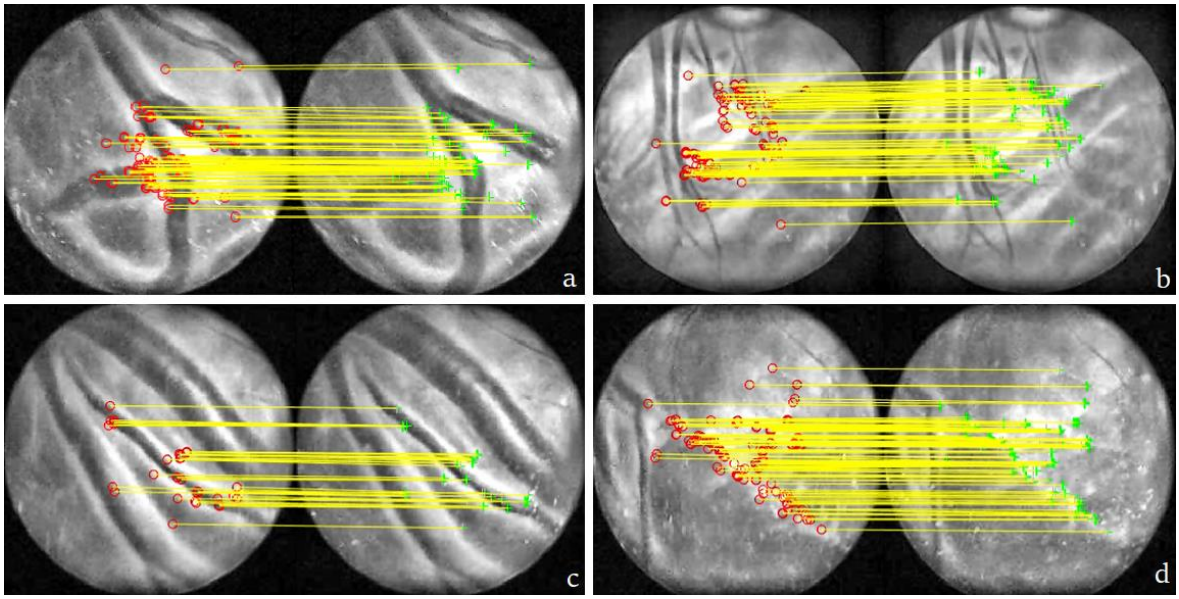


Figure 11. Examples of inlier feature matches (IFM) that are considered true positives based on visual inspection.

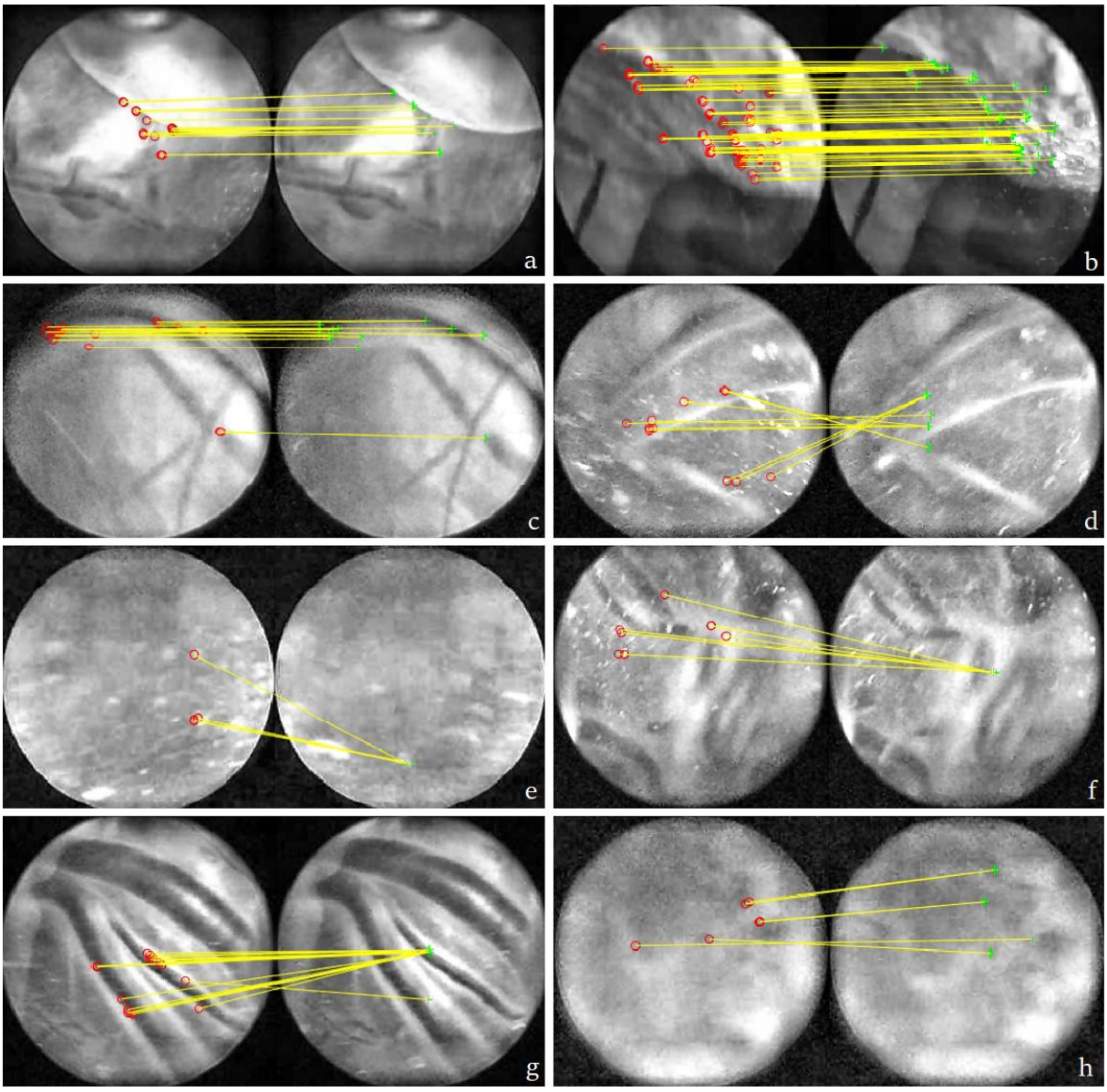


Figure 12. Examples of inlier feature matches (IFM) that are considered false positives based on visual inspection.

4.4 Discussion

When comparing the number of IFM found in the frames from the five different classes, most IFM were found in the first class (*vessels*), followed by the fourth class (*partly vessels*). This is in accordance with our hypothesis. The qualitative analysis showed that class 3 (*no vessels*) resulted in the most false positive IFM, followed by class 2 (*vessels + bad view*) and 5 (*partly vessels + bad view*). Moreover, the qualitative analysis showed that the edges of the occluding structures in class 4 and 5 led to false positives as well.

For class 2, 3 and 5, the median was five IFM, which is marginal considering the fact that at least four IFM are needed to calculate a geometric transformation matrix, in case of projective

transformations. However, it is important to mention that the number of IFM is affected by the chosen input parameters of MSAC, which were not optimized in this study.

Comparing our findings to literature is limited since, to the best of our knowledge, no similar study is published. Nonetheless, multiple studies that focused on feature-based image stitching of fetoscopic video sequences affirm the challenges related to *in-vivo* data, including both the overall low number and quality of features as well as the disruptiveness of occlusions, caused by fetal body parts and the turbidity of the amniotic fluid.[25,27,32,35,41]

4.4.1 Study Limitations & Future Recommendations

The input parameters that affect the geometric transformation matrix estimation and MSAC were chosen based on trial and error and no further tests were performed for parameter optimization. Since these parameters affect the number of IFM, parameter optimization is recommended in future development of a feature-based image stitching algorithm for *in-vivo* fetoscopic video sequences. Fortunately, our focus was mainly on the differences in the number of IFM found between the classes, rather than the absolute number of IFM. Therefore, our results are still assessable for evaluating the effect of the frame content on the relative number of IFM.

Although the qualitative analysis provided additional information on the true and false positives, no test was performed to explore the true and false negatives. In order to get optimal insights in the effect of the frame content on the number and quality of the IFM, additional investigation into the IFM classified as outliers is recommended.

It is important to be aware of the subjective nature of data labeling. The process of data labeling is sensitive to both inter and intra-observer variability. The latter can be reduced by repeatedly checking the labeled dataset until satisfied with all labeled frames. This was not done because of the limited amount of time available for this chapter and the time-consuming nature of this task.

Furthermore, one of the drawbacks of labeling video content is that there are always frames that display the transition between two classes. For example, when a video shows clearly visible vessels, followed by poorly visible vessels, classification of the so-called *transition frames* can be challenging. The observer performing the data labeling has to decide where to draw the line, and has to be consistent for comparable cases. When using this labeled dataset for training an image classification DL network, we suggest to experiment with excluding the transition frames from the training set and explore the network's output when confronted with these frames.

Based on our findings, it is recommended to actively exclude the frames from class 2, 3 and 5 when attempting feature-based image stitching. The frames from class 1 are deemed most suitable for image stitching. Whether the frames from class 4 are suitable for feature-based image stitching is inconclusive and has to be further explored. It is thus suggested to use either only the frames from class 1 or the frames from both class 1 and 4 for feature-based image stitching.

4.5 Conclusion

The aim of this chapter was to determine which *in-vivo* fetoscopic video frames are suitable for feature-based image stitching. This is done by evaluating the number and quality of the detected ORB-feature matches for different frame classes. Frames in which the vessels are visible without any occlusions result in the highest number of IFM and true positive IFM, followed by frames that show vessels that are partly occluded. Frames that show no vessels, fetal body parts or suffer from the turbidity of the amniotic fluid result in the lowest number of true positive IFM. It is recommended to actively exclude these frames for image-stitching purposes, since they are generally more sensitive to wrong geometric transformation estimation.

5. Automatic vessel identification using a deep learning approach

5.1 Introduction

This chapter focuses on training a binary classification network for vessel identification in *in-vivo* fetoscopic video frames. The goal of the DL network is to identify frames in which placental vessels are clearly visible, since these frames are considered useful for image stitching. Figure 13 provides a schematic overview of the purpose of this network, which will be referred to as *VesDet*.

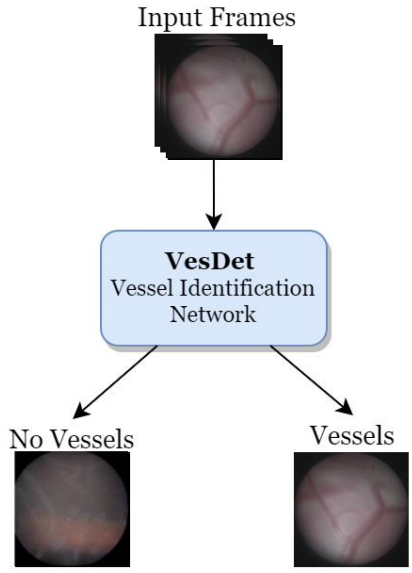


Figure 13. Illustration of the purpose of the VesDet network

Quite recently, two studies were published on deep learned classification of *in-vivo* fetoscopic video frames. First, Vasconcelos et al. compared multiple ResNet-based DL networks for ablation detection in fetoscopic video sequences to evaluate the surgical procedure timeline. They used a ResNet101 network with pre-trained weights, which was fine-tuned using their data. They showed that it is possible to incorporate existing network-architectures and pre-trained weights for image classification of fetoscopic images.[57]

Thereafter, Bano et al. trained a deep learning network to extend and improve the work done by Vasconcelos et al.[57], by adding additional fetoscopic events. They trained a spatio-temporal network to identify the following events: *Clear View, Occlusion, Tool* and *Vessel Ablation*. They combined a VGG-16 based network with pre-trained weights with a long short-term memory recurrent neural network (LSTM-RNN).[38]

The findings from both Vasconcelos et al. and Bano et al. suggest that existing network architectures, initialized with pre-trained weights (based on ImageNet), can be used for image classification of fetoscopic video sequences.

In this chapter, a deep learning network will be trained for binary image classification of *in-vivo* fetoscopic video frames. Based on both the findings from the above-mentioned studies and the results from the annual ImageNet visual recognition challenge (image classification, 2015)[58], the VGG-16 network architectures seems most suitable for image classification of fetoscopic video frames. Moreover, a transfer learning approach using pre-trained weights will be used since this is associated with early convergence and requiring less training data.[50,51] To our knowledge, we are the first to train a DL network for binary classification of vascular structures in fetoscopic video sequences.

5.2 Method

The frames of ten *in-vivo* fetoscopic videos (Chapter 3, Table I) were used, resulting in a total of 62,422 RGB with a resolution of 256 x 256 x 3 pixels. The frames were labeled according to the findings of Chapter 4: frames in which the vessels are clearly visible and frames in which vessels are 'partly' visible were assigned to the *Vessels* class. These frames are considered useful for image-stitching. The remaining frames were assigned to the *No Vessels* class. Table VI provides an overview of the two classes used in this chapter, including example frames. The number of frames for each class, per video, is provided in Table VII.

Table VI. The two classes used for labeling the in-vivo fetoscopic video frames.

