

Unsupervised Representation Learning from Intravascular Ultrasound Videos

by

Lay Jain

S.B., Computer Science and Engineering and Economics, Massachusetts
Institute of Technology (2022)

Submitted to the Department of Electrical Engineering and Computer
Science

in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

© 2023 Lay Jain. All rights reserved.

The author hereby grants to MIT a nonexclusive, worldwide, irrevocable, royalty-free license to exercise any and all rights under copyright, including to reproduce, preserve, distribute and publicly display copies of the thesis, or release the thesis under an open-access license.

Authored by: Lay Jain
Department of Electrical Engineering and Computer Science
May 12, 2023

Certified by: Polina Golland
Professor
Thesis Supervisor

Accepted by: Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Unsupervised Representation Learning from Intravascular Ultrasound Videos

by

Lay Jain

Submitted to the Department of Electrical Engineering and Computer Science
on May 12, 2023, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Vascular diseases such as atherosclerosis are a leading cause of mortality and morbidity worldwide. Intravascular Ultrasound (IVUS) is an imaging technology that has the distinctive ability to offer real-time endovascular information of the coronary vasculature. However, its low signal-to-noise ratio, low data availability, and numerous artifacts make it challenging to use both for humans and automated methods. This work explores the use of representation learning and de-noising techniques to address these challenges and aid in the diagnosis of vascular diseases. We test our methods on the task of stent malapposition detection, where naive approaches fail discouragingly. We improve the naive baseline accuracy by 16%. In addition, we develop a deep learning approach for real-time stabilization of the IVUS videos, which performs registration 20-fold faster than the classical ANTs approach. Our results demonstrate the importance of incorporating domain knowledge in performance improvement while still indicating the limitations of current systems for achieving clinically ready performance.

Thesis Supervisor: Polina Golland

Title: Professor

Acknowledgments

First and foremost, I would like to thank my advisor, Dr. Polina Golland, for her support and thoughtful guidance. I would also like to express my gratitude to my immediate supervisor, Dr. Neerav Karani, for his invaluable ideas and prompt feedback throughout the duration of the project. I would also like to acknowledge my lab-mates at the Medical Vision Group for their help. In particular, I want to express my gratitude to Dr. Neel Dey for his excellent advice and assistance during multiple meetings throughout the project. I am also grateful to Ray Wang for running experiments closely related to the project. The project could not have been completed without their help.

I would also like to extend my gratitude to collaborators outside of MIT. I would like to thank Dr. Satyandra Kashyap and Dr. Niharika D'Souza at IBM Research for their invaluable insights. Their expertise and research results have significantly enhanced the quality of my work. I also express my sincere gratitude to our collaborators at Boston Scientific for their incredible help in connecting us with clinicians. Their contributions helped me gain useful insights about the data and incorporate domain knowledge, which greatly benefited the project.

Finally, I would like to thank my parents – Bhavna Jain and Anil Kumar Jain – for their encouragement and support towards pursuing my education. This experience has been enriching and rewarding.

Contents




1	Introduction	21
2	Background and Related Work	25
2.1	Background	25
2.1.1	Blood vessel diseases	25
2.1.2	Intravascular ultrasound imaging	26
2.2	Related Work	27
2.2.1	Supervised Learning from IVUS	27
2.2.2	Unsupervised representation learning from videos	27
2.2.3	Jitter removal from IVUS	28
3	Dataset	29
3.1	Labels and Dataset Split	29
3.2	Motion in IVUS pullbacks	33
4	Methods	35
4.1	Supervised malapposition detection	36
4.1.1	Loss functions	37
4.1.2	Encoding invariance in the classifier	38

4.1.3	Representation choice: Cartesian vs. polar coordinates	39
4.1.4	Including temporal information	40
4.1.5	Optimization choices	40
4.2	Unsupervised representation learning from unlabelled pullbacks . . .	41
4.2.1	Contrastive Random Walks on Video	42
4.2.2	VideoMAE - Masked Autoencoders	44
4.3	Jitter removal from pullbacks	45
4.3.1	Registration using ANTs	46
4.3.2	Registration using deep-learning	48
5	Experiments and Results	51
5.1	Supervised malapposition detection	51
5.1.1	Loss functions	51
5.1.2	Encoding invariance in the classifier	53
5.1.3	Representation choice: Cartesian vs. polar coordinates	54
5.1.4	Including temporal information	54
5.1.5	Optimization choices	54
5.2	Unsupervised representation learning from unlabelled pullbacks . . .	55
5.2.1	Contrastive Random Walks on Video	55
5.2.2	VideoMAE - Masked Autoencoders	58
5.3	Jitter removal from pullbacks	59
5.3.1	Registration using ANTs	59
5.3.2	Registration using deep-learning	62
6	Discussion and Conclusion	65

6.1	Discussion	65
6.1.1	Future Directions	66
6.2	Conclusion	67
A	Classification Metrics	69
B	Fold-wise Breakdown of the Dataset	73

List of Figures

- 1-1 Malapposition in intravascular ultrasound images. Figure 1-1a shows a healthy lumen, with blood showing as speckles in the middle. The bright spots in 1-1b are stent struts. On the bottom right, one can see blood speckles outside of the stent boundary. This is the stent gap, highlighted in figure 1-1c. Thus, the figure 1-1b has malapposition. Chapter 2 provides more background on IVUS, stents and malapposition. 22
- 1-2 A preview of our motion correction results for IVUS videos. The figure shows lateral cross-section of the input and output videos. The black region in the centre is the ultrasound catheter. In the input video, the catheter doesn't move but the vessel around it moves, exhibiting random jitter (due to the motion of catheter) superimposed with periodic motion (caused by heartbeat) that can be seen as saw-teeth in the bright lumen boundary. The output transfers the motion and jitter back to the catheter, making the lumen boundary more stationary across frames. 23
- 2-1 Longitudinal section of an artery showing stent placement to prevent narrowing down of the vessel due to calcification. The areas marked by \star show stent gap (malapposition). Adapted from [41]. 25

3-1	Some arterial intravascular ultrasound images from the dataset. The bottom sub-figure 3-1b shows images where the stent is not placed incorrectly (malapposed), whereas the top figure (3-1a) shows pictures of healthy lumen and stent. Without explicit training, it is almost impossible for the human eye to distinguish between the two classes.	30
3-2	Cartesian images (top) and their corresponding polar representations (bottom). Here, figures (a)-(c) show healthy lumen and stent (label (-)) and (d)-(f) show malapposed stent placement (label (+)).	31
3-3	Some stent gap annotations. Recall that the stent gap is the gap between a malapposed stent and the vessel wall (section 2.1). Unfortunately, only few malapposed frames are annotated with the stent gap.	31
3-4	The number of malapposed and unlabelled frames present across the 28 labelled pullbacks. Notice that only around 10% of the frames are malapposed, resulting in significant class imbalance.	32
3-5	Assigning labels to the frames. Each subfigure shows one pull-back, with frames arranged in order of acquisition. The frames colored in teal  are marked as malapposed and assigned the label (+). The frames colored in purple  are within a distance of 100 from some malapposed frame, and are therefore discarded due to ambiguity. The rest of the frames, colored yellow  are assigned the label (-). Notice that malapposition occurs in contiguous segments.	32
3-6	200 × 200 correlation matrix of the first 200 frames of an input intravascular video. Notice the periodic stripes.	33

3-7	(a) A plot of the correlation of the first frame with the rest of the video. Note the distinctive peaks caused by frames in the same phase of the cardiac cycle. In this figure, the peaks occur at frames 24, 48, and 72. (b) Cross-sectional view of the video. The red lines mark the frames in phase.	34
4-1	Intensity and geometric augmentations. Figure 4-1a shows the original image in the dataset. Figure 4-1b shows a blurred and jittery version (intensity augmentation). Figure 4-1c shows a version after flipping and affine transforms (geometric augmentation).	39
4-2	A Cartesian image and its polar counterpart, with and without stent gap segmentation. Notice that in 4-2c, the two red parts actually refer to nearby real locations, but a simple CNN doesn't take this into account. Using reflect padding and augmentations partly remedy this issue.	40
4-3	Evolution of learning rate with the step-decay, cosine-annealing, and one-cycle schedules.	41
4-4	Contrastive Random Walks on Video. Every frame is divided into patches and the video is represented as a graph where the nodes are patches and edges are patch affinities. A contrastive loss encourages the paths that reach a target, implicitly supervising latent correspondence along the path. Learning proceeds without labels on palindrome frame sequences. Taken from [26].	42
4-5	VideoMAE architecture. Input videos are divided into patches and patches are randomly masked with a high masking ratio (80-95%). The mask is kept same for a tubelet of frames, and a vision transformer model is trained using an in-painting task with a mean squared error loss. Taken from [57].	44