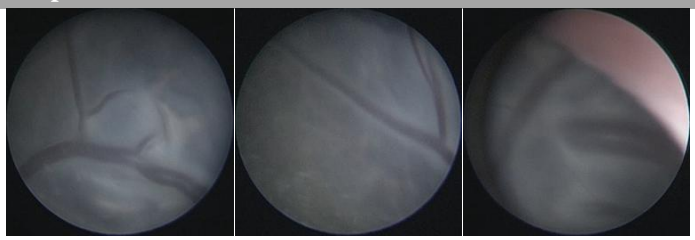
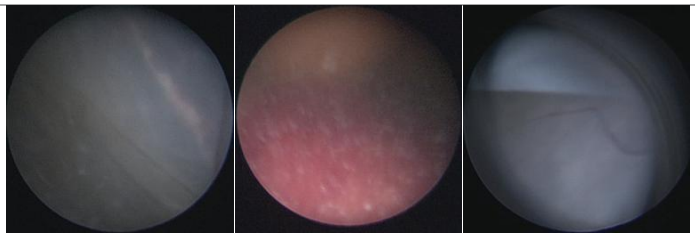
Class	Class description	Example frames*
Vessels	These frames are considered useful for image stitching. Vessels are clearly visible with acceptable image quality. The vessels might be partly blocked by occlusions (e.g. fetal structures, umbilical cord).	
No Vessels	These frames are considered *un*useful for image stitching. The frames either contain poorly visible vascular structures or no vessels at all.	

Table VII. Number of frames per class, per video.

Fetoscopic video #	*Vessels* class	*No Vessels* class	Total
1	1,388	2,123	3,511
2	771	1,224	1,995
3	442	1,496	1,938
4	117	658	775
5	7,291	5,387	12,678
6	1,899	3,409	5,308
7	18,817	5,049	23,866
8	158	1,488	1,646
9	1,104	4,227	5,331
10	1,961	3,413	5,374
Total	33,948	28,474	**62,422**

5.2.1 Network Architecture

A CNN with a VGG-16 architecture, as introduced by Simonyan and Zisserman[59], is used (Figure 14). To convert this multi-class classification network to a binary classification network, the output layer is converted to have a single output feature (node) with a sigmoid activation. Consequently, the output layer will generate a value between 0 and 1. The closer the output value to 0 or 1, the 'more certain' the network is that the input frame should be classified as *No Vessels* or *Vessels*, respectively.

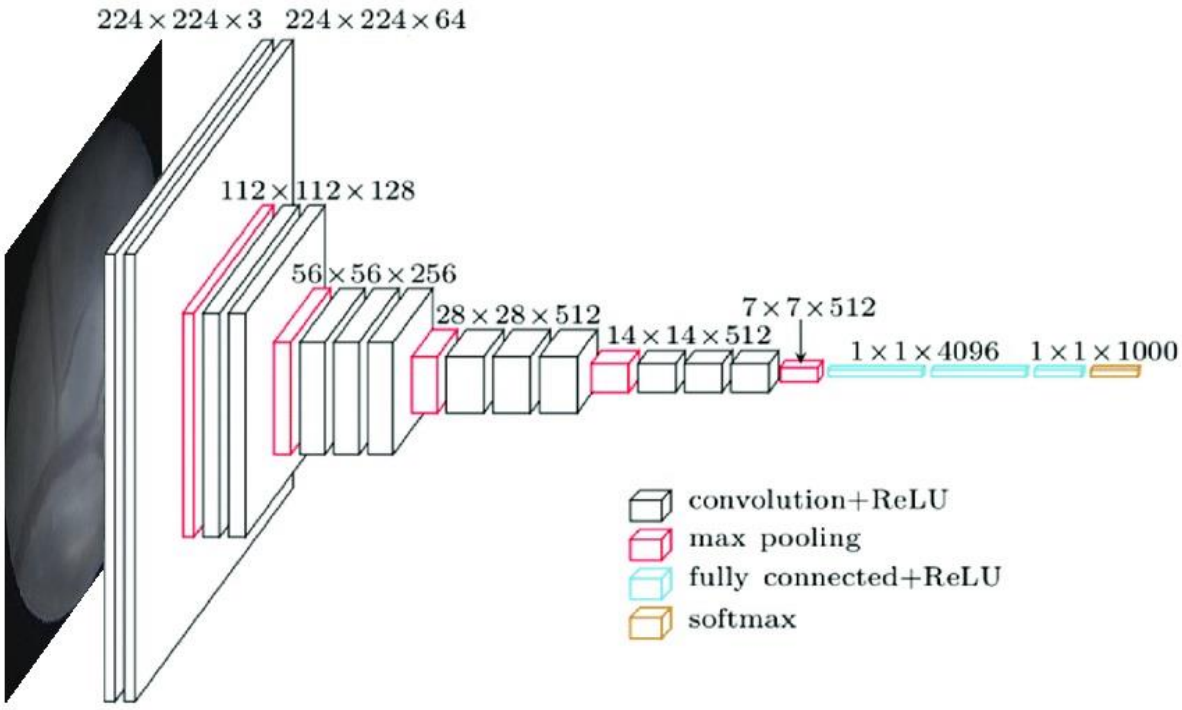


Figure 14. The architecture of a VGG-16 convolutional neural network (taken from Nash et al.[60] and modified).

5.2.2 Training

The weights of a pre-trained VGG-16 network (pre-trained on ImageNet) were imported from the TorchVision library (PyTorch framework)[a]. Data transformations were applied to the labeled frames, including a center crop to 224 x 224 pixels, conversion from RGB image to Tensor and image normalization (with mean = [0.485, 0.456, 0.406] and std = [0.229, 0.224, 0.225]).

The frames from video six were kept separate to serve as the test set. This is done to test whether the network is able to correctly classify the frames from an unseen FLOVA procedure or patient. The frames from the remaining nine videos were used for training. These frames were randomly split into a training and validation set, with an 80:20 split respectively.

The network was trained using a batch size of 6, an initial learning rate of 10^{-3} with a factor 0.1 drop every seven epochs. The network was trained for 15 epochs with early stopping. The network is trained using Google Colab's GPU.

5.2.3 Testing

The performance of the trained network was tested on one unseen video. A threshold value of 0.5 was applied to the network's predictions to classify the frames as *Vessels* (prediction > 0.5) or *No Vessels* (prediction < 0.5). The results were collected in a confusion matrix, which was used to calculate performance metrics. Moreover, the performance of the classification network was visualized by calculating and plotting the true and false positive rates for various threshold values, also known as the Receiver Operating Characteristics (ROC) curve.

Additionally, a random set of input frames and corresponding predictions were plotted and visually inspected to evaluate the network's performance. In case of a wrong classification, the prediction value is assessed to find out how 'sure' the network was of its prediction. Ideally, wrong classifications have values surrounding the threshold value of 0.5. Moreover, wrongly classified frames were plotted and visually inspected to see what frame content leads to wrong predictions and what the prediction value was.

For additional qualitative analysis of the performance of VesDet, saliency maps were generated to determine what spatial region of the frame is most relevant to the network when determining its predictions. This is done using guided backpropagation[61], using the FlashTorch[b] visualization toolkit. Lastly, the prediction speed of the network was measured.

[a] https://pytorch.org/vision/0.8/models.html
[b] https://github.com/MisaOgura/flashtorch

5.3 Results

Due to overfitting after seven epochs, the network of the seventh epoch is considered the best network. For this network, the training and validation accuracy were 0.955 and 0.951 respectively.

Table VIII shows a confusion matrix of the performance VesDet on the unseen video. Here, a threshold of 0.5 is applied to determine the predictions. This results in a sensitivity, specificity and accuracy of 0.958, 0.811 and 0.863 respectively. Figure 15 shows the ROC curve of the same results for various thresholds between 0 and 1. The area under the ROC curve (AUC) is 0.95. The saliency map of a frame from the video six is shown in Figure 16. Lastly, Figure 17 and Figure 18 show saliency maps and predictions for random frames from the *Vessels* and *No Vessels* classes, respectively.

The measured average prediction speed of the network was 0.0014 seconds per frame, which corresponds with a prediction rate of about 714 fps.

Table VIII. Confusion matrix of the predicted classes, when tested on an unseen video

		Predicted Class		
		Vessels	*No Vessels*	Total
Actual Class	*Vessels*	1,820	79	1,899
	No Vessels	646	2,763	3,409
	Total	2,466	2,842	5,308

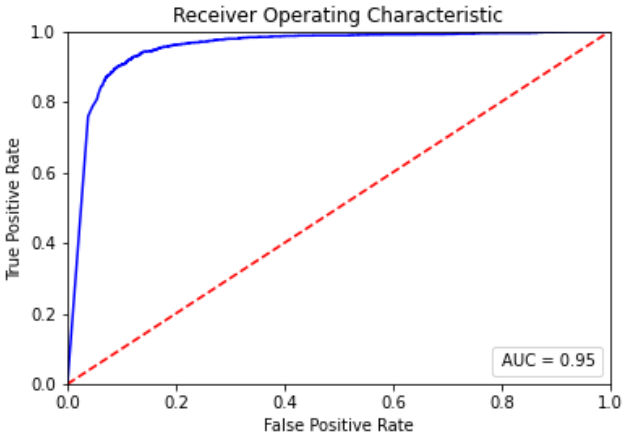


Figure 15. Receiver Operating Characteristics curve of the performance of VesDet, when tested on an unseen video.

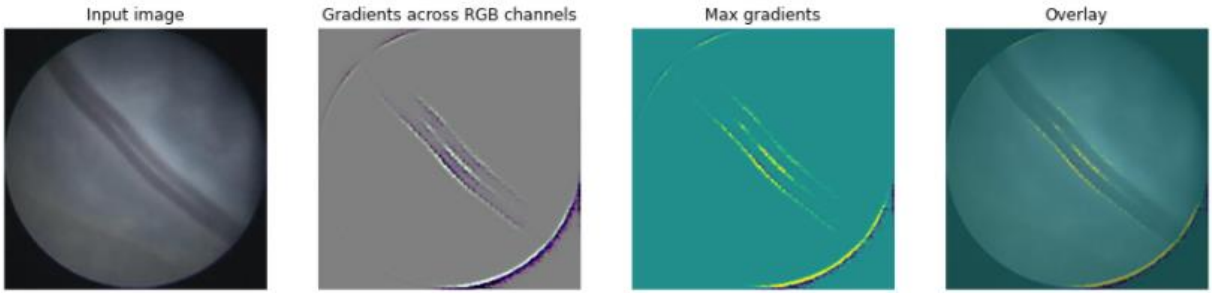


Figure 16. VesDet saliency map on an unseen input frame (video 6) from the 'Vessels' class.

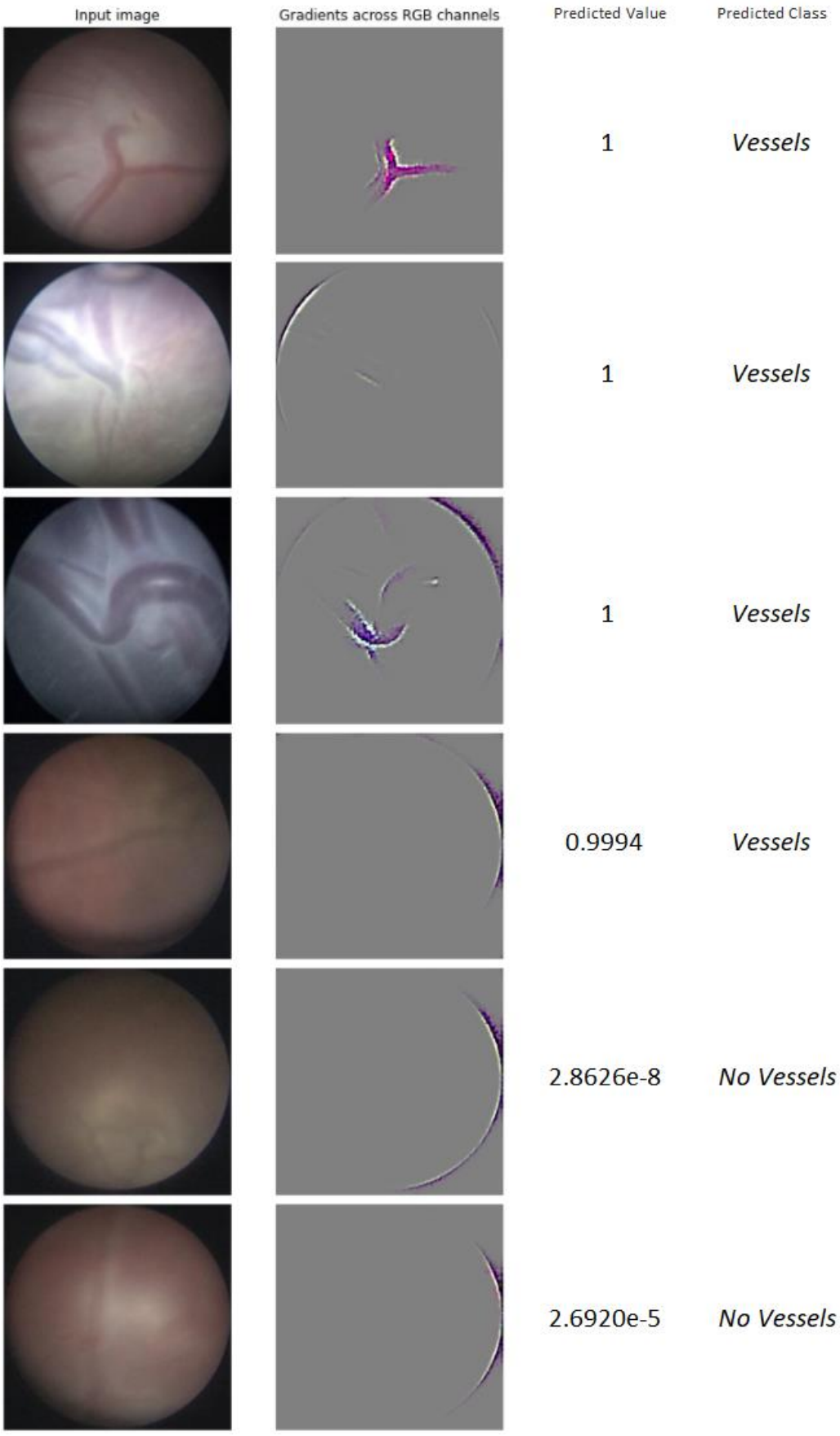


Figure 17. Predictions of VesDet on frames from the 'Vessels' class. A threshold of 0.5 is used to determine the predicted class.