4-6	Proposed registration network architecture. Input video is passed through a an encoder f_θ with two linear heads for rotation and translation parameters respectively. The video is then transformed by sampling from a transformation grid generated using the predicted parameters. Another network, g_ψ , independently predicts the template, which is used to evaluate the model performance in the loss function. Additionally, we apply a regularization penalty on the predicted transformation parameters.	49
5-1	Out-of-sample performance on the pre-training task of finding correspondences in a cycle. We plot the average fraction of correctly mapped patches with training step across cycles of different lengths. The model in figure 5-1a was trained on clips of three frames, thus finding correspondences on cycles of length 2 and 4. On the other hand, the model in figure 5-1b was trained on the task of finding correspondences on cycles of length up to 20. For both the models, the input frames were divided into $N = 49$ overlapping patches using a 7×7 grid. Notice that the second task is much harder than the first, as also demonstrated by the low probabilities of correct return in figure 5-1b.	57
5-2	Frames X_i of an input clip (top) and their corresponding propagated labels \hat{L}_i (bottom). Here, column (a) is the annotated first frame of the clip (X_0, L_0). Columns (b)-(f) are the succeeding frames, and their propagated labels X_i, \hat{L}_i	58
5-3	Pre-training using VideoMAE. The figure shows a sample output from the VideoMAE pre-training in-painting task applied on the IVUS images. Figure 5-3a shows the original frame, which is masked (figure 5-3b) and fed as an input to the model. Figure 5-3c shows the output of a trained model on the given input. Notice the checkerboard artifacts in the output. This particular model was trained over clips of length $T = 12$ with a masking ratio $\rho_{mask} = 50\%$	58

5-4	Results of applying procedure in algorithm 2 to an intravascular ultrasound video of length $T = 600$. Notice how the algorithm is able to reduce the saw-tooth pattern in figure 5-4a to a much smoother vessel boundary in figure 5-4b, thus de-noising the jitters caused by the cardiac cycle.	60
5-5	Results of applying procedure in algorithm 2 to an intravascular ultrasound video of length $T = 600$. Notice how the algorithm is able to significantly increase the correlation between the frames of the video. Notice also that the procedure is able to significantly reduce but not eliminate the periodic peaks caused by the cardiac cycle.	61
5-6	Results of using registration network for the task of aligning digits. A ResNet-18 encoder was used with a rotation head of depth 2 and a U-Net with 3 channels. The loss consisted of the MSE registration loss with respect to the U-Net output template and an l_2 penalty loss with $\lambda = 0.1$. The clip length was chosen to be $T = 30$	62

List of Tables

5.1	The effect of using a weighted loss functions on various classification metrics. Notice that weighting the positive samples leads to higher TPR and BA, by increasing true and false positives and reducing false negatives.	52
5.2	The effect of using Focal Loss (section 4.1.1) on the task of malapposition detection. All the experiments were run with intensity and geometric augmentations (section 4.1.2).	52
5.3	The effect of data augmentations on classification accuracy. The numbers are average over four folds. Notice that both intensity and geometric transformations are useful for our task.	53
5.4	The effect of using a E(2)-equivariant steerable convolutional neural network for the malapposition classification task.	53
5.5	The effect of using polar coordinates for the malapposition classification task. Notice the jump in accuracy, and the similarity with table 5.4.	54
5.6	The effect of using clips of different lengths T_{clip} on the malapposition classification task. All the experiments are performed with One-Cycle cosine annealing learning rate schedule, and intensity and geometric augmentations.	54

5.7	The effect of using different popular learning rate schedulers for the malapposition classification task.	55
5.8	The effect of pre-training using contrastive random walks on videos on the classification accuracy. A ResNet-18 encoder was trained using the random walk objective [26], and a depth-2 MLP was added for the downstream classification fine-tuning. The pre-training was done on cycle-lengths ≤ 4 , and the patches and frames were randomly cropped and resized. The accuracy values are average over 4 folds. Notice that pre-training doesn't help once we add data augmentation.	56
5.9	The effect of pre-training using VideoMAE on the balanced accuracy for the malapposition task. The transformer encoder was pretrained using clips of length $T = 12$, with a masking ratio of $\rho_{mask} = 95\%$ using tubelets of length $t = 2$. During pre-training, clips were augmented with intensity and geometrical transforms.	59
5.10	Results of clip stabilization using ANTs (algorithm 1). Notice the benefit of using multi-scale registration (listing 5.1).	60
5.11	Results of stabilization of a single clip using ANTs and the registration network. The first row shows the statistics of the original clip (without any stabilization). In the last two rows, the registration network is trained on MSE loss with no penalty ($\lambda = 0$).	63
5.12	Results of clip stabilization using Registration network for different penalties on the predicted parameters. Each row reports the results from best of three values of the regularization hyperparameter λ : 0.01, 0.1, 1. Compare this to the ANTs results, table 5.10.	64
5.13	A comparison of performance on unseen data. Note that while ANTs performs a new optimization on every input, the registration network simply performs one forward pass in order to predict the transformation parameters.	64

5.14	Run-time for the three stabilization approaches. Reported values are an average over 3000 clips of length $T = 30$	64
B.1	The patient IDs included in the four folds.	73
B.2	Fold-wise breakdown of malapposed (label (+)), well-apposed (label (-)), and discarded frames. The labels were assigned based on the strategy depicted in figure 3-5.	73

Chapter 1

Introduction

Vascular diseases, such as atherosclerosis, are a leading cause of mortality and morbidity globally [14]. Intravascular Ultrasound (IVUS) is an imaging technology that has the distinctive ability to offer real-time endovascular information of the coronary vasculature [35, 1] and has been shown to have a positive impact on clinical outcomes for percutaneous coronary interventions [51, 58]. Despite this, it is not yet widely adopted owing to its low signal-to-noise ratio as well as numerous imaging artifacts that make diagnosis time consuming, error prone and hard, both for machines and the trained human eye. The resulting high cost of manual annotations limits the availability of labelled data for training machine learning methods, making the task of automated analysis even more challenging.

One such challenging task is that of malapposition detection. Malapposition refers to the lack of contact between a stent and the underlying intimal surface of the arterial wall [49]. Identifying malapposition from IVUS images is not a straightforward task (figure 1-1) and requires a significant amount of time and effort, even for experienced practitioners. Indeed, straightforward classification approaches only achieve almost random accuracy on this task.

To address these challenges, this work explores the use of representation learning and de-noising techniques to aid in the diagnosis of vascular diseases. We evaluate

our learned representations by focusing on the downstream task of malapposition detection. This thesis consists of two major parts. In the first part, we explore various methods such as self-supervised pre-training, data augmentations, and temporal methods, and improve performance on malapposition detection. In the second part, we propose a stabilization method for the IVUS videos in order to mitigate motion-related noise, which can be a major barrier to diagnosis.

Contributions: With these methods, we are able improve the balanced accuracy for malapposition classification from 58% to 74%. We also quantify and correct for motion and heartbeat artifacts in the acquired intravascular ultrasound videos (figure 1-2). Furthermore, our registration network is capable of achieving video stabilization comparable to traditional methods, but with a 20-fold speedup in running time, paving the way for real-time noise reduction during acquisition. Our results address the importance of incorporating domain knowledge in performance improvement while still indicating the limitations of current systems for achieving clinically ready performance.

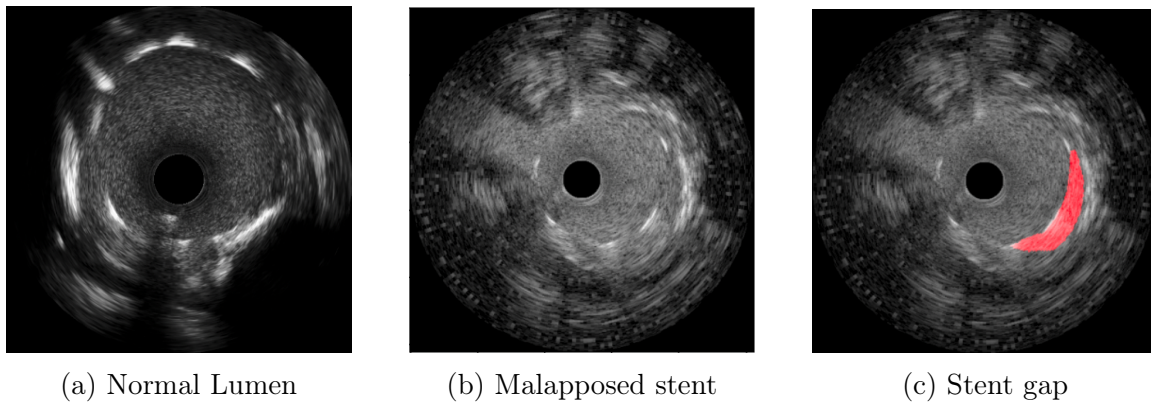


Figure 1-1: Malapposition in intravascular ultrasound images. Figure 1-1a shows a healthy lumen, with blood showing as speckles in the middle. The bright spots in 1-1b are stent struts. On the bottom right, one can see blood speckles outside of the stent boundary. This is the stent gap, highlighted in figure 1-1c. Thus, the figure 1-1b has malapposition. Chapter 2 provides more background on IVUS, stents and malapposition.