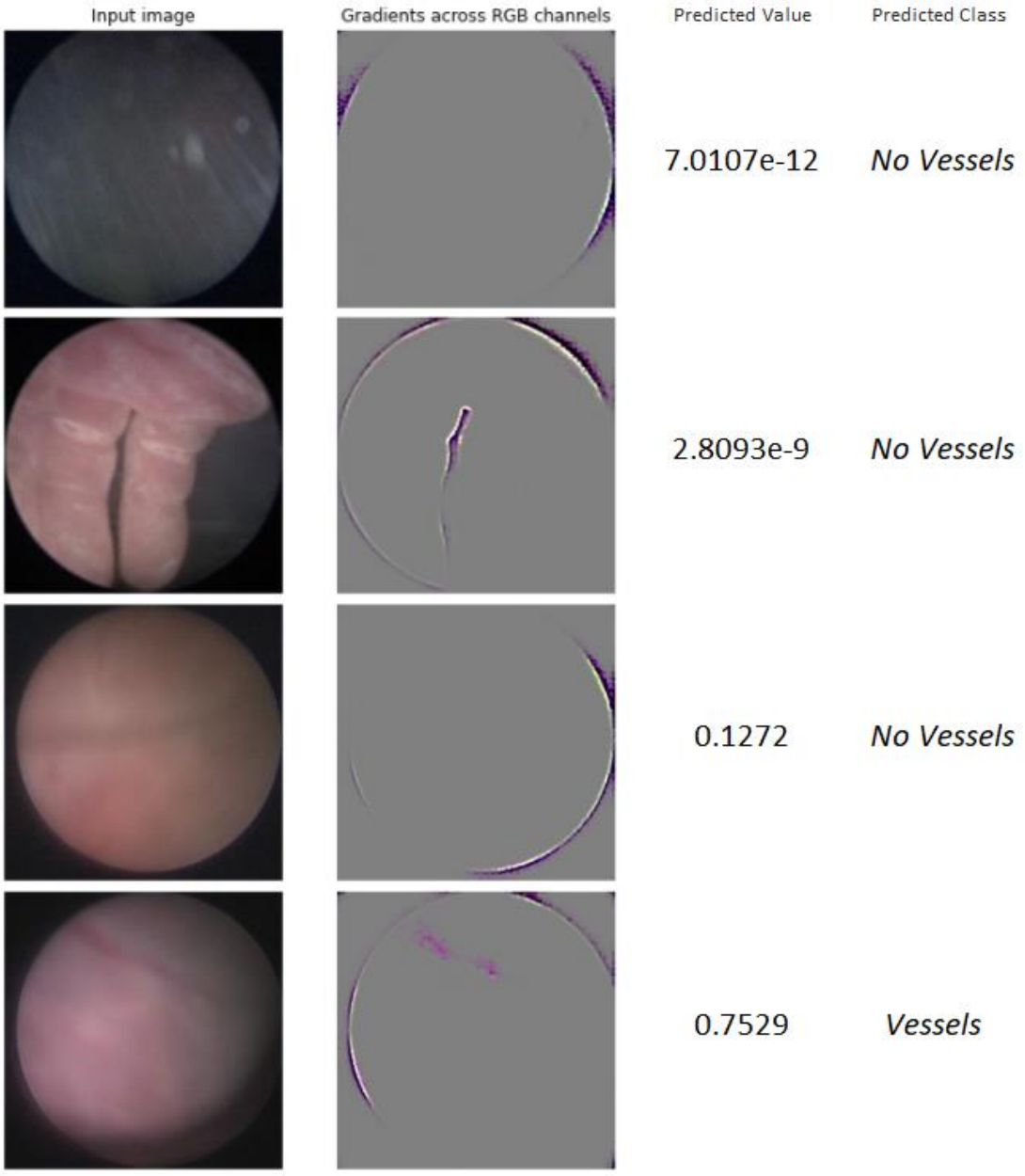


Figure 18. Predictions of VesDet on frames from the 'No Vessels' class. A threshold of 0.5 is used to determine the predicted class.

5.4 Discussion

The results show that our DL network was able to detect frames with clearly visible vessels in an unseen fetoscopic video. The network shows promising results for future applications with a ROC-AUC of about 0.95. The predictions were calculated at a high framerate. Moreover, qualitative analysis support the idea that the network is capable of recognizing vascular structures. This is based on the highlighted areas in the saliency maps at the positions of vascular structures. Therefore, the network is considered suitable further development of an image stitching algorithm for FOV expansion during fetoscopic surgery.

The difference in accuracy (0.86) and AUC (0.95) can be explained by the fact that the accuracy is based on a single threshold value rather than a range of threshold values. The high AUC value indicates that wrongly predicted classes have prediction values surrounding the threshold value, meaning that the network was 'not so sure' about its prediction. Furthermore, the AUC value is considered more relevant for clinical applications since a fixed optimal threshold value does not exist for each setting. In other words, the appropriate threshold value is expected to vary between patients and technical settings, such as lighting conditions.

The measured average prediction rate 714 fps suggest that this network is suitable for real-time implementations. However, it is important to mention that the calculations were performed using Google Colab's GPU. Google Colab offers free usage of multiple different GPUs. Which GPUs are available vary over time and users are not allowed to select the GPU they want to use. Therefore, we do not know what GPU was precisely used to perform our calculations. Additionally, the usage of Google Colab's GPU is limited to a certain amount of time. It is therefore not a sustainable solution.

Some of the saliency maps show unexpected or surprising results. One surprising finding is that some correctly classified frames have 'empty' saliency maps. One possible explanation could be that the saliency maps only show the maximal values. Moreover, the fact that the borders of the FOV are highlighted is an unexpected result, since the borders are present in all frames from both classes and should therefore not play an important role in the network's decision-making.

Comparison of our results to findings in literature is limited, since we were the first to train a binary classification network for vessel identification in fetoscopic video sequences. Nevertheless, one study focused on multi-class classification of fetoscopic frames: Bano et al. trained a classification network for identifying multiple events, including *Clear View*, which is thought to be similar to our *Vessels* class. The frames from the *Clear View* class were detected with an F1-score and AUC of 0.85 and 0.91, respectively.[38] This is quite similar to our findings (F1 score of 0.83; AUC of 0,95). However, comparing the results of one class from a multi-class classification network with our binary classification network is unfounded.

5.4.1 Study Limitations & Future Recommendations

In this chapter, the labeled frames from Chapter 4 were directly used. The results from Chapter 4 showed that the frames with vessels (partly blocked by occlusions) result in more inlier ORB-feature matches (IFM). However, it is not demonstrated that these frames make feature-based image

stitching feasible yet. Since the ultimate goal of the network is to identify frames that are useful for image stitching, it would be appropriate to label the frames based on tests involving image stitching.

Another limitation is the fact that the data was unbalanced, meaning that there was no equal distribution between the number of frames in each fetoscopic video and in the two classes. This was not corrected for by undersampling to keep the dataset as large as possible. An alternative would be to remove frames until a balance is reached, followed by data augmentation, such as mirroring, to enlarge the dataset.

The performance of the network is tested on one unseen video. A more robust way would be to train and test the network using a *k*-fold cross-validation method, in which *k* is the number of fetoscopic videos and each fetoscopic video acts as the test set for *k* folds. The frames from the remaining *k*-1 videos will be subdivided in a training and validation set and used for training. Once trained, the network is evaluated and this whole process is repeated *k* times. This approach is not executed because of the limited time available in this study. However, this is highly recommended for further development of the network.

For future research and development of the binary classification network, it is recommended to include temporal information. In other words, prediction values from previous frames should be taken into account in the prediction of the current frame. This is thought to be relevant since the type of frame content is commonly consistent for a certain number of consecutive frames. Another recommendation for further development is to compare different network architectures, including ResNet, GoogLeNet and Inception-v3/v4.

5.5 Conclusion

The goal of this chapter was to train a binary classification network for vessel identification in *in-vivo* fetoscopic video frames. These frames can be used to create an overview of the placental vasculature by performing image stitching. Our network was successful in classifying frames of an unseen FLOVA procedure video (AUC = 0.95), supporting its potential for application in future clinical settings. Further research is could focus on optimization of the network's performance, including experiments with different network architectures and the inclusion of temporal information.

6. Automatic vessel segmentation using U-Net

6.1 Introduction

This chapter focuses on the training of an automatic segmentation network. The resulting vessel segmentations or can be used for either 1) selective regional image enhancement for improved feature detection or 2) for intensity-based image stitching, respectively. Figure 19 provides a schematic overview of the purpose of this network, which will be referred to as *VesSeg*.

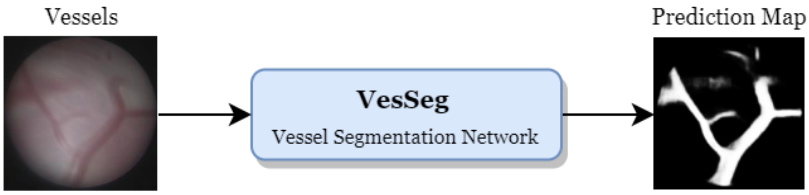


Figure 19. Illustration of the purpose of the VesSeg network

Sadda et al. reported the first attempt in deep-learned placental vessel segmentation of *in-vivo* fetoscopic videos using a fully-connected neural network (FCNN).[19] Casella et al. proposed a FCNN network for automatic inter-fetus membrane segmentation to help the surgeon identify the membrane, which can serve as a reference for finding vascular anastomoses.[42] The proposed FCNN networks were variations of the U-Net architecture, which was introduced by Ronneberger, Fischer and Brox.[53]

Recently, Bano et al. proposed a network based on U-Net combined with a pre-trained ResNet101 backbone for placental vessel segmentation in fetoscopic videos. The prediction maps were used for intensity-based image stitching. After performing 6-fold cross-validation, a Dice score of 0.78 ± 0.13 was found. They manually annotated 483 *in-vivo* fetoscopic video frames with vascular structures and offered both the input frames and the ground truths (GTs) for public use.[62]

Since limited work is published on deep-learned segmentation of placental vessels, inspiration is taken from literature in the field of fundus imaging. Multiple articles have been published on automatic vessel segmentation of retinal vasculature.[63–67]

For our vessel segmentation DL network, a U-Net will be trained from scratch. The traditional U-Net architecture, as introduced by Ronneberger, Fischer and Brox, will be used since it is known for its robustness and requiring relatively little training data. [53] Moreover, U-Net is the most widely used encoder-decoder network for segmentation tasks in the field of medical imaging. [54]

6.2 Method

First, a summary of the methodology is provided: the dataset offered by Bano et al.[62] is used for training our first network. This network was then used to generate GTs for the *in-vivo* frames from the Radboudumc (referred to as *Rumc*). This was done by saving the outputs from the network and

fine-tuning them manually. Thereafter, the network was trained two more times, using the *Rumc* dataset and a combination of both datasets (*BanoRumc*).

6.2.1 Data

The total dataset consisted of 729 *in-vivo* fetoscopic video frames and GT segmentations, of which 483 from the published dataset from Bano et al.[62]. The remaining 246 frames came from our *in-vivo* fetoscopic videos (see Chapter 3, Table I). Only frames in which the vessels were clearly visible (see Chapter 4) were included. It is important to mention that the 246 frames were randomly selected to yield the most diverse dataset.

Before training the networks, the GTs from the *Bano* dataset were 'cleaned'. This means that the GTs were manually fine-tuned to yield even more precise GTs. This was done because 1) some very small or large vessels were not included in the GTs, 2) some vessels were 'drawn outside the FOV' and 3) the light reflections on the vessels were excluded in some frames. Lastly, the frames were binarized by thresholding because some GTs contained more than two pixel values, which is inconvenient for a binary segmentation task. Table IX show example frames of the *Bano* dataset, including the original GTs and the GTs after 'cleaning'. Table X show examples frames from the *Rumc* dataset.

Table IX. Example frames from the 'Bano' dataset. The cleaned GTs are the GTs used in this chapter.

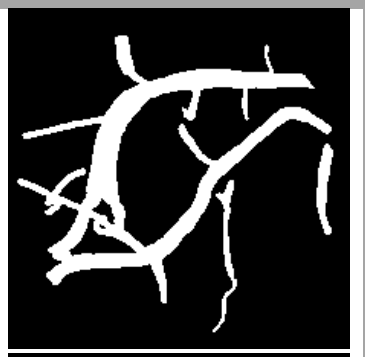
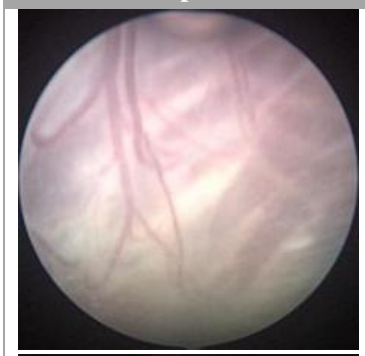
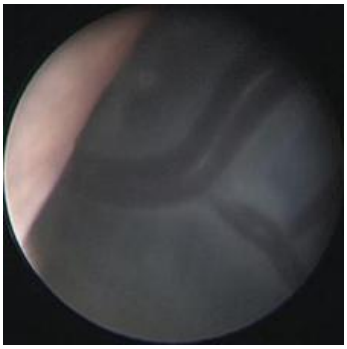
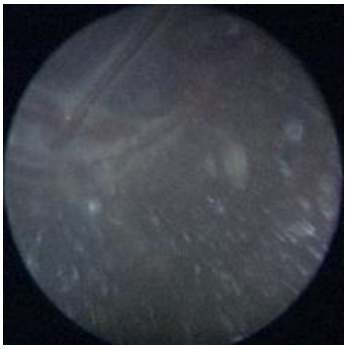
Input	Ground Truth (original)	Ground Truth (cleaned)
		
		

Table X. Example frames from the 'Rumc' dataset.

Input	Ground Truth
	
	
	

After splitting the test sets (~15%) from the datasets, data augmentation was performed: the frames were flipped horizontally, vertically and horizontally and vertically combined (Figure 20), which quadrupled our datasets. This is done because DL networks generalize better when trained using more data.[68] Table XI provides an overview of the number of frames used in this chapter.

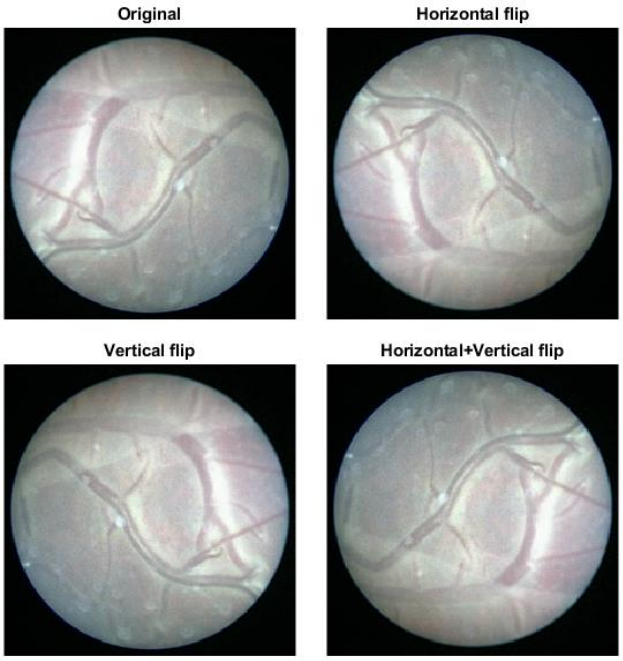


Figure 20. The data augmentation used in this chapter.

Table XI. Number of frames for each dataset used in this chapter, after splitting the total dataset in train and test sets.

Dataset	Train	Test	Total
Bano	1,640	288	1,928
Rumc	844	140	984
BanoRumc	2,484	428	2,912

6.2.2 Network Architecture & Training

The U-Net architecture (Figure 7) is used with input and output layers of size 224 x 224 x 3 and 224 x 224 x 1, respectively. The network was built in python using the Keras open-source library. The output layer has a sigmoid activation layer. Therefore, the network generates prediction values between 0 and 1 for each individual pixel. The higher the prediction value, the more 'certain' the network is that the pixel belongs to a vessel. All prediction values combined in one grid form a prediction map.

Before training the network, the frames were downsized to 224 x 224 pixels in height and width and split in a training and validation set, with an 80:20 split respectively. The frames were randomly split with controlled shuffling, for a reproducible output for each training and testing session. The network was trained three times, using the *Bano*, *Rumc* and *BanoRumc* datasets. The trained networks will be referred to as *Unet_Bano, Unet_Rumc* and *Unet_BanoRumc* respectively.

A batch size of 8 was used and the network was trained for 30 epochs, with early stopping. An initial learning rate of 10^{-3} was used, with a factor 0.1 drop after 3 epochs of no improvement in validation

loss, with a lower bound of 10^{-5}. The network was trained using Keras with TensorFlow as backend. The training was performed on a Lenovo ThinkPad with an Intel® Core™ i5-4200U processor.

6.2.3 Testing

The performance of the trained networks was tested using both the test sets that correspond to the training set and other test sets. For example, the network trained using the *Bano* training set was tested on the *Bano, Rumc* and *BanoRumc* test sets. This is done to test the performance of a trained network when applied to data from different FLOVA procedures (e.g. different scopes, patients, settings, etc.). It is important to mention that the test sets were not used for any of the training sessions.

The performance of the segmentation network was visualized by calculating the true and false positive rates for various threshold values applied to the prediction maps. The true and false positive rates were plotted in a ROC curve. Moreover the networks' performances were measured by calculating a Sørensen–Dice, also known as a Dice Score, DSC or F1 score. A Dice Score takes two times the area of overlap, divided by the total number of pixels in both images:

$$Dice\ Score = \frac{2|X \cap Y|}{|X| + |Y|} \quad (1)$$

The quality of the resulting prediction maps was assessed by visual inspection. This is done by plotting a randomly selected set of input frames with the corresponding prediction maps from the three networks. The prediction maps were judged based on false negatives, such as missing vessels, and false positives, such as flakes of the amniotic fluid. Moreover, the prediction speed of our network was measured. Lastly, the performance of our best network (highest Dice Score) was qualitatively compared to the networks proposed by Sadda et al.[19] and Bano et al.[62].

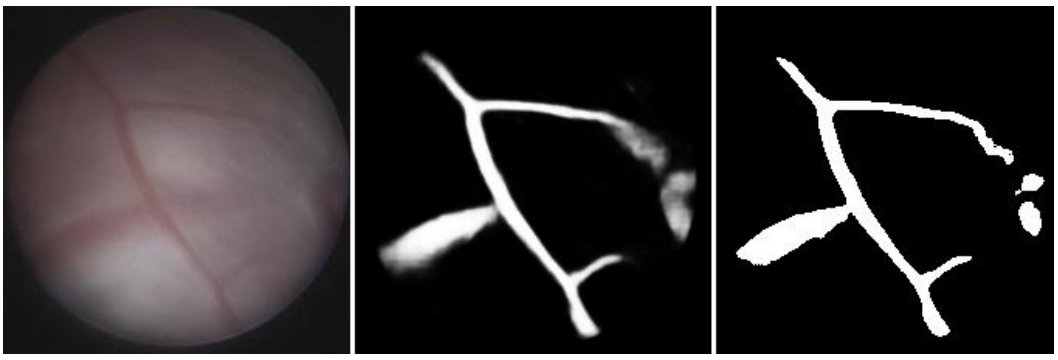


Figure 21. Example frame with its corresponding prediction map (middle) and segmentation map (right). A threshold value of 0.5 was used to generate the segmentation map.

6.3 Results

The loss, accuracy and Dice Score of the trained networks, tested on all three test sets, can be found in Table XII. The *Unet_BanoRumc* has the best Dice Score for all three test sets. The values marked in blue refer to cases in which the data used for training and testing originate from the same dataset. For these tests, the area under the ROC curve was 0.98, 0.97 and 0.98 for *Unet_Bano, Unet_Rumc* and *Unet_BanoRumc*, respectively. Figure 22 shows the ROC curve of the *Unet_BanoRumc* network when tested on the *BanoRumc* test set. The measured average prediction speed of the network was 0.14 seconds per frame, which corresponds with a prediction rate of about 7 fps.

Figure 23 and Figure 24 provide some example frames for qualitative comparisons. Figure 23 focuses on comparisons between our own networks which were trained using different datasets. Figure 24 provides insights in the performance of our best network (*Unet_BanoRumc*) compared to networks proposed by colleagues.

Table XII. The performance of the trained U-Nets on multiple test sets. The rows marked in blue are the cases where the training and test set correspond.

Network	Test set	Loss	Accuracy	Dice Score (± std)
Unet_Bano	Bano	0.09	0.97	**0.79 (± 0.12)**
	Rumc	0.20	0.94	0.53 (± 0.36)
	BanoRumc	0.13	0.96	0.70 (± 0.26)
Unet_Rumc	Bano	0.17	0.95	0.60 (± 0.23)
	Rumc	0.12	0.96	**0.66 (± 0.28)**
	BanoRumc	0.15	0.95	0.62 (± 0.25)
Unet_BanoRumc	Bano	0.09	0.97	**0.80 (± 0.13)**
VesSeg	Rumc	0.10	0.97	0.72 (± 0.26)
	BanoRumc	0.09	0.97	0.77 (± 0.18)

Figure 22. Receiver Operating Characteristics (ROC) curve of the performance of the 'BanoRumc' network, when tested on the 'BanoRumc' test set.

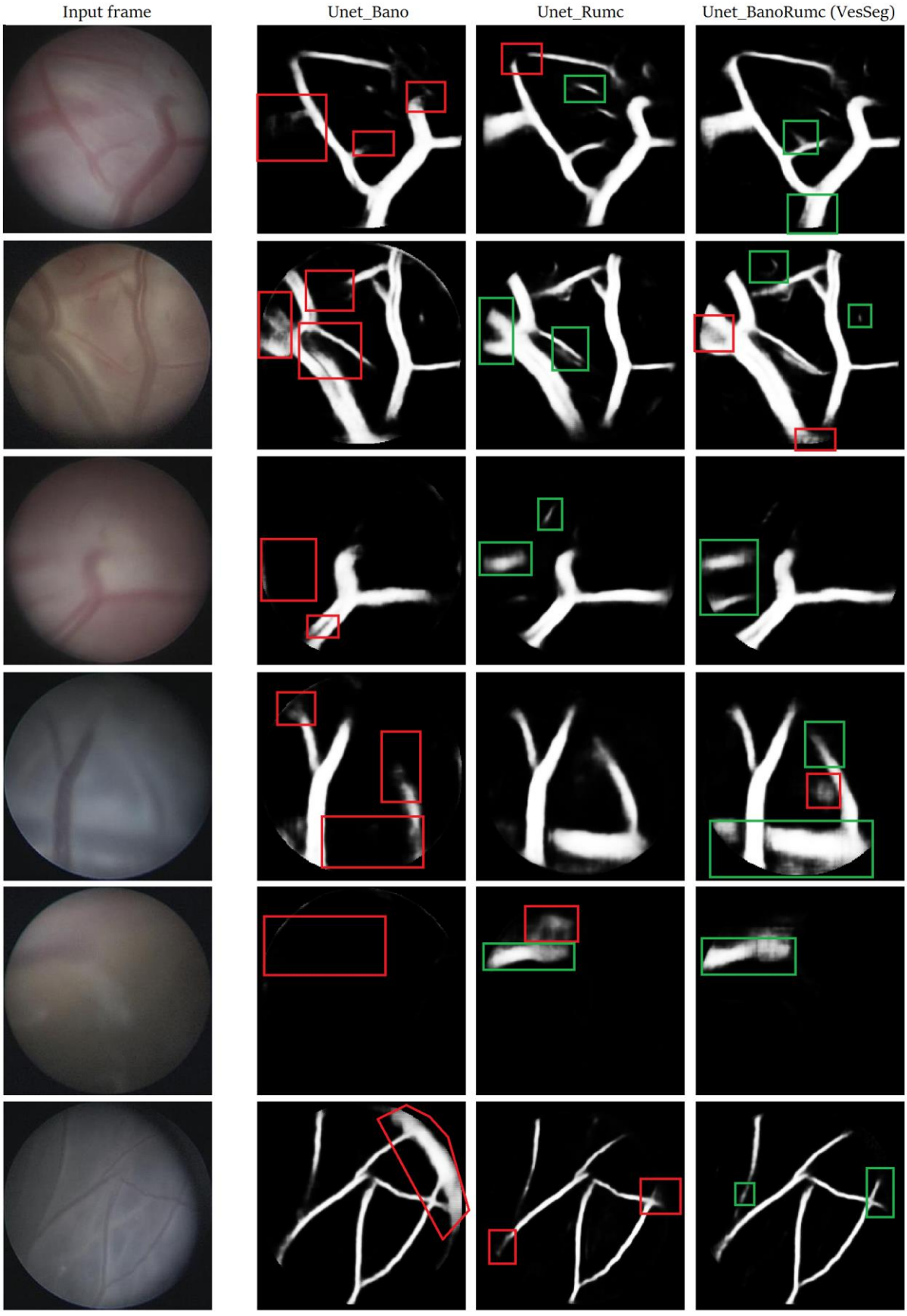


Figure 23. Prediction maps generated by our networks, which have been trained using different datasets ('Bano'; 'Rumc'). The red boxes mark false positives or negatives; the green boxes mark true positives or negatives.