Thesis Organization: The rest of this thesis is organized as follows. Chapter 2 provides a brief background of blood vessel diseases and intravascular ultrasound

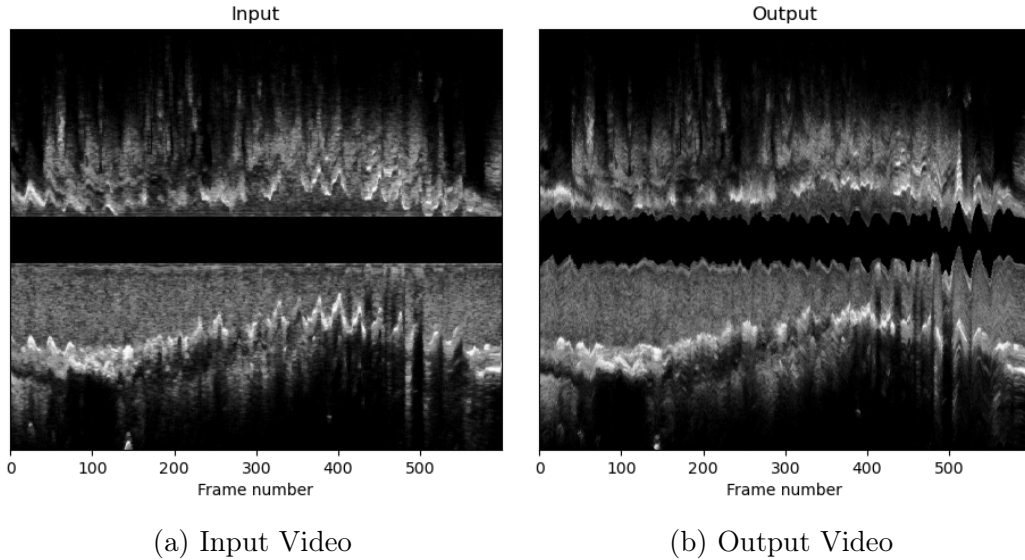


Figure 1-2: A preview of our motion correction results for IVUS videos. The figure shows lateral cross-section of the input and output videos. The black region in the centre is the ultrasound catheter. In the input video, the catheter doesn't move but the vessel around it moves, exhibiting random jitter (due to the motion of catheter) superimposed with periodic motion (caused by heartbeat) that can be seen as saw-teeth in the bright lumen boundary. The output transfers the motion and jitter back to the catheter, making the lumen boundary more stationary across frames.

imaging (section 2.1), and discusses related work (section 2.2). Chapter 3 provides an overview of the IVUS dataset and some preliminary profiling. Chapter 4 establishes methods to create strong baselines (section 4.1), and discusses methods for learning representations using various pre-training (section 4.2) and registration (section 4.3) approaches. We then provide the results of experiments using these methods in chapter 5. We discuss these results in chapter 6 and conclude in section 6.2.

Chapter 2

Background and Related Work

2.1 Background

2.1.1 Blood vessel diseases

All blood vessels have a *lumen*, which is a hollow passageway through which blood flows. Sometimes, cholesterol plaque builds up in the walls of the arteries, obstructing blood flow. This condition is known as *atherosclerosis*. In later stages, atherosclerosis can cause deposition of calcium in the arteries, making the walls hard and unable to expand. This condition is referred to as *calcification*. Atherosclerosis and calcification in the arteries can lead to serious blockages in blood flow, and can lead to strokes and heart attacks.

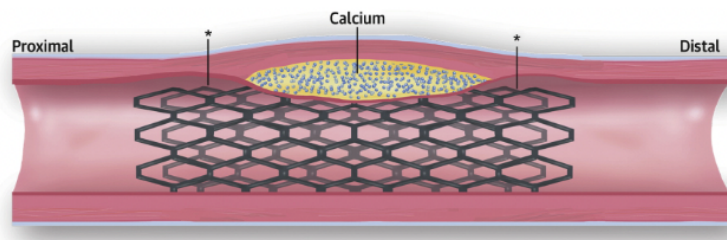


Figure 2-1: Longitudinal section of an artery showing stent placement to prevent narrowing down of the vessel due to calcification. The areas marked by \star show stent gap (malapposition). Adapted from [41].

In order to prevent further narrowing of the vessel and to expand the walls to maintain normal blood flow, a wire-mesh tubular support called a *stent* is inserted into the vessel. A stent is said to be *malapposed* if it fails to adhere to the walls of the vessel (figure 2-1). Most operators strictly advocate complete apposition of the stent struts to the wall, because of the risk of thrombosis caused by protrusion of stents into the blood field [44]. *Thrombosis* refers to the formation of a blood clot in a blood vessel, and can lead to serious illnesses and disabilities.

2.1.2 Intravascular ultrasound imaging

Intravascular ultrasound (IVUS) is used to image the interior of blood vessels. It provides real-time images of the vessel's interior and helps identify blood-vessel diseases and stent malapposition. IVUS imaging supplements angiography (based on x-rays), which shows 2D silhouettes of vessels. During the IVUS image acquisition, an ultrasound catheter is taken to a particular location inside a blood vessel and frames are acquired as it is pulled back either manually or using a motor. The acquired temporal sequence of frames is called a *pullback*.

Despite its many advantages, IVUS imaging has some limitations. The quality of the images can be affected by several factors, including the size and shape of the blood vessel, the presence of blood or tissue, and the patient's anatomy. Furthermore, the IVUS images are prone to various artifacts such as elliptic distortions (caused due to angulated position of catheter), phantom walls or plaques (caused by multiple echoes), impulse responses, near field (Fresnel zone) impacts, ring down artifacts, and air bubble and guidewire artifacts [16]. This makes diagnosis of blood vessel diseases hard and time-consuming.

2.2 Related Work

2.2.1 Supervised Learning from IVUS

Only a little works exists on supervised learning of IVUS images and detection of malapposition. Most of the existing work on IVUS analysis focuses on dense prediction tasks such as segmentation of lumen and media–adventitia border [66, 12, 5, 43], plaque [67, 10, 33, 45, 4], and stent [11, 43] and uses classical approaches [2]. To the best of our knowledge, only one study [40] has focused on malapposition detection. However, their approach requires access to paired pre-stenting and post-stenting IVUS pullbacks and is therefore inapplicable in a vast number of clinical use cases. Furthermore, no existing work considers the very real problem of limited annotations in the context of IVUS images.

2.2.2 Unsupervised representation learning from videos

While there has been a flurry of advances in learning image representations [30, 27, 38, 62, 21, 7], video has often simply been treated as a simple extension of image into another, time dimension. However, this approach is limiting since it fails to account for temporal-correspondence – “what went where” [15]. Recent approaches to self-supervised learning from videos have started taking into account the notion of temporal correspondence and high auto-correlation in the time dimension. For instance, [26, 34, 60, 63] use the idea of temporal correspondence by learning correspondences via cycle-consistency of time. Another line of work [46, 53] predicts video clips with autoencoders in pixel-space for representation learning by using CNN or LSTM backbones. Pre-training on videos done by using masked modelling [23, 61, 56] has also demonstrated success with various architectures, most recently by using a vision transformer backbone [57].

Unsupervised representation learning has been used for medical data of various modalities [8, 28, 25] including ultrasound[29, 9, 19]. To the best of our knowledge,

no works have looked at the application of pre-training methods to intravascular ultrasound pullbacks and malapposition detection.

2.2.3 Jitter removal from IVUS

Numerous methods have been developed to register image pairs [59, 6] and to reconstruct 3D volumes from slices [31, 20]. For IVUS images, most of the methods take traditional, non deep-learning, approaches. In [54], same cardiac phase frames are selected through clustering, leading to discarding of a majority of the frames and a consequent loss of useful information. In [18], the IVUS images are heavily blurred and the vessel wall is modelled as a circle, which is then used to estimate the rigid parameters. However, the method is sensitive to the blurring threshold and fails to generalize on pullbacks with cross sections of high eccentricity. Recently, [55] proposed the first deep learning method for motion correction in IVUS, which is based on a generative adversarial network. However, their method is trained on a simulated dataset which fails to characterize the motion of coronary vascular structures in different scenarios such as after stent deployment or under atherosclerosis.

Chapter 3

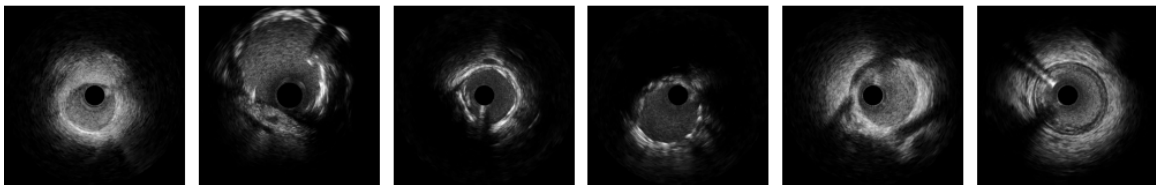
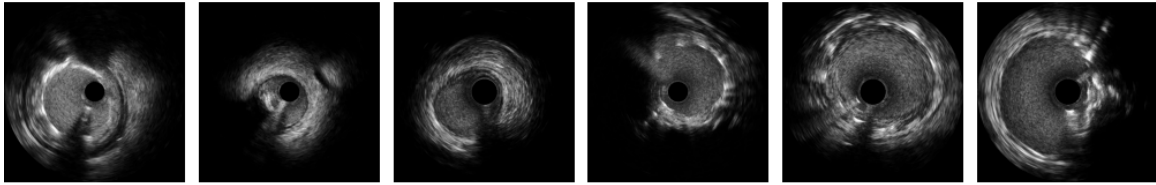
Dataset

In this chapter, we describe and characterize the intravascular ultrasound dataset on which our methods are evaluated.