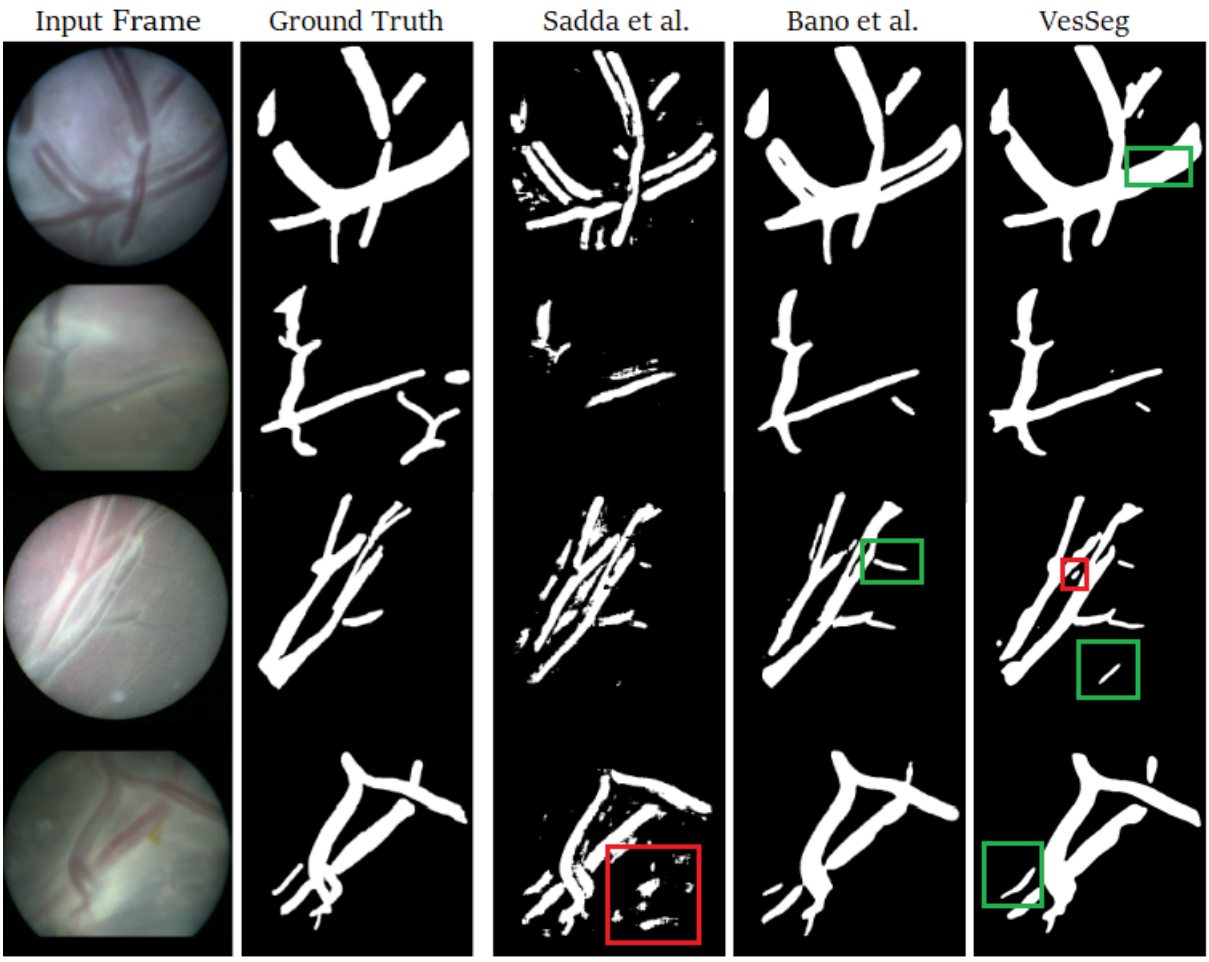


Figure 24. Qualitative comparison of our best network (Unet_BanoRumc, VesSeg) with the DL networks proposed by Sadda et al.[19] and Bano et al.[62]. The four input frames originate from the 'Bano' dataset. The green and red boxes mark true and false positives respectively.

6.4 Discussion

Our results show that our U-Net, which was trained from scratch, was able to generate segmentations of vascular structures in *in-vivo* fetoscopic video sequences, with Dice Scores up to 0.80 and an area under the ROC curve of 0.98. The U-Net trained using the largest dataset (*Unet_BanoRumc*) outperformed the other networks.

The qualitative analysis showed that the *Unet_BanoRumc* network performed the most accurate segmentations of vascular structures. Based on visual inspection, the *Unet_BanoRumc* was most sensitive to both smaller and larger vessels. The *Unet_Bano* was most susceptible to the light reflections on the vessels and led to the most false negatives.

The *Unet_BanoRumc* network outperformed the network proposed by Sadda et al.[19] This was seen both quantitatively (they reported a Dice Score of 0.55 (± 0.22) and qualitatively. Moreover, our *Unet_Bano* and *Unet_BanoRumc* networks slightly outperformed the network trained by Bano et al.: they reported an overall Dice Score of 0.78 (± 0.13)[62], compared to 0.79 (± 0.12) and 0.80 (± 0.13)

for our networks. However, these differences are marginal. The qualitative comparisons also showed a similar performance of our network compared to that of Bano et al. However, our network seemed to perform better at detecting relatively small vessels compared to their network. This can be explained by the fact that extra attention was payed to precise annotation of the GTs, which included more inclusions of small vessels.

Besides the fact that the *Unet_BanoRumc* network (slightly) outperformed the other networks based on the quantitative and qualitative analysis, additional motivation for using our network for clinical applications can be found in the fact that the network is trained using the most diverse dataset. The dataset is most diverse in terms of number of patients, vessel size (very small and large vessels), image quality, used fetoscopes and frame content. The latter refers to the frames in which the vessels are partly blocked by structures such as fetal parts or the umbilical cord.

Lastly, the measured average prediction rate of about 7 fps is considered sufficient for real-time image stitching purposes. This is because not all frames are necessary for image stitching due to the relatively large overlapping areas in the consecutive frames.[62] Furthermore, the predictions were calculated using an Intel Core i5 processor. Using a more advanced processing unit is thought to improve the prediction rates.

6.4.1 Study Limitations & Future Recommendations

It is important to be aware of the subjective nature of data labeling, making it sensitive to both inter- and intra-observer variability. In our study, the data was labeled once, by one person. To improve the reproducibility, it is recommended to label the data multiple times, by multiple individuals.

The data from multiple fetoscopic videos were merged and shuffled before splitting the test sets from the datasets. As a result, some frames in the test set might be similar in appearance to frames in the training set. In future clinical applications, the network will be confronted with new clinical data. Therefore, it is recommended to re-train the network while leaving one fetoscopic video out and use this video as the test set, which is known as the hold-out method. Another, more robust evaluation method is the *k*-fold cross-validation (see section 5.4.1).

In our study, the traditional U-Net architecture[53] was used. For further improvement of the network's performance, it is recommended to experiment with different network architectures. Another recommendation would be to include temporal information, since consecutive frames show largely similar scenes. To do so, the prediction map from the previous frame is used as additional input for the DL network.

The motivation for training a vessel segmentation network (VesSeg) was to use the resulting prediction or segmentation maps for image stitching by either 1) using the segmentation maps for selective regional image enhancement of the frames for improved feature detection or 2) using the prediction maps from VesSeg for intensity-based image-stitching. Besides this, the segmentation maps might be of additional use during FLOVA procedures. For example, after preprocessing the

frames using selective regional image enhancement, these frames can be presented to the surgeon for improved image quality and vessel visibility.

Another potential use for VesSeg is presenting the prediction or segmentation maps directly to the surgeon. While discussing the results of the network with a gynecologist from the Radboudumc, he mentioned that the network was able to detect some very small vessels, which were not noticed by the gynecologist at fist. In other words, the prediction maps could potentially supply the surgeon with additional information. Nonetheless, it should be pointed out that the network might fail to detect some (small) vessels. Therefore, the original *in-vivo* fetoscopic frames should remain the main source of information. Another potential application could be to superimpose the segmentation or prediction maps onto the original frames during FLOVA and allow for manual adjustments of the transparency of the masks if desired.

Finally, the segmentation maps could be used as a region of interest (ROI) during or after feature detection. When using a segmentation mask *after* feature detection, the features located outside the mask are discarded. However, using the segmentation mask as a ROI *during* feature detection is thought to be more effective because of how the feature detection function operates: feature detection continues until a certain number of features are found. Using the segmentation mask while detecting features, rather than after, is thought to lead to more features within the ROI. This has not been tested in our project since the feature detection function used only allows for rectangular ROIs.

6.5 Conclusion

The aim of this chapter was to train a deep learning network for automatic vessel segmentation in *in-vivo* fetoscopic video sequences. Our trained network showed promising results, including Dice Scores up to 0.80 and a ROC-AUC of 0.98. Moreover, when visually comparing the input frames with the resulting prediction maps, the results appear convincing. Future research is required to determine for what potential image stitching approach the prediction maps are most useful.

7. Technical note: vessel segmentation maps for image stitching

7.1 Introduction

This chapter discusses the use of the vessel segmentation maps for both feature-based and intensity-based image stitching of *in-vivo* fetoscopic video sequences. The previously trained deep learning networks will be incorporated: VesDet will be used to select frames with vessels and VesSeg will be used to create vessel prediction and segmentation maps. The potential uses of both the prediction and segmentation maps for image stitching will be explained and briefly demonstrated. Recommendations for future development of an image stitching algorithm will be given based on our observations gained through experimentation. The main goal of this chapter is to explore the possibilities for image stitching using VesSeg and provide recommendations for future development of the image stitching algorithm.

In the field of fetoscopic surgery, most research focusses on feature-based image stitching.[21–24,34–36] Although feature-based approaches are generally associated with a high robustness and computational efficiency[69,70], the visual properties of the *in-vivo* fetoscopic video sequences can lead to an inadequate number of reliable features. This is thought to be the result of a series of factors, including the poor lighting conditions, the lack of texture of the placental surface and the presence of occluding and dynamic structures in the intrauterine environment.[26,38] The latter can lead to the detection of dynamic features, which is problematic since feature-based stitching relies on static or stable features.

An alternative to feature-based image stitching is intensity-based image stitching, also known as direct image registration or alignment. Quite recently, Intensity-based image stitching has gained attention in the field of fetoscopic surgery. In 2018, Peter et al.[26] proposed the first intensity-based approach for image stitching of *in-vivo* fetoscopic video sequences. They performed dense pixelwise alignment of the image gradient orientations of *in-vivo* fetoscopic video frames.[26] Thereafter, in September 2020, Bano et al. published an article that proposes a direct registration method using the probability (prediction) maps from a vessel segmentation network.[62] Their results were very promising and support our hypothesis, being that vessel segmentations can be used for intensity-based image stitching of *in-vivo* fetoscopic video sequences. Therefore, one of our goals is to see whether we can develop a similar algorithm for the department of Department of Obstetrics & Gynaecology at the Radboudumc.

7.1.2 Image Stitching Algorithm

In image stitching, two or more frames with overlapping content are combined together to create one larger image or overview. This process involves image registration, warping and blending.[71] In the image registration step, the geometric transformation between two images is estimated. When using data from a fetoscope, which is a monocular system, transformation matrices are estimated based solely on the information from the individual video frames. Feature-based and intensity-based image stitching approaches differ in the way they determine the geometric transformation between two consecutive frames.

Feature-based image registration

When using a feature-based approach, the transformation matrix is estimated based on the matching feature points. As demonstrated in an earlier project (*M2-3*), the number of inlier feature matches (IFM) can be increased using vessel segmentation maps combined with a so-called selective regional image enhancement (SRIE) step. In this chapter, we will experiment with both the prediction and segmentation maps from VesSeg for improved feature-based image stitching of *in-vivo* fetoscopic video sequences.

Intensity-based image registration

Intensity-based image stitching approaches use direct pixel-to-pixel comparisons with gradient descent or another optimization technique in order to estimate the geometric transformation between two frames. This is done by applying numerous transformations to the so-called 'moving' image in order to align it with the 'fixed' image. The smaller the sum of the pixelwise intensity value comparisons, the better the alignment or registration of the two images. We hypothesize that the prediction maps from VesSeg are suitable for intensity-based image stitching.

7.2 Experimentations & Observations

Multiple experimentations with feature-based and intensity-based image stitching were performed to get insights in the potential and usage of the prediction and segmentation maps generated by VesSeg. All experimentations were performed using MATLAB R2020b.

7.2.1 Data Acquisition

The frames from two *in-vivo* fetoscopic videos (Chapter 3, Table I) were used: video 1 and 6. These videos were selected because of their relatively good image quality and the fact that the placentas were located posteriorly, which is thought to make image stitching more feasible compared to frames with anterior placentas.

VesDet was used to select frames with clearly visible vascular structures. Thereafter, four sets of 50 consecutive frames were manually selected and extracted (Figure 25). VesSeg was then used to generate the prediction maps (Figure 26), which were later converted to segmentation maps by image binarization.

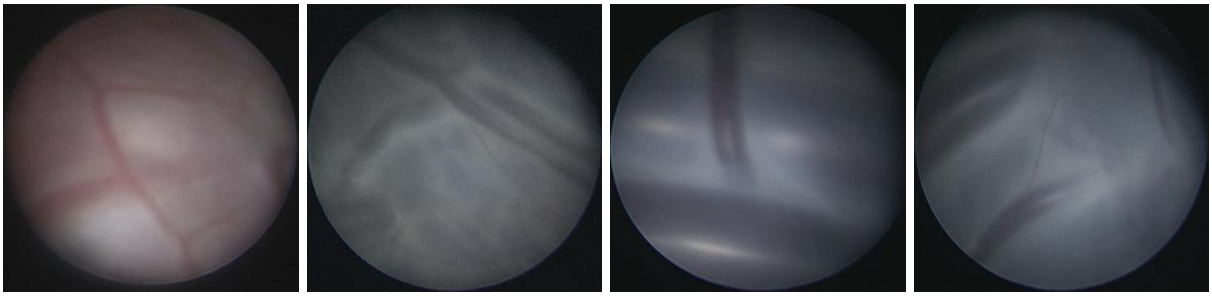


Figure 25. Example frames from the four sets of in-vivo fetoscopic video frames.

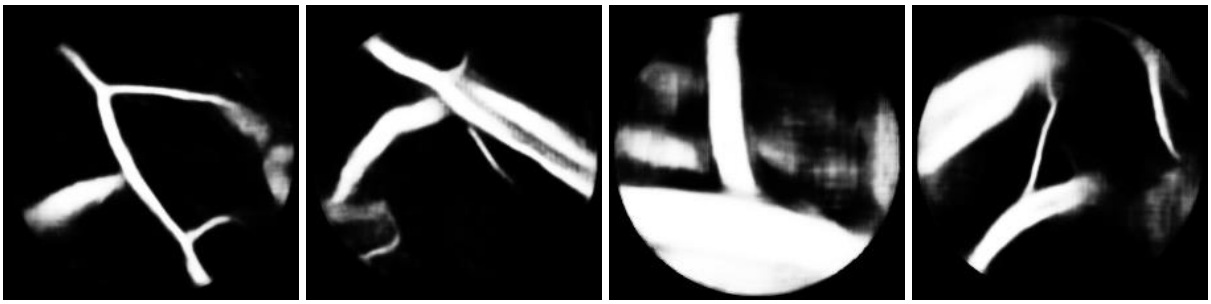


Figure 26. Examples of prediction maps generated by VesSeg.