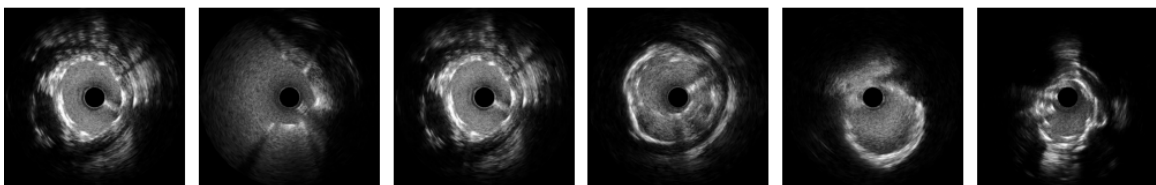
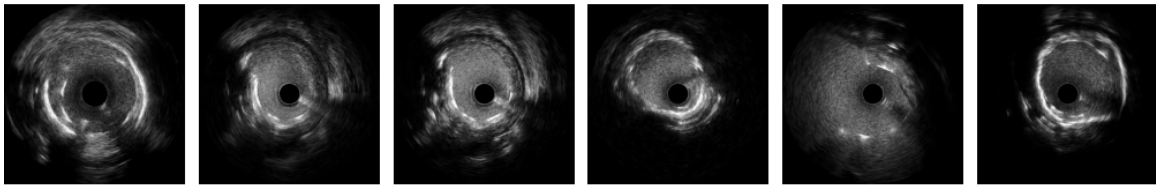
The overall dataset consists of 72 arterial pullbacks. Each pullback consists of between 1000 and 3000 frames, each of size 256×256 . Figure 3-1 shows frames from some of the pullbacks in the data. The black circle in the center represents the catheter. Note that the images are converted from original polar images taken by the catheter. Figure 3-2 compares the polar and Cartesian representations. Note that the catheter center is simply radial padding in the polar picture.

3.1 Labels and Dataset Split

Out of the 72 arterial pullbacks, only 28 are labelled for malapposition. Figure 3-1 compares normal and malapposed lumen. In total, there are 3695 malapposed frames. Out of these, 83 have pixel level stent gap annotations. Figure 3-3 shows some of these annotated frames with the stent gap overlaid in red on top of the corresponding image.



(a) Healthy lumen and stent, label (-)



(b) Malapposed stent, label (+)

Figure 3-1: Some arterial intravascular ultrasound images from the dataset. The bottom sub-figure 3-1b shows images where the stent is not placed incorrectly (malapposed), whereas the top figure (3-1a) shows pictures of healthy lumen and stent. Without explicit training, it is almost impossible for the human eye to distinguish between the two classes.

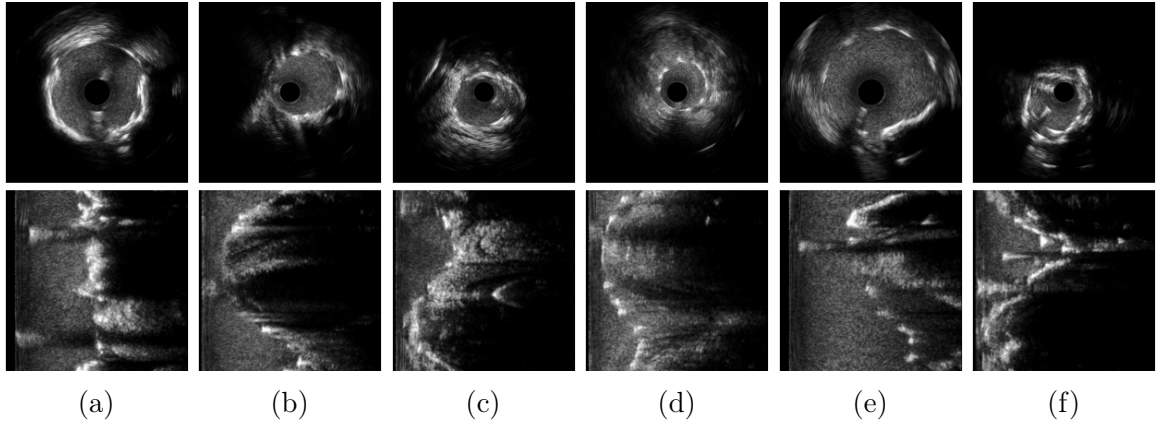


Figure 3-2: Cartesian images (top) and their corresponding polar representations (bottom). Here, figures (a)-(c) show healthy lumen and stent (label $(-)$) and (d)-(f) show malapposed stent placement (label $(+)$).

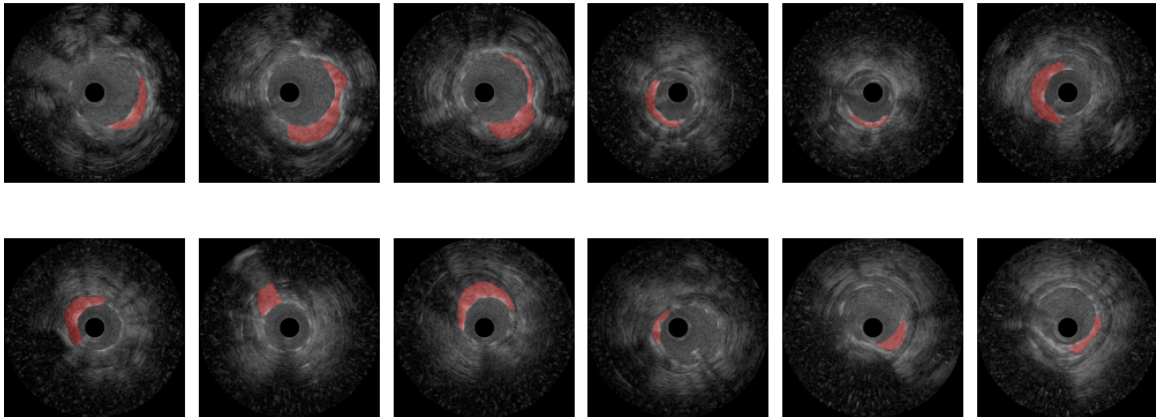


Figure 3-3: Some stent gap annotations. Recall that the stent gap is the gap between a malapposed stent and the vessel wall (section 2.1). Unfortunately, only few malapposed frames are annotated with the stent gap.

In this work, we use the downstream classification task of malapposition detection in order to evaluate our learned representations. Frames labelled as $(+)$ are malapposed, whereas the remaining frames may or may not be non-malapposed. Figure 3-4 shows the number of malapposed labels present across the 28 pullbacks.

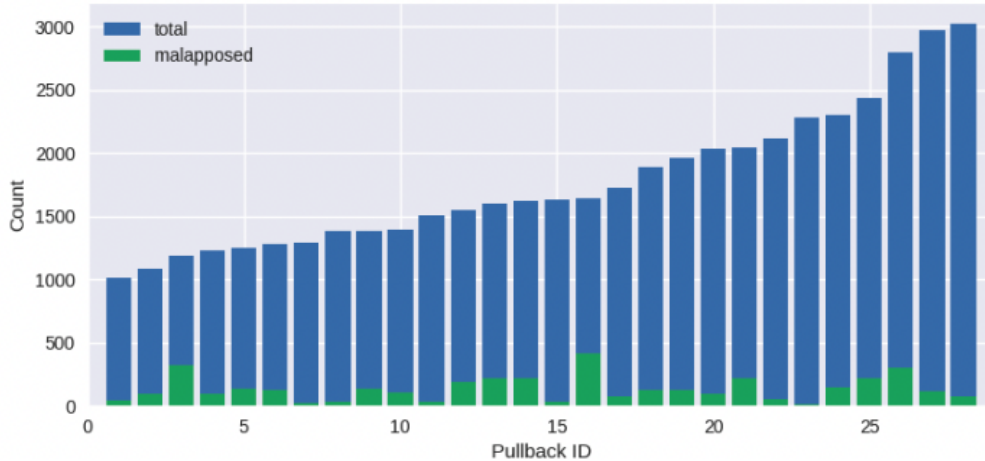


Figure 3-4: The number of malapposed and unlabelled frames present across the 28 labelled pullbacks. Notice that only around 10% of the frames are malapposed, resulting in significant class imbalance.

In our experiments, we assume that any frame not within a distance of 100 frames from any malapposed frame is non-malapposed, and label it as (-). Figure 3-5 pictorially portrays this labeling strategy. The resulting dataset has 3695 malapposed and 37259 well-apposed frames, resulting in a class imbalance (-)/(+) of ≈ 10 . As figure 3-5 depicts, the malapposed frames are clustered into temporal segments. Informally speaking, the high auto-correlation between adjacent frames means that the effective sample size is much lower.

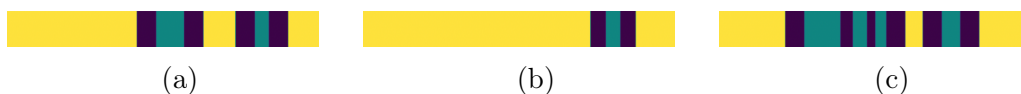


Figure 3-5: Assigning labels to the frames. Each subfigure shows one pull-back, with frames arranged in order of acquisition. The frames colored in teal are marked as malapposed and assigned the label (+). The frames colored in purple are within a distance of 100 from some malapposed frame, and are therefore discarded due to ambiguity. The rest of the frames, colored yellow are assigned the label (-). Notice that malapposition occurs in contiguous segments.

3.2 Motion in IVUS pullbacks

A major challenge in the applications of IVUS imaging is the presence of motion artifacts. Contraction and relaxation of the heart and the pulsatile blood flow in the lumen cause the catheter tip to move laterally relative to the lumen in a periodic fashion [54]. In addition to the cardiac motion, the videos also exhibit artifacts arising from arterial vasomotion, and catheter motion such as bending, longitudinal oscillations, and non-uniform rotational distortion [55, 48].

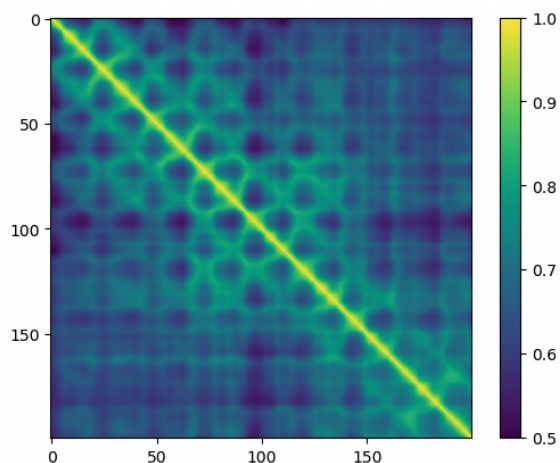


Figure 3-6: 200×200 correlation matrix of the first 200 frames of an input intravascular video. Notice the periodic stripes.

The videos in our dataset also exhibit a noticeable periodic motion caused by the cardiac cycle. Here we profile this motion by looking at normalized cross correlation (NCC) between frames. The NCC between two images $A, B \in [0, 1]^{H \times W}$ is defined as:

$$\text{NCC}(A, B) := \frac{1}{H \times W} \cdot \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \frac{(A_{ij} - \mu(A))(B_{ij} - \mu(B))}{\sigma(A) \cdot \sigma(B)} \quad (3.1)$$

where $\mu(\cdot), \sigma(\cdot)$ denote, respectively, the intensity mean and standard deviation.

For a video $X \in [0, 1]^{T \times H \times W}$ of T frames, each of dimension $H \times W$, the cross-correlation matrix is the $T \times T$ matrix of frame-wise normalized cross correlations:

$$\text{Corr}(X)_{ij} = \text{NCC}(X[i], X[j]) \quad (3.2)$$

Figure 3-6 shows the correlation matrix of an IVUS pullback clip. We notice that the correlations exhibit a periodic pattern. Figure 3-7a explicitly illustrates this pattern by plotting the first row of the correlation matrix. We observe a periodic trend superimposed on a decaying correlation. The periodic trend is caused by the cardiac cycle. A simple way to find the time period of the cardiac cycle is to extract the peaks and look at distance between them. In the pullback depicted in figure 3-7a, the time-period is $T_{peak} = 24$ frames per cycle. The ultrasound videos are taken at 30 frames per second. Thus, the time period is $60 \times 30/24 = 75$ beats per minute, which matches with the average human resting heart rate.

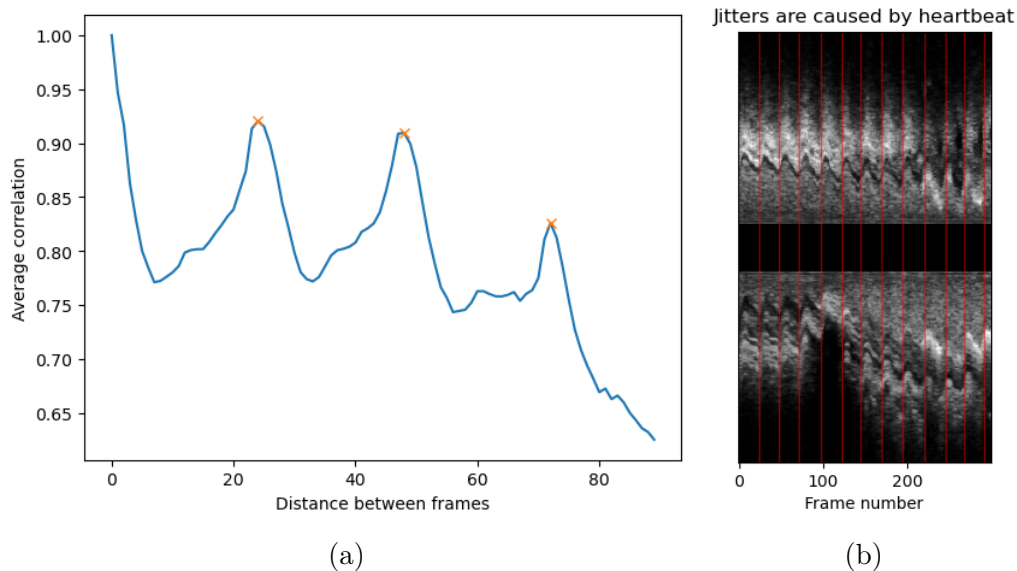


Figure 3-7: (a) A plot of the correlation of the first frame with the rest of the video. Note the distinctive peaks caused by frames in the same phase of the cardiac cycle. In this figure, the peaks occur at frames 24, 48, and 72. (b) Cross-sectional view of the video. The red lines mark the frames in phase.

Figure 3-7b shows the lateral view of the IVUS video. The red lines represent peaks extracted by this method. As expected, the peaks coincide with the periodic sawteeth caused by periodic motion. In section 4.3, we will discuss methods to remove these jitters and use correlation as a registration metric for quantifying motion correction.

Chapter 4

Methods

We start by investigating fully supervised methods. To tackle challenges such as class imbalance and the small sized annotated dataset, we augment the simple supervised learning with different techniques aimed at learning from small amounts of data or exploiting various symmetries, which are discussed in detail in section 4.1.

However, as discussed in chapter 3, a majority of our data is unannotated. In order to leverage this unlabelled data, we explore various unsupervised representation learning methods. We aim to use the unlabelled IVUS videos in order to learn representations which can be efficiently fine-tuned for downstream tasks using a relatively small annotated dataset. A common way to learn such representations is to carry out pre-training on unlabelled videos, by defining suitable self-supervised tasks. These tasks are designed so as to capture a-priori known properties of the data at hand, such as temporal consistency in videos. We explore some pre-training approaches in section 4.2.

As we noted in section 3.2, the IVUS videos have substantial jitter. We believe that removing this noise from the videos may be beneficial for subsequent learning. Therefore, we develop methods to reduce motion from the input video. The jitter free video thus generated may be thought of as a domain specific representation of the raw data that is invariant to the irrelevant motion present in the input. Section

4.3 discusses some stabilization methods.

In the following subsections, we first describe procedures to mitigate class imbalance (sections 4.1.1), and set up strong baselines using augmentations, equivariant networks, and other transformations (sections 4.1.2, 4.1.3, 4.1.4). We then discuss two pre-training approaches – one using the cycle consistency of time (section 4.2.1) and another using temporal redundancy (section 4.2.2). Then, we dive deeper into registration (section 4.3) and discuss two approaches to create stabilized representations of the input videos (section 4.3.1, 4.3.2).

Notation: Throughout this chapter we will adopt the following notation. We use upper-case X for the raw input IVUS video or frame. If the input consists of T frames, each of size $H \times W$, then $X \in [0, 1]^{T \times H \times W}$ is a three-dimensional input tensor of intensity values. If the input has only one frame, $T = 1$. For $0 \leq t < T$, we denote by $X_t \equiv X[t] \in [0, 1]^{H \times W}$ the t -th frame of X . We denote the *labels* by $y \in \{(+), (-)\}$, where $(+)$ indicates malapposed input and $(-)$ indicates healthy input. We use f_θ to denote a classification network, parameterized by θ . The classification network outputs *logits*, \hat{x} , $f_\theta(X) = \{\hat{x}_{(-)}, \hat{x}_{(+)}\}$. The *predicted label* is $\hat{y} = (+)$ if $\hat{x}_{(+)} \geq \hat{x}_{(-)}$ and $\hat{y} = (-)$ otherwise. Finally, we introduce indicator functions for the labels $y_{(+)} = \mathbb{1}[y = (+)]$ and similarly for $y_{(-)}$.