7.2.2 Feature-Based Image Stitching

Multiple different attempts were made for the feature-based image stitching. First, ORB feature detection and matching was performed using the *in-vivo* frames without performing any preprocessing to the frames. This resulted in little to none features being detected, regardless of the input parameters used for ORB feature detection. This is in accordance with earlier experiences with our *in-vivo* dataset. Therefore, the frames were fist preprocessed to increase the number of features being detected.

Preprocessing

Multiple different preprocessing approaches were tested to evaluate the effect of preprocessing on feature detection. Based on our observations gained through experimentations, the following preprocessing method led, in general, to the highest number of features, feature matches and inlier feature matches (IFM).

First, the RGB frames were converted to the HSV (Hue, Saturation, Value) color space and contrast enhancement was applied to the Value component to enhance the luminance.[55] This was done using the CLAHE[56]. The frames were converted back to the RGB color space and the three color channels were separated. Additionally, a grayscale copy of the RGB image was created. Then, element wise matrix multiplication of the green channel and grey image was performed, followed by element wise division using the red channel. This was done to emphasize the impact of the green channel, since most contrast between the vessels and the placental surface is seen in the green channel.[72] Thereafter, image noise was reduced by applying pixelwise adaptive low-pass Wiener filtering with a neighborhood size of 3 x 3 pixels. Lastly, additional contrast enhancement was applied to the resulting image using histogram equalization.

Selective Regional Image Enhancement

The prediction and segmentation maps from VesSeg were used for selective regional image enhancement (SRIE). Prediction maps, which are the initial outputs from VesSeg, were binarized to create segmentation maps (Figure 27). For all prediction maps, a threshold value of 0.5 was used.

First, the segmentation maps were used for SRIE: the vessels (foreground) were separated from the background (Figure 27) and noise reduction was applied to the background. This was done using pixelwise adaptive low-pass Wiener filtering with a neighborhood size of 3 x 3 pixels. Thereafter, the foreground and background were fused back together. For additional experimentations, the same

steps were repeated using the prediction maps instead of the segmentation maps. Moreover, we experimented with different SRIE approaches, such as contrast enhancement in the foreground and applying a Gaussian blur ($\sigma = 3$) to the background. However, based on our experiences, these steps led to a lower number of IFM.

It was noticed that usage of the segmentation and prediction maps for SRIE leads a decrease in the number of detected and matched features, while simultaneously leading to an increase in the number of IFM. This is thought to be the result of the noise reduction which was selectively applied to the background. As a result, the total number of detected features decreases. Fortunately, the number of IFM remained stable or increased, suggesting that the 'lost' features were (mostly) false positives.

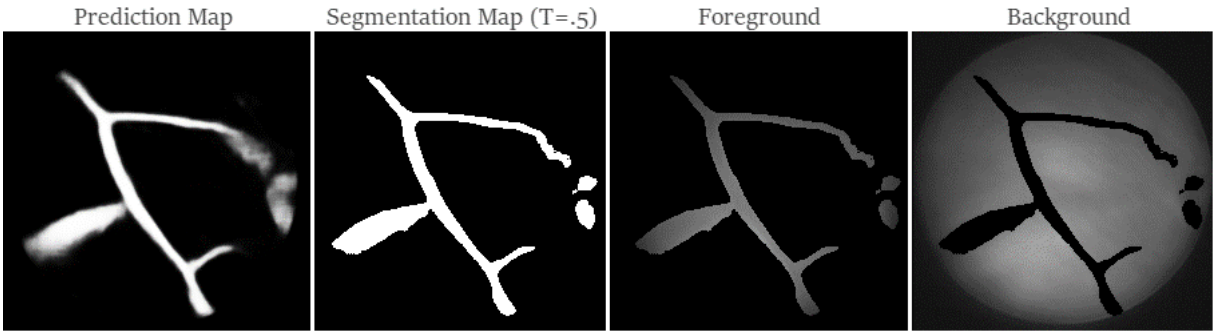


Figure 27. Prediction map, segmentation map, foreground and background of the same frame. A threshold value (T) of 0.5 was used to generate the segmentation map.

Feature Detection and Image Stitching

ORB feature detection and matching was performed using the same method and parameters as described in section 4.2.3. For the four sets of frames, image stitching was not successful for all 50 frames, meaning that either not enough IFM were found to calculate the transformation matrices or that the resulting reconstruction was visually corrupted. The input parameters for ORB feature detection and matching were manually adjusted in an attempt to improve the resulting reconstruction, but to no avail.

Figure 28 shows five consecutive frames stitched together using the prediction maps from VesSeg for SRIE. It is important to mention that only a binary circular mask was used in the blending process and no further blending techniques were used. As a result, the edges or seams of the stitched frames are clearly visible.

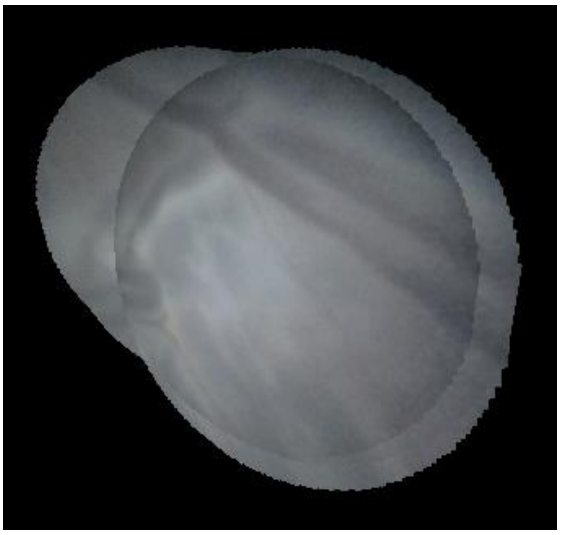


Figure 28. Feature-based image stitching of five consecutive in-vivo frames using the prediction maps from VesSeg for selective regional image enhancement (SRIE).

The use of the prediction and segmentation maps for SRIE was briefly tested for a visual SLAM[c] approach, which is designed to create 3D reconstructions. Unfortunately, we did not manage to get past the map initialization step, since not enough IFM were found.

Evaluation

Although not all 50 frames were successfully stitched together, image stitching was repeatedly successful for about five to ten consecutive frames, interchanged with incorrectly stitched frames. Here, image stitching was considered *successful* when the resulting reconstruction was in accordance with our expectations, based on visual inspection.

Based on our observations through experimentations, using the prediction or segmentation maps from VesSeg for SRIE, improved the performance of our feature-based image stitching approach. More specifically, the prediction maps seemed to outperform the segmentation maps. For now, only a couple of consecutive *in-vivo* frames were stitched together successfully, which is naturally inadequate for clinical implementation. Nevertheless, some first steps are taken into the use of prediction and segmentations maps for improved feature-based image stitching.

Limitations and Alternatives

One important limitation of our experiments is the fact that the parameters involved in the image stitching algorithms were chosen based on trial and error. Thorough parameter optimization is thought to positively affect the performance of the algorithm. The same can be said for the preprocessing and SRIE algorithms. However, it is important to be aware of the high variability of image appearances between different fetoscopic frames and videos[40], making it challenging to develop a *one size fits all* algorithm.

[c] https://mathworks.com/help/vision/ug/monocular-visual-simultaneous-localization-and-mapping.html

In our code, the frames were first preprocessed, followed by the SRIE step. One limitation here is the fact that the preprocessing step converts the color image into a grayscale image before the SRIE step, which reduces the dimensionality of the image. As a result, potential useful information might get lost. Conversely, applying SRIE to a color image is thought to allow for more image enhancement possibilities, such as saturation enhancement. Therefore, one future recommendation is to perform a SRIE method to a color image instead of a grayscale image and to further explore the possibilities for SRIE to color images.

One of the main drawbacks of feature-based image stitching of *in-vivo* fetoscopic video frames is the lack of features or IFM. As a result, the algorithm fails to estimate (appropriate) geometric transformation matrices for all sets of consecutive frames. Therefore, it is highly recommended to force the algorithm to skip 'not useful' frames. This can be done, for example, by increasing the *Confidence* argument of MATLAB's geometric transformation matrix estimation function and skipping frames in case the algorithm fails to find a transformation matrix with respect to this new confidence value. Fortunately, our fetoscopic videos are recorded at a minimum frame rate of 25 fps and consecutive frames usually contain large overlapping areas. This allows for 'safely' skipping one or multiple frames. Moreover, other studies reduced the frame rate from 25 fps to 12.5 fps[38] or 1 fps[62] for image stitching purposes, suggesting that image stitching is still feasible with less frames.

7.2.3 Intensity-Based Image Stitching

For the intensity-based approach, the transformation matrices of two consecutive frames were calculated through intensity-based image registration using the prediction maps. The image registration was done using the so-called *imregtform* function[d] build in MATLAB. Unfortunately, explicit documentation on this function is missing.

Optimization configuration was done using regular step gradient descent (RSGD)[e], for multi-resolution pyramids. Gradient descent, which is a first-order iterative optimization algorithm, aims to find local minima of the cost function. RSGD starts with a constant step length and reduces the step length by a given relaxation factor every time the gradient descent changes direction. MATLAB's default values for step length (0.0625) and relaxation factor (0.5) were used.

The number of maximum iterations was set to 200. For the other parameters, MATLAB's default settings were used. Affine geometric transformations were applied to the *moving* image, thus allowing for translation, rotation, scale, and shear transformations. Figure 29 demonstrates two consecutive frames before and after direct image registration.

Image Stitching

After performing image registration of the prediction maps, the resulting transformation matrices were used to stitching both the prediction maps and *in-vivo* fetoscopic frames together. An example of five consecutive frames is shown in Figure 30.

[d] https://mathworks.com/help/images/ref/imregtform.html
[e] https://mathworks.com/help/images/ref/registration.optimizer.regularstepgradientdescent.html

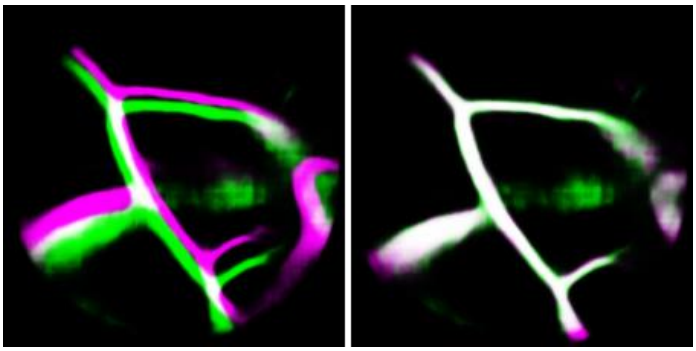


Figure 29. An overlay of two consecutive frames before (left) and after (right) direct image registration.

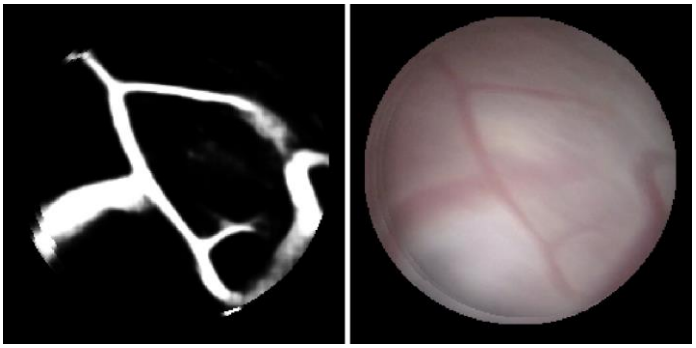


Figure 30. Intensity-based image stitching of five consecutive prediction maps from VesSeg (left). The corresponding transformation matrices were used for image stitching of the in-vivo fetoscopic frames (right).

Evaluation

The performance of our intensity-based image stitching approach was similar to that of the feature-based approach: the algorithm failed to stitch all 50 consecutive frames from the four datasets successfully. Here, image stitching was considered *successful* when the resulting reconstruction was in accordance with our expectations, based on visual inspection. Nonetheless, the intensity-based approach was also able to stitch about five to fifteen consecutive frames successfully, interchanged with incorrectly stitched frames. Although the number of successfully stitched frames were slightly higher compared to the feature-based approach, the differences were minimal.

Limitations and Alternatives

One important disadvantage of the intensity-based approach is that it is more computational costly compared to the feature-based method. However, using different programming languages and more powerful processors is thought to improve the computational time. Although the promising results reported by Bano et al.[62] provides confidence in the potential use of VesSeg for image stitching, they did not report on the speed of their algorithm. Therefore, it is unclear whether real-time intensity-based image stitching is feasible. Despite this, we consider intensity-based image stitching as a potential solution for *in-vivo* fetoscopic video frames.

In our preliminary experimentations, only one intensity-based image stitching approach is tested. It is therefore recommended to experiment with different approaches. Firstly, the approach reported by Bano et al. shows great potential. Moreover, inspiration can be taken from other methods

proposed in other fields of research. Examples are optical flow for miniature microscopic images[71], image alignment using a second-order minimization method[73] and the combination of intensity- and feature-based image stitching.[74,75]

7.4 Summary & Recommendations

Based on our experimentations, the use of VesSeg improved the performance of both the feature-based and intensity-based image stitching algorithms. In case of the feature-based approach, the prediction maps combined with SRIE showed most potential. Likewise, using the prediction maps for intensity-based showed a higher potential compared to usage of the segmentation maps.