4.1 Supervised malapposition detection

Evaluation Metric: As mentioned in chapter 3, the malapposition dataset has high class imbalance $(-)/(+)$ of ≈ 10 . As such, a dummy classifier, $\hat{y}_{dummy} = (-)$, which simply predicts $(-)$ for every input achieves an accuracy of ≈ 0.9 . In such situations, it is customary to look at other metrics of classifier performance such as true positive rate (*TPR* or *sensitivity*), true negative rate (*TNR* or *specificity*), *F1* score, and balanced accuracy (*BA*) [65]. Appendix A defines these metrics. We choose to compare models using the balanced accuracy values. For example, the aforementioned dummy classifier achieves a balanced accuracy of 0.5.

4.1.1 Loss functions

The simplest approach for constructing a supervised classifier is to train a neural network f_θ to minimize the total loss $\sum_i \mathcal{L}^{supervised}(X_i, y_i; \theta)$ over the samples (X_i, y_i) . The most common choice of loss function is the cross entropy loss. For logits x , defining for convenience, the model's estimated probabilities for the two classes:

$$p_{(+)} = \frac{\exp(x_{(+)})}{\exp(x_{(+)}) + \exp(x_{(-)})} \quad , \quad p_{(-)} = \frac{\exp(x_{(-)})}{\exp(x_{(+)}) + \exp(x_{(-)})} \quad (4.1)$$

Given true labels y , the cross entropy loss is given by:

$$\mathcal{L}_{CE}(x, y) = -\log p_{(+)} \cdot y_{(+)} - \log p_{(-)} \cdot y_{(-)} \quad (4.2)$$

Weighted loss

A popular approach in situations of high class imbalance is to use a *weighted loss function*, assigning different weights to the two classes [47]. For logits x and true labels y , the w -weighted loss function is given by:

$$\mathcal{L}_{weighted}(x, y; w) = -w \cdot \log p_{(+)} \cdot y_{(+)} - \log p_{(-)} \cdot y_{(-)} \quad (4.3)$$

Increasing the value of w increases the weight assigned to the (+) class. Plugging $w = 1$ gives back the unweighted binary cross-entropy loss.

Focal loss

Recently, it has been proposed to address the class imbalance problem by using a *Focal Loss* function that is designed to focus the training on a sparse set of hard examples and prevent the vast number of easy negatives from overwhelming the detector during training [36]. Given the true labels y , the focal loss is given by:

$$L_{focal}(x, y; \alpha, \gamma) = -\alpha \cdot p_{(-)}^\gamma \cdot \log p_{(+)} \cdot y_{(+)} - (1 - \alpha) \cdot p_{(+)}^\gamma \cdot \log p_{(-)} \cdot y_{(-)} \quad (4.4)$$

When $\gamma = 0$, focal loss is equivalent to weighted loss with weight $w = \frac{\alpha}{1-\alpha}$. The focusing parameter γ smoothly adjusts the rate at which easy examples are down-weighted. A value of $\gamma > 0$ shifts the focus from easier examples to harder examples.

4.1.2 Encoding invariance in the classifier

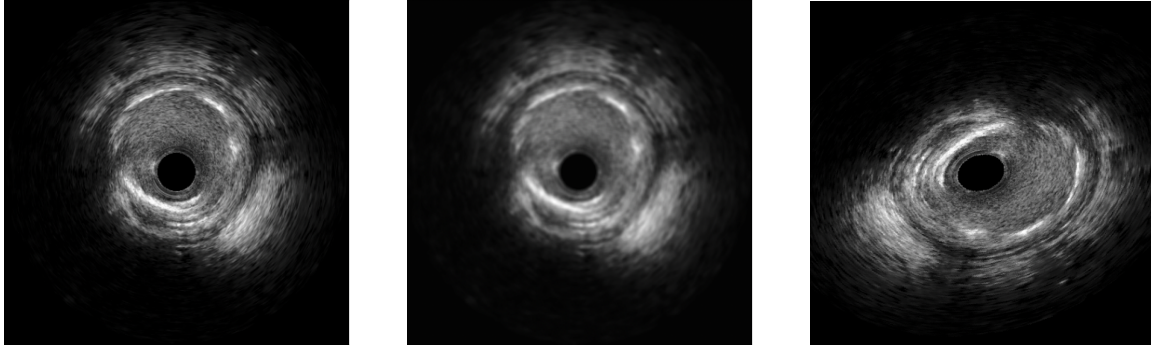
Invariance to certain variations in the data (such as rotations, translations, etc.) can lead to improved generalization and reduced sensitivity to noise. We may encode such invariances either in a data driven way (using data augmentations) or an architecture driven way (using a function class that is invariant to certain transformations). We describe the two approaches here.

Data augmentations

Data augmentations allow models to generalize to unseen examples by capturing more variation in the input data. Because our dataset is relatively small, data augmentations are especially useful in our case. We consider two types of augmentations: intensity and geometrical. Intensity augmentations randomly perturb the image by changing intensity attributes such as brightness, hue, contrast, blur, sharpness, etc. The geometric transformations randomly perturb the image with affine transformations: shear, flip, and rotation. During all these transforms we keep the catheter center unchanged. Figures 4-1b and 4-1c shows a sample of the images resulting from intensity and geometrical transformations of the original image, figure 4-1a.

$E(2)$ -equivariance

The Euclidean group $E(2)$ is the group of isometries of the plane \mathbb{R}^2 , consisting of translations, rotations and reflections. $E(2)$ -equivariant steerable convolution neural networks [64] constrain the latent representation of the inputs to be equivariant under the transformations of the group $E(2)$. This is especially relevant to our images without a strongly preferred global orientation. Indeed, the malapposition may be



(a) Original Image (b) Intensity Transformed (c) Geometry Transformed

Figure 4-1: Intensity and geometric augmentations. Figure 4-1a shows the original image in the dataset. Figure 4-1b shows a blurred and jittery version (intensity augmentation). Figure 4-1c shows a version after flipping and affine transforms (geometric augmentation).

found at arbitrary positions and in arbitrary orientations in the intravascular images. Therefore, just like augmentations that transform the input by elements of the group $E(2)$ - such as rotations, flips, and translations - the equivariance prior is expected to be useful for malapposition detection.

4.1.3 Representation choice: Cartesian vs. polar coordinates

The intravascular ultrasound catheter acquires polar images of blood vessels which are then warped into Cartesian plane with a padding for the catheter (resulting in the black catheter center) for the physicians to view and diagnose from. As such, it is plausible to perform all analysis on the polar images themselves. Moreover, since convolutional neural networks are approximately translation equivariant, models trained on polar representation will lead to representations that are (approximately) equivariant to rotations around the chosen origin. Hence we hope that models trained on polar representations achieve similar desirable properties as achieved by $E(2)$ -equivariant networks and rotation transforms (section 4.1.2). Figure 3-2 shows several Cartesian images and their polar counterparts. Figure 4-2 shows a sample Cartesian image with stent gap annotations and its polar counterpart.

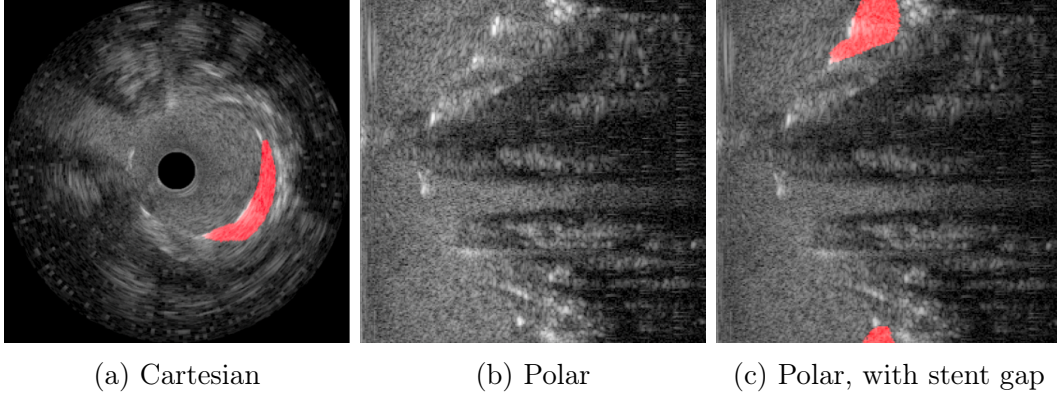


Figure 4-2: A Cartesian image and its polar counterpart, with and without stent gap segmentation. Notice that in 4-2c, the two red parts actually refer to nearby real locations, but a simple CNN doesn't take this into account. Using reflect padding and augmentations partly remedy this issue.

4.1.4 Including temporal information

A frame-wise classification model fails to account for the temporal nature of the dataset. Therefore, using clips of consecutive frames instead of just one frame may increase accuracy by adding temporal information. Indeed, having multiple consecutive frames makes it easier to spot blood flow around the stent, and therefore malposition of the stent. A higher clip length may also help reduce the effect of noise in frames. On the flip side, a higher clip length leads to increased memory and computation, longer training times, and owing to the small size of our dataset, possible over-fitting.

We study the effect of varying clip length on the performance of a model f_θ that employs 2D-convolutions by treating the time axis as separate channels. For an input $X \in [0, 1]^{T \times H \times W}$, the classifier aims to predict the malapposition annotation (y) of the central frame $X[[T/2]]$.

4.1.5 Optimization choices

Learning rate schedules can significantly impact the model's accuracy and rate of convergence. Since there is no one-size-fits-all approach to scheduling, we experiment

with different popular methods such as step-decay, cosine-annealing [37] and one-cycle [50]. Figure 4-3 plots these schedules.

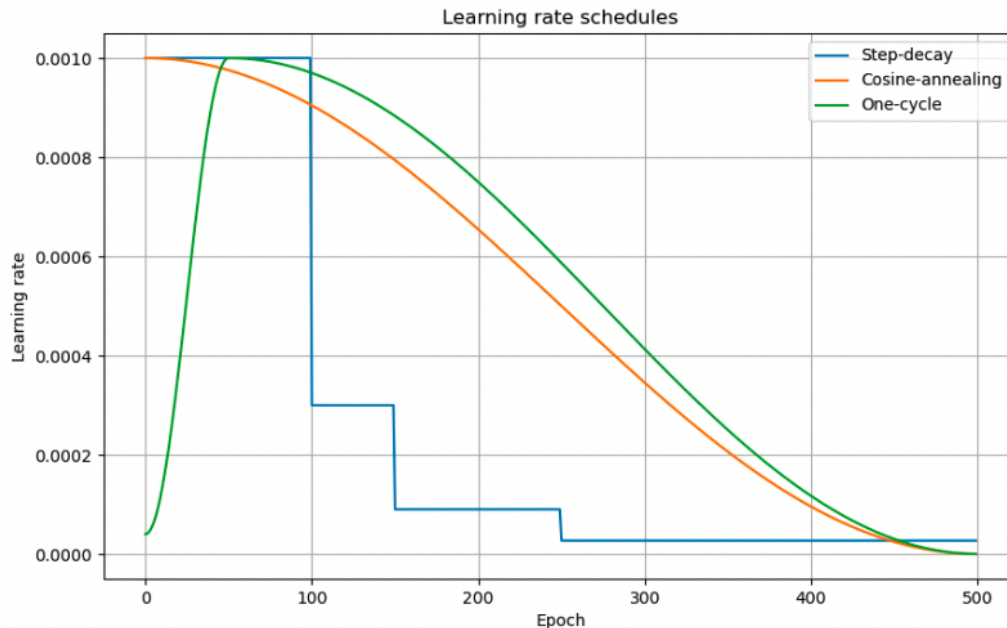


Figure 4-3: Evolution of learning rate with the step-decay, cosine-annealing, and one-cycle schedules.

4.2 Unsupervised representation learning from unlabelled pullbacks

Numerous methods for unsupervised learning on video have been proposed which use classical temporal methods [46, 53], cycle-consistency of time [26, 34, 60, 63] or utilize the high auto-correlation by masked modelling [23, 61, 56, 57]. Here, we choose two methods: contrastive random walks (section 4.2.1) and VideoMAE (section 4.2.2); and study their representation learning performance on the IVUS pullbacks. In particular, we use unsupervised methods to learn a useful intermediate representation from the unlabelled IVUS pullbacks, and then fine-tune this representation in tandem with a smaller model for the downstream task of malapposition classification using the smaller annotated dataset.

Notation: During *pre-training*, we learn an encoder $h_\phi : [0, 1]^{H \times W} \rightarrow \mathbb{R}^d$ that takes the raw input to an intermediate representation. We then use the labelled data to learn a simpler classifier $g_\psi : \mathbb{R}^d \rightarrow \mathbb{R}^2$ that outputs logits for classification. Thus, in accordance with our previous notation (section 4.1), the classifier network is $f_\theta = g_\psi \circ h_\phi$ with parameters $\theta = \{\phi, \psi\}$.

4.2.1 Contrastive Random Walks on Video

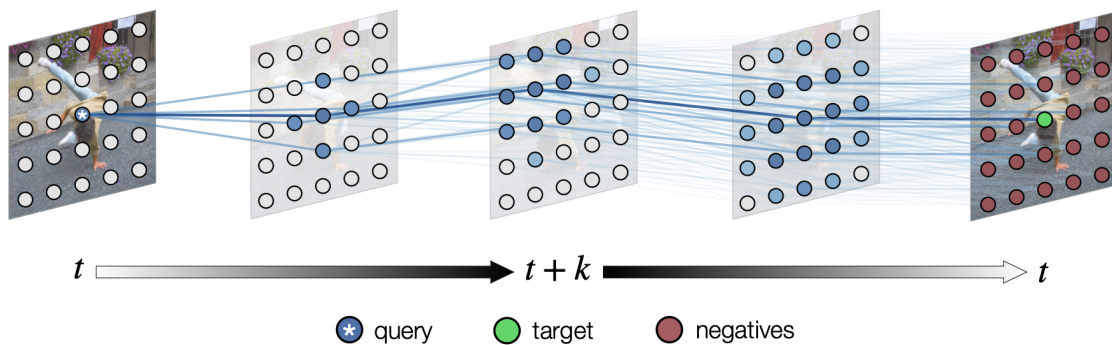


Figure 4-4: Contrastive Random Walks on Video. Every frame is divided into patches and the video is represented as a graph where the nodes are patches and edges are patch affinities. A contrastive loss encourages the paths that reach a target, implicitly supervising latent correspondence along the path. Learning proceeds without labels on palindrome frame sequences. Taken from [26].

To begin our exploration, we use contrastive random walks [26], which is the state-of-the-art self-supervised method on video object segmentation, pose keypoint propagation, and semantic part propagation.

Overview: The task of correspondence is cast as prediction of links in a space-time graph constructed from video. In this graph, the nodes are patches sampled from each frame, and nodes adjacent in time can share a directed edge. They learn a representation in which pairwise similarity defines transition probability of a random walk, so that long-range correspondence is computed as a walk along the graph (See figure 4-4). The representation is optimized to place high probability along paths of similarity. Targets for learning are formed without supervision, by cycle-consistency: the objective is to maximize the likelihood of returning to the initial node when

walking along a graph constructed from a palindrome of frames. Thus, a single path-level constraint implicitly supervises chains of intermediate comparisons.

Formulation: Each video X is represented as a directed graph where the nodes are patches, and weighted edges connect nodes in the neighboring frames. Let \mathbf{q}_t be the set of N nodes extracted from the frame $X[t]$ by sampling patches in a grid. We learn an encoder h_ϕ that maps patches to l_2 -normalized d -dimensional vectors, with similarity $d_\phi(q^i, q^j) = \langle h_\phi(q^i), h_\phi(q^j) \rangle$. A softmax (with temperature τ) converts patch similarities to the stochastic affinity matrix between timesteps t and $t + 1$:

$$A_t^{t+1}(i, j) = \frac{\exp(d_\phi(q_t^i, q_{t+1}^j)/\tau)}{\sum_{l=1}^N \exp(d_\phi(q_t^i, q_{t+1}^l)/\tau)} \quad (4.5)$$

The k -range ($k \leq T$) correspondence matrix can then be formulated as:

$$\bar{A}_0^k = \prod_{t=1}^{k-1} A_{t-1}^t \quad (4.6)$$

and the matrix for the traversal back \bar{A}_k^0 can be constructed using the reversed video with frames $X[k-1], X[k-2], \dots, X[0]$. Since each patch should correspond to itself after traversing the cycle of length $2k$, the cycle-consistency objective is:

$$\mathcal{L}_{cyc}^k = \mathcal{L}_{CE}(\bar{A}_0^k \cdot \bar{A}_k^0, I_{N \times N}) \quad (4.7)$$

The pre-training loss is sum over cycles of all lengths: $\mathcal{L}_{crw} = \sum_{k=1}^T \mathcal{L}_{cyc}^k$.

Fine-tuning: Once an encoder, h_ϕ , is pre-trained for the contrastive random walk correspondence task on our intravascular ultrasound clips using appropriate value of the parameters like maximum half-cycle length T , softmax temperature τ , number of patches N , and edge dropout, the resulting encodings are tested on the downstream classification task of malapposition detection. To this end, we use a linear head, g_ψ , for classification and fine-tune the trained encoder parameters (ϕ) in tandem with those of the linear head (ψ).