To summarize, the pseudocodes of the feature-based and intensity-based image stitching approaches using VesSeg that showed most potential based on our experimentations are provided here:

```
Pseudocode - feature-based image stitching using VesSeg
 1  Load images
 2  Get prediction maps using VesSeg
 4  For two consecutive frames
 5        Image preprocessing
 6        Selective regional image enhancement (SRIE) using the
            prediction maps
 7        Detect features
 8        Match features
 9        Outlier rejection using MSAC
10        Estimate geometric transformation matrix (H) using the inlier
            feature matches
11        Warp one image using H
12        Combine both images
13        Blending to remove the seams
```

```
Pseudocode - intensity-based image stitching using VesSeg
 1  Load images
 2  Get prediction maps using VesSeg
 3  For two consecutive frames
 4        Estimate geometric transformation matrix (H) using direct image
    registration (optimization using regular step gradient descent)
 5        Warp one image using H
 6        Combine both images
 7        Blending to remove the seams
```

It is important to mention that our algorithms were able to stitch about five to fifteen consecutive frames, which is naturally inadequate for clinical implementation. Therefore, further research on and development of these stitching algorithms is necessary. Nevertheless, the results of our experimentations suggest that VesSeg can be of added value for feature-based and intensity-based image stitching approaches.

Finally, one of the main goals was to provide recommendations for future development of the image stitching algorithm. To summarize the recommendations provided in this chapter, an overview is provided in Table XIII.

Table XIII. Recommendations for further development of the feature-based and intensity-based image stitching algorithms.

Image stitching in general	Feature-based approach	Intensity-based approach
- Use only a fraction of the frames to reduce computational time (e.g. from 25 fps to 1-5 fps) - Experiment with other programming languages and/or functions	- Use the prediction maps combined with SRIE - Parameter optimization for ORB feature detection - Apply SRIE to color image instead of grayscale image and further explore the possibilities for SRIE - Only allow for F matrices predicted with a high confidence value	- Use prediction maps, rather than segmentation maps - Evaluate possibilities for increased computational speed - Experiment with other intensity-based methods, including the method proposed by Bano et al.[62]

8. General Discussion

This thesis focused on the training and added value of deep learning networks for improved image stitching of *in-vivo* fetoscopic video frames. First, an image classification network was trained for identification of frames with clearly visible vessels. Our network showed promising results when tested using an unseen *in-vivo* video. Moreover, high prediction rates were measured. This offers promising prospects for future clinical applications of the network. To our knowledge, we were the first the train a binary classification network for vessel detection of *in-vivo* fetoscopic video frames. A second DL network was trained for vessel segmentation. Both the quantitative and qualitative analysis showed promising results, supporting the network's potential for future applications.

Comparisons of our networks with similar networks proposed in literature showed that our networks performed either similar to these networks or outperformed them. This was an unexpected result, since our approaches were relatively simple compared to their approaches. For example, Bano et al. trained a network for fetoscopic event identification using a LSTM-RNN. to incorporate spatio-temporal information.[38] For vessel segmentations, Bano et al. trained and tested multiple U-Nets with different pre-trained backbones.[62] Although one-to-one comparisons of our networks to the networks trained by colleagues is limited due to the fact that different datasets and evaluation methods were used, our findings might suggest that more advances network architectures will not have a significant effect on the network's performance.

After training and evaluating the deep learning networks, multiple different potential applications of VesSeg for improved image stitching were explored. While doing so, a wider variety of potential uses of the network were found than initially anticipated. However, our experimentations lacked a scientific approach and thus limited conclusions can be drawn from our findings. Therefore, further research on these (and possibly other) image stitching approaches is strongly recommended.

8.1 Clinical Implementation & Applications

Before the deep learning networks can be used for image stitching in clinical practice, further development of the image stitching algorithm is required. Chapter 7 provides an overview of recommendations for development of the algorithm. Nevertheless, the deep learning networks are considered *ready* for potential other clinical implementations, based on the good performance of the networks when tested using unseen *in-vivo* data and the reported prediction rates of 7 and 714 fps.

One example of a potential application of VesSeg is to provide the surgeon with additional insights or information during FLOVA. Here, the prediction maps from VesSeg are generated and displayed to the surgeon. The idea is that the network might 'see' vessels that were otherwise missed by the human eye. This can be done using multiple different approaches, including 1) directly showing the prediction or segmentation maps or 2) superimpose the prediction or segmentation maps onto the original frames and allow for manual adjustments of the transparency of the masks if desired.

These approaches can be executed (near) real-time, depending on the processor used to perform the calculations. Moreover, it is important to mention that the goal of these approaches is to provide

additional information, while the images from the fetoscope remain the main source of information for identification of the vascular anastomoses.

8.2 Future Recommendations & Perspectives

Firstly, more research is required for further development of the image stitching algorithm. Moreover, additional experimentations with different network architectures may further improve the networks' performances. Also, investing in an advanced processor is thought to allow for faster training and application of the deep learning algorithms and faster development of the image stitching algorithm.

It is important to be aware of the fact that the deep learning networks need to be retrained after new clinical data is collected from future FLOVA procedures, for optimal performance of the networks.

Another suggestion for faster development of the image stitching algorithm is related to the test setups. Investing in a test setup that mimics the *in-vivo* situation more closely is thought to be of great value. Currently, most studies are limited to videos from previously performed FLOVA procedures, which can make retrospective image stitching more challenging compared to real-time image stitching. When testing the algorithm real-time, the operator can adjustment the scope's movements based on the progress of the real-time generated reconstruction. For example, when the signal gets lost, the operator will return to a previous position and reduce the speed of the scope's movements.

Two suggestions for improved test setups will be described. The first suggestion is to invest in a test setup that mimics the *in-vivo* FLOVA setting more realistically. This includes a hyper-realistic placenta phantom, a dark environment and a liquid that mimics to amniotic fluid. One requirement for the latter is the presence of flakes that mimic fetal skin flakes. Figure 31 demonstrates the differences in appearances of placental phantoms and an *ex-vivo* placentas compared to an *in-vivo* fetoscopic video frame. The second suggestion is to train a deep learning network that is able to convert frames from a test setup to frames that mimic the appearance of *in-vivo* fetoscopic video frames.

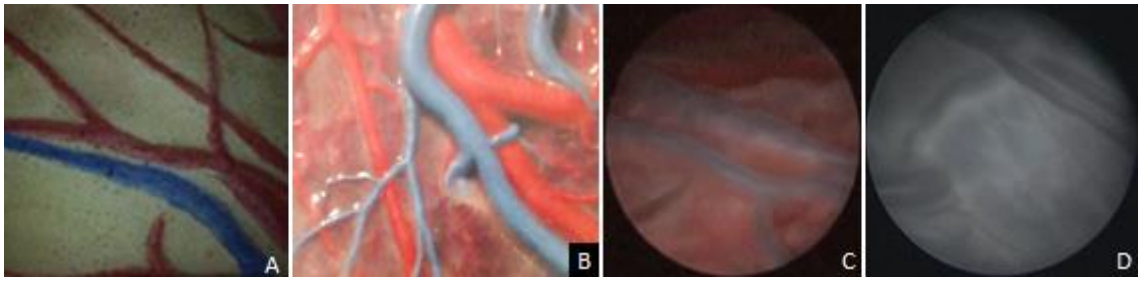


Figure 31. Example frames of different test setups and an in-vivo fetoscopic video frame. A) placenta phantom used by Yang et al.[23]; B) dye-injected placenta; C) dye-injected placenta in dark container to mimic intrauterine environment; D) frame from a FLOVA procedure (in-vivo).

Although research on the topic of FOV expansion during fetal surgery is gaining more and more attention, studies on technological innovations related to the instruments and equipment remain absent.[15] Therefore, a suggestion for future research could be to investigate the possibilities of technological innovations regarding the instruments for improves image stitching. For example, usage of a binocular would allow for more robust image stitching. A second camera could perhaps be incorporated in a fetoscope without increasing the outer diameter significantly by combining one regular sized camera with a relatively small camera equipped with a fisheye lens.

Lastly, a future perspective for the image stitching algorithm is navigation support during FLOVA. Here, the idea is that, after the placental surface reconstruction is generated, the algorithm keeps tracking the fetoscope's relative position to the placenta. The fetoscope's position is then real-time visualized.

9. General Conclusion

The aim of this thesis was to evaluate the potential use of deep learning approaches for image stitching of *in-vivo* fetoscopic video frames, for field of view expansion during fetal surgery. A classification network was trained for vessel identification, followed by a vessel segmentation network. Both networks showed promising results for future clinical applications and for further development of the image stitching algorithm. Despite the promising results, additional experiments for network optimization is recommended, including the use of different network architectures. Lastly, before the networks can be implemented in clinical practice, further development of the image stitching algorithm is required.

References

1. Bahtiyar MO, Emery SP, Dashe JS, Wilkins-Haug LE, Johnson A, Paek BW, et al. The North American Fetal Therapy Network consensus statement: prenatal surveillance of uncomplicated monochorionic gestations. Obstet Gynecol [Internet]. 2015 Jan [cited 2019 Sep 13];125(1):118–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25560113
2. Faye-Petersen OM, Crombleholme TM. Twin-to-Twin Transfusion Syndrome. Neoreviews [Internet]. 2008 Sep 1 [cited 2019 Sep 13];9(9):e370–9. Available from: http://neoreviews.aappublications.org/cgi/doi/10.1542/neo.9-9-e370
3. Cordero L, Franco A, Joy SD, O'Shaughnessy RW. Monochorionic Diamniotic Infants Without Twin-to-Twin Transfusion Syndrome. J Perinatol [Internet]. 2005 Dec 10 [cited 2019 Sep 13];25(12):753–8. Available from: http://www.nature.com/articles/7211405
4. Heineman MJ (Maas J. Obstetrie en Gynaecologie: de voortplanting van de mens. 6th ed. Elsevier gezondheidszorg; 2007. 341–351 p.
5. Cunningham FG, Williams JW (John W. Williams obstetrics. 22nd ed. McGraw-Hill Professional; 2005. 1441 p.
6. Senat M-V, Deprest J, Boulvain M, Paupe A, Winer N, Ville Y. Endoscopic Laser Surgery versus Serial Amnioreduction for Severe Twin-to-Twin Transfusion Syndrome. N Engl J Med [Internet]. 2004 [cited 2019 Sep 2];351:136–44. Available from: www.nejm.org
7. Kontopoulos E, Chmait RH, Quintero RA. Twin-to-Twin Transfusion Syndrome: Definition, Staging, and Ultrasound Assessment. Twin Res Hum Genet. 2016;19(3):175–83.
8. Emery SP, Bahtiyar MO, Moise KJ, North American Fetal Therapy Network. The North American Fetal Therapy Network Consensus Statement: Management of Complicated Monochorionic Gestations. Obstet Gynecol [Internet]. 2015 Sep [cited 2019 Sep 13];126(3):575–84. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26244534
9. Slaghekke F, Lopriore E, Lewi L, Middeldorp JM, Van Zwet EW, Weingertner AS, et al. Fetoscopic laser coagulation of the vascular equator versus selective coagulation for twin-to-twin transfusion syndrome: An open-label randomised controlled trial. Lancet. 2014;383(9935):2144–51.
10. Donepudi R, Akkermans J, Mann L, Klumper FJ, Middeldorp JM, Lopriore E, et al. Impact of cannula size on recurrent twin–twin transfusion syndrome and twin anemia–polycythemia sequence after fetoscopic laser surgery. Ultrasound Obstet Gynecol [Internet]. 2018 Dec 7 [cited 2020 Apr 29];52(6):744–9. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/uog.18904
11. Peeters SHP, Akkermans J, Westra M, Lopriore E, Middeldorp JM, Klumper FJ, et al. Identification of essential steps in laser procedure for twin-twin transfusion syndrome using the Delphi methodology: SILICONE study. Ultrasound Obstet Gynecol [Internet]. 2015 Apr 1 [cited 2020 Mar 9];45(4):439–46. Available from: http://doi.wiley.com/10.1002/uog.14761
12. Klaritsch P, Albert K, Van Mieghem T, Gucciardo L, Done' E, Bynens B, et al. Instrumental requirements for minimal invasive fetal surgery. BJOG An Int J Obstet Gynaecol [Internet]. 2009 Jan [cited 2020 May 27];116(2):188–97. Available from: http://doi.wiley.com/10.1111/j.1471-0528.2008.02021.x
13. Gautier Scientific Illustration. Tweeling Transfusie Syndroom (TTS) [Internet]. [cited 2019 Sep 17]. Available from: http://gautierillustration.com/portfolio/tweeling-transfusie-syndroom-tts-2/
14. Petersen SG, Gibbons KS, Luks FI, Lewi L, Diemert A, Hecher K, et al. The impact of entry technique and access diameter on prelabour rupture of membranes following primary fetoscopic laser treatment for twin-twin transfusion syndrome. Fetal Diagn Ther. 2016;40(2):100–9.
15. Akkermans J, Peeters SHP, Klumper FJ, Lopriore E, Middeldorp JM, Oepkes D. Twenty-Five Years of Fetoscopic Laser Coagulation in Twin-Twin Transfusion Syndrome: A Systematic Review. Fetal Diagn Ther. 2015;38(4):241–53.
16. Diehl W, Diemert A, Grasso D, Sehner S, Wegscheider K, Hecher K. Fetoscopic laser coagulation in 1020 pregnancies with twin-twin transfusion syndrome demonstrates improvement in double-twin survival rate. Ultrasound Obstet Gynecol [Internet]. 2017 Dec 1 [cited 2020 Apr 28];50(6):728–35. Available from: http://doi.wiley.com/10.1002/uog.17520
17. Cheng H, Clymer JW, Po-Han Chen B, Sadeghirad PhD B, Ferko NC, Cameron CG, et al. Prolonged operative duration is associated with complications: a systematic review and meta-analysis. J Surg Res [Internet]. 2018;229:134–44. Available from: https://doi.org/10.1016/j.jss.2018.03.022
18. Jackson TD, Wannares JJ, Todd Lancaster R, Rattner DW, Hutter MM. Does speed matter? The impact

of operative time on outcome in laparoscopic surgery. Surg Endosc. 2011;25(7):2288–95.
19. Sadda P, Imamoglu M, Dombrowski M, Papademetris X, Bahtiyar MO, Onofrey J. Deep-learned placental vessel segmentation for intraoperative video enhancement in fetoscopic surgery. Int J Comput Assist Radiol Surg [Internet]. 2018/11/30. 2019;14(2):227–35. Available from: https://doi.org/10.1007/s11548-018-1886-4
20. Yang L, Wang J, Kobayashi E, Liao H, Yamashita H, Sakuma I, et al. Ultrasound image-based endoscope localization for minimally invasive fetoscopic surgery. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2013. p. 1410–3.
21. Yang L, Wang J, Kobayashi E, Ando T, Yamashita H, Sakuma I, et al. Image mapping of untracked free-hand endoscopic views to an ultrasound image-constructed 3D placenta model. Int J Med Robot Comput Assist Surg. 2014/05/08. 2015;11(2):223–34.
22. Yang L, Wang J, Ando T, Kubota A, Yamashita H, Sakuma I, et al. Vision-based endoscope tracking for 3D ultrasound image-guided surgical navigation. Comput Med Imaging Graph [Internet]. 2014/09/30. 2015;40:205–16. Available from: http://dx.doi.org/10.1016/j.compmedimag.2014.09.003
23. Yang LJ, Wang JC, Ando T, Kubota A, Yamashita H, Sakuma I, et al. Self-contained image mapping of placental vasculature in 3D ultrasound-guided fetoscopy. Surg Endosc Other Interv Tech. 2016;30(9):4136–49.
24. Yang LJ, Wang JC, Ando T, Kubota A, Yamashita H, Sakuma I, et al. Towards scene adaptive image correspondence for placental vasculature mosaic in computer assisted fetoscopic procedures. Int J Med Robot Comput Assist Surg. 2016;12(3):375–86.
25. Gaisser F, Peeters SHP, Lenseigne BAJ, Jonker PP, Oepkes D. Stable image registration for in-vivo fetoscopic panorama reconstruction. J Imaging. 2018;4(1).
26. Peter L, Tella-Amo M, Shakir DI, Attilakos G, Wimalasundera R, Deprest J, et al. Retrieval and registration of long-range overlapping frames for scalable mosaicking of in vivo fetoscopy. Int J Comput Assist Radiol Surg [Internet]. 2018;13(5):713–20. Available from: https://doi.org/10.1007/s11548-018-1728-4
27. Sayols N, Hernansanz A, Parra J, Eixarch E, Gratacos E, Amat J, et al. Vision Based Robot Assistance in TTTS Fetal Surgery. Proc Annu Int Conf IEEE Eng Med Biol Soc EMBS. 2019;5855–61.
28. Ahmad MA, Ourak M, Gruijthuijsen C, Deprest J, Vercauteren T, Vander Poorten E. Deep learning-based monocular placental pose estimation: towards collaborative robotics in fetoscopy. Int J Comput Assist Radiol Surg [Internet]. 2020 [cited 2020 May 4]; Available from: https://doi.org/10.1007/s11548-020-02166-3
29. Torrents-Barrena J, López-Velazco R, Piella G, Masoller N, Valenzuela-Alcaraz B, Gratacós E, et al. TTTS-GPS: Patient-specific preoperative planning and simulation platform for twin-to-twin transfusion syndrome fetal surgery. Comput Methods Programs Biomed. 2019;179:104993.
30. Maneas E, Aughwane R, Huynh N, Xia W, Ansari R, Kuniyil Ajith Singh M, et al. Photoacoustic imaging of the human placental vasculature. J Biophotonics. 2019;
31. Liao H, Tsuzuki M, Kobayashi E, Dohi T, Chiba T, Mochizuki T, et al. Fast Image Mapping of Endoscopic Image Mosaics with Three-Dimensional Ultrasound Image for Intrauterine Treatment of Twin-to-Twin Transfusion Syndrome. In: International Workshop on Medical Imaging and Virtual Reality [Internet]. 2008. p. 329–338. Available from: http://link.springer.com/10.1007/978-3-540-79982-5
32. Tella-Amo M, Daga P, Chadebecq F, Thompson S, Shakir DI, Dwyer G, et al. 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. 2016.
33. Tella-Amo M, Peter L, Shakir DI, Deprest J, Stoyanov D, Iglesias JE, et al. Technical note: probabilistic visual and electromagnetic data fusion for robust drift-free sequential mosaicking. Application to fetoscopy. 2018;5(2):103.
34. Daga P, Chadebecq F, Shakir DI, Herrera LCG, Tella M, Dwyer G, et al. Real-time mosaicing of fetoscopic videos using SIFT. Med Imaging 2016 Image-Guided Proced Robot Interv Model. 2016;9786:97861R.
35. Gaisser F, Peeters SHP, Lenseigne B, Jonker PP, Oepkes D. Fetoscopic panorama reconstruction: Moving from ex-vivo to in-vivo. Commun Comput Inf Sci. 2017;723:581–93.
36. Reeff M, Gerhard F, Cattin P, Székely G. Mosaicing of Endoscopic Placenta Images. In: GI Jahrestagung. Dresden; 2006. p. 467–74.

37. Mur-Artal R, Tardós JD. ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. 2017;
38. Bano S, Vasconcelos F, Poorten E Vander, Vercauteren · Tom, Ourselin · Sebastien, Deprest J, et al. FetNet: a recurrent convolutional network for occlusion identification in fetoscopic videos. Int J Comput Assist Radiol Surg [Internet]. 2020 [cited 2020 May 4]; Available from: https://doi.org/10.1007/s11548-020-02169-0
39. Sadda P, Onofrey J, Imamoglu M, Papademetris X, Qarni B, Bahtiyar MO. Real-time computerized video enhancement for minimally invasive fetoscopic surgery. Laparosc Endosc Robot Surg [Internet]. 2018 Sep 1 [cited 2020 Mar 5];1(2):27–32. Available from: https://doi.org/10.1016/j.lers.2018.06.001
40. Vasconcelos F, Brandão P, Vercauteren T, Ourselin S, Deprest J, Peebles D, et al. Towards computer-assisted TTTS: Laser ablation detection for workflow segmentation from fetoscopic video. Int J Comput Assist Radiol Surg. 2018;13(10).
41. Bano S, Vasconcelos F, Amo MT, Dwyer GG, Gruijthuijsen C, Deprest J, et al. Deep learning-based fetoscopic mosaicking for field-of-view expansion. Int J Comput Assist Radiol Surg [Internet]. 2019 [cited 2020 Sep 1];11764. Available from: https://doi.org/10.1007/s11548-020-02242-8
42. Casella A, Moccia S, Frontoni E, Paladini D, De Momi E, Mattos LS. Inter-foetus Membrane Segmentation for TTTS Using Adversarial Networks. Ann Biomed Eng. 2020;48(2):848–59.
43. Rublee E, Rabaud V, Konolige K, Bradski G. ORB: an efficient alternative to SIFT or SURF.
44. Adel E, Elmogy M, Elbakry H. Image Stitching System Based on ORB Feature-Based Technique and Compensation Blending. Int J Adv Comput Sci Appl. 2015;6(9):55–62.
45. Tyagi D. Introduction to ORB (Oriented FAST and Rotated BRIEF) [Internet]. 2019 [cited 2019 Sep 20]. Available from: https://medium.com/software-incubator/introduction-to-orb-oriented-fast-and-rotated-brief-4220e8ec40cf
46. Rosten E, Drummond T. Fusing Points and Lines for High Performance Tracking. Tenth IEEE Int Conf Comput Vis Vol 1. 2005;2:1508–15.
47. Calonder M, Lepetit V, Strecha C, Fua P. BRIEF: Binary robust independent elementary features. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2010;6314 LNCS(PART 4):778–92.
48. Kampakis S. What deep learning is and isn't - The Data Scientist [Internet]. [cited 2021 Mar 4]. Available from: https://thedatascientist.com/what-deep-learning-is-and-isnt/
49. Schmidhuber J. Deep Learning in neural networks: An overview. Neural Networks [Internet]. 2015;61:85–117. Available from: http://dx.doi.org/10.1016/j.neunet.2014.09.003
50. Goodfellow I, Bengio Y, Courville A. Chapter 15.2 Transfer Learning and Domain Adaptations. In: Deep Learning [Internet]. MIT Press; 2016. p. 534–9. Available from: http://www.deeplearningbook.org
51. Pal L. Image classification: A comparison of DNN, CNN and Transfer Learning approach [Internet]. 2019 [cited 2020 Jun 9]. Available from: https://medium.com/analytics-vidhya/image-classification-a-comparison-of-dnn-cnn-and-transfer-learning-approach-704535beca25
52. Saha S. A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way [Internet]. 2018 [cited 2021 Mar 22]. Available from: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53
53. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2015;9351:234–41.
54. Zhou T, Canu S, Ruan S. Automatic COVID-19 CT segmentation using U-Net integrated spatial and channel attention mechanism. Int J Imaging Syst Technol. 2021;31(1):16–27.
55. Manju RA, Koshy G, Simon P. Improved Method for Enhancing Dark Images based on CLAHE and Morphological Reconstruction. Procedia Comput Sci [Internet]. 2019;165(2019):391–8. Available from: https://doi.org/10.1016/j.procs.2020.01.033
56. Zuiderveld K. Contrast Limited Adaptive Histogram Equalization. In: Heckbert PS, editor. Graphics gems IV [Internet]. San Diego, CA, USA: Academic Press Professional, Inc.; 1994. p. 474–85. Available from: http://cas.xav.free.fr/Graphics Gems 4 - Paul S. Heckbert.pdf
57. Vasconcelos F, Brandao P, Vercauteren T, Ourselin S, Deprest J, Peebles D, et al. Towards computer-assisted TTTS: Laser ablation detection for workflow segmentation from fetoscopic video. Int J Comput Assist Radiol Surg. 2018/06/29. 2018;13(10):1661–70.

58. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis. 2015;115(3):211–52.
59. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc. 2015;1–14.
60. Nash W, Drummond T, Birbilis N. A review of deep learning in the study of materials degradation. npj Mater Degrad [Internet]. 2018;2(1):1–12. Available from: http://dx.doi.org/10.1038/s41529-018-0058-x
61. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: The all convolutional net. 3rd Int Conf Learn Represent ICLR 2015 - Work Track Proc. 2015;1–14.
62. Bano S, Vasconcelos F, Shepherd LM, Poorten E Vander, Vercauteren T, Ourselin S, et al. Deep Placental Vessel Segmentation for Fetoscopic Mosaicking. Int Conf Med Image Comput Comput Interv [Internet]. 2020;1:1–11. Available from: http://arxiv.org/abs/2007.04349
63. Liskowski P, Krawiec K. Segmenting Retinal Blood Vessels With Deep Neural Networks. IEEE Trans Med Imaging. 2016;35(11):2369–80.
64. Chudzik P, Al-Diri B, Caliva F, Hunter A. DISCERN: Generative Framework for Vessel Segmentation using Convolutional Neural Network and Visual Codebook. Conf Proc . Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf. 2018;2018:5934–7.
65. Guo Y, Budak Ü, Şengür A. A novel retinal vessel detection approach based on multiple deep convolution neural networks. Comput Methods Programs Biomed. 2018;167:43–8.
66. Jiang Z, Zhang H, Wang Y, Ko SB. Retinal blood vessel segmentation using fully convolutional network with transfer learning. Comput Med Imaging Graph [Internet]. 2018;68(April):1–15. Available from: https://doi.org/10.1016/j.compmedimag.2018.04.005
67. Noh KJ, Park SJ, Lee S. Scale-space approximated convolutional neural networks for retinal vessel segmentation. Comput Methods Programs Biomed. 2019;178:237–46.
68. Goodfellow I, Bengio Y, Courville A. Chapter 7.4 Dataset Augmentation. In: Deep Learning [Internet]. MIT Press; 2016. p. 236–8. Available from: http://www.deeplearningbook.org
69. Peter L, Tella-Amo M, Shakir DI, Deprest J, Ourselin S, Iglesias JE, et al. Active Annotation of Informative Overlapping Frames in Video Mosaicking Applications. 2020; Available from: http://arxiv.org/abs/2012.15343
70. Zhu M, Wang W, Liu B, Huang J. A Fast Image Stitching Algorithm via Multiple-Constraint Corner Matching. Math Probl Eng. 2013;2013.
71. Loewke KE, Camarillo DB, Piyawattanametha W, Mandella MJ, Contag CH, Thrun S, et al. In vivo micro-image mosaicing. IEEE Trans Biomed Eng. 2011;58(1):159–71.
72. Lurie KL, Angst R, Zlatev D V, Liao JC, Ellerbee Bowden AK. 3D reconstruction of cystoscopy videos for comprehensive bladder records. Biomed Opt Express [Internet]. 2017 Apr 1 [cited 2019 Sep 20];8(4):2106–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28736658
73. Lovegrove S, Davison AJ. Real-time spherical mosaicing using whole image alignment. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2010;6313 LNCS(PART 3):73–86.
74. Richa R, Linhares R, Comunello E, Von Wangenheim A, Schnitzler JY, Wassmer B, et al. Fundus image mosaicking for information augmentation in computer-assisted slit-lamp imaging. IEEE Trans Med Imaging. 2014;33(6):1304–12.
75. Papademetris X, Jackowski AP, Schultz RT, Staib LH, Duncan JS. Integrated intensity and point-feature nonrigid registration. Lect Notes Comput Sci. 2004;3216(PART 1):763–70.