4.2.2 VideoMAE - Masked Autoencoders

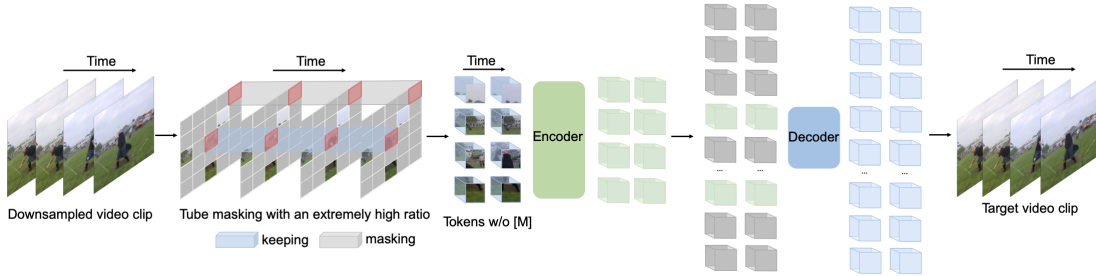


Figure 4-5: VideoMAE architecture. Input videos are divided into patches and patches are randomly masked with a high masking ratio (80-95%). The mask is kept same for a tubelet of frames, and a vision transformer model is trained using an in-painting task with a mean squared error loss. Taken from [57].

Overview: VideoMAE [57] is state-of-art self-supervised method on the relatively small-scale video datasets such as Something-Something [22], UCF101 [52], and HMDB51 [32]. Self supervised pre-training is performed by training a transformer model on the task of in-painting video clips with tube masking with an extremely high ratio. This simple design makes video reconstruction a more challenging and meaningful self-supervision task, thus encouraging extracting more effective video representations during the pre-training process.

VideoMAE has been shown to achieve impressive results on down-stream tasks even with pre-training on very small datasets. This property makes it a good fit for our downstream task of malapposition detection with just a limited number of samples.

Formulation: The input video $X \in [0, 1]^{T \times H \times W}$ is first divided into non-overlapping cubes of size $t \times h \times w$, and each cube is represented with a token embedding, resulting in $\frac{T}{t} \times \frac{H}{h} \times \frac{W}{w}$ 3D tokens. Then a fraction ρ_{mask} of the cubes are masked by using *tube masking*. Mathematically, the tube masking mechanism can be expressed as:

$$\mathbb{I}[p_{x,y,..} \in \Omega] \sim \text{Bernoulli}(\rho_{mask}) \quad (4.8)$$

where Ω is the set of masked tokens. Notice that different times t share the same

mask. The unmasked tokens are fed into the transformer encoder h_ϕ . Finally a shallow decoder \tilde{g} is placed on top of the visible tokens from the encoder and the learnable mask tokens to reconstruct the video. The loss function is the mean squared error between the normalized masked tokens and the re-constructed ones in the pixel space:

$$\mathcal{L}_{VideoMAE} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \|X(p) - \hat{X}(p)\|_2^2 \quad (4.9)$$

Fine-tuning: Once pre-trained on the VideoMAE in-painting objective, the shallow decoder \tilde{g} is discarded and replaced with a linear classification head g_ψ , which is then fine-tuned along with the encoder h_ϕ using the IVUS pullbacks which are labelled for malapposition.

4.3 Jitter removal from pullbacks

As discussed in chapter 3, the intravascular ultrasound videos suffer from jitters caused by the cardiac cycle as well as the motion of the catheter within the blood vessels. A video-stabilization approach that minimizes such motion may therefore make it easier for a human to diagnose blood vessel diseases effectively and for a temporal model to perform better. Indeed, stabilized videos reduce spurious noise and enable convolutions to better capture temporal variations at corresponding spatial locations [39, 69]. We find that basic non-medical approaches to digital video stabilization such as point-feature matching [3] fail on our data.

In this section, we seek to achieve video-stabilization by leveraging methods from the well-studied field of medical image *registration*. Registration is the process of aligning two or more images of the same scene or object, taken from different view-points, times, or sensors. The goal of image registration is to find a transformation that maps the pixels of one image to the corresponding pixels in the other image, such that the images are spatially aligned. Numerous approaches for image registration have been proposed [59, 6, 17, 68]. We study a classical approach, using Advanced

Normalization Tools (ANTs) (section 4.3.1) and a deep learning registration approach (section 4.3.2). While the motion of the catheter, blood, and the cardiac cycle in general cause the blood vessels to undergo non-rigid transformations, we visually notice that rigid transformations (translation, rotation) are a major contributor to motion noise in the IVUS videos. Therefore, we restrict our attention to rigid transforms for simplicity. However, both of these methods can be extended to arbitrary choice of transforms in a straightforward manner.

Notation: In addition to the notation introduced at the beginning of the chapter, we introduce the following simplifying notation. We denote by $X \in [0, 1]^{T \times H \times W}$ the video input to the stabilization algorithm. $\hat{X} \in [0, 1]^{T \times H \times W}$ represents the output (stabilized) video. For a video X , we denote by $\bar{X} \in [0, 1]^{H \times W}$ the temporal pixel-wise mean:

$$\bar{X}_{i,j} = \frac{1}{T} \cdot \sum_{t=0}^{T-1} X_{t,i,j} \quad (4.10)$$

Further, for $A \in \mathbb{R}^{T \times H \times W}$ and $B \in \mathbb{R}^{H \times W}$, we overload the subtraction operation by implicit broadcasting, defining $(A - B) \in \mathbb{R}^{T \times H \times W}$ by:

$$(A - B)_{t,i,j} := A_{t,i,j} - B_{i,j} \quad (4.11)$$

Finally, we define the squared l_2 norm of a tensor $X \in \mathbb{R}^{T \times H \times W}$ as:

$$\|X\|_2^2 = \sum_{t=0}^{T-1} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} X_{t,i,j}^2 \quad (4.12)$$

4.3.1 Registration using ANTs

Basic ANTs-based stabilization: Advanced Normalization Tools (ANTs) Registration [59] implements a pairwise gradient-based multi-scale registration procedure with a vast choice of transforms. A straightforward extension of the ANTs registration method from two images to a clip of T images is to construct a template from the clip (for instance by taking the pixel-wise mean), and registering each frame to

the template. Algorithm 1 illustrates this basic stabilization approach.

Algorithm 1 Basic clip-stabilization using ANTs

Input

X Video array $\triangleright X : T \times H \times W$

Output

\hat{X} Stabilized video

$\hat{X} \leftarrow X$

`template` \leftarrow `ContrsuctTemplate`(X)

for $i = 0$ to T **do**

$\hat{X}[i] \leftarrow$ `ANTsRegistration`(`template`, $X[i]$)

end for

return \hat{X}

However, clip length T highly influences the performance of this basic method. Indeed, the intravascular ultrasound videos consist of thousands of frames encompassing multiple distinct anatomies such as lumbar spine bones, bowel and psoas muscles, lymph nodes, venous valves, ilio caval confluence and vascular branches. As such, the approximation of a single evolving structure under which registration methods work fails on temporally far-apart frames. Therefore, while algorithm 1 is suitable for short clips, stabilization of longer clips and full videos requires additional refinements.

Iterative ANTs-based stabilization: In order to stabilize videos with a large number of frames, we employ an iterative moving average approach calling ANTs as a subroutine for pairwise registration. In the iterative approach, each frame of the input is registered to a template constructed from the frames within its temporal neighborhood. In order to approximately maintain the cardiac phase through different templates, we use a window size of a whole integer multiple of the cardiac cycle duration (chapter 3). The process is repeated until the resulting transformations become smaller than a threshold. Algorithm 2 illustrates the iterative moving average approach in pseudo-code.

Evaluating registration quality: The registration quality may be quantitatively evaluated by the amount of mean squared error it can explain away. For input and output clips $X, \hat{X} \in [0, 1]^{T \times H \times W}$, using the notation introduced at the beginning

Algorithm 2 Iterative video-stabilization using ANTs

Input

X Video array $\triangleright X : T \times H \times W$
 w Template window half-size

Output

\hat{X} Stabilized video
 $\hat{X} \leftarrow X$

while stopping condition not achieved **do**

for $i = 0$ to T **do**

 template \leftarrow ContrsuctTemplate($X[i - w : i + w]$)

$\hat{X}[i] \leftarrow$ ANTsRegistration(template, $X[i]$)

end for

end while

return \hat{X}

of this section, the MSE-reduction per pixel is given by:

$$\Delta\text{MSE}(X, \hat{X}) = \frac{1}{T \cdot H \cdot W} \cdot (\|X - \bar{X}\|_2^2 - \|\hat{X} - \bar{\hat{X}}\|_2^2) \quad (4.13)$$

A higher value of ΔMSE indicates a better stabilization.

4.3.2 Registration using deep-learning

Taking inspiration from recent deep learning approaches to medical image registration [24, 42], we implement a CNN-based architecture for video stabilization (figure 4-6). One benefit of using a trained neural-net based model over classical optimization approaches is the ability to do real-time registration.

Architecture: We use an encoder f_θ with two fully connected heads to predict the rotation ($\phi \in (-\pi, \pi]^T$) and translation ($v \in (-1, 1)^{T \times 2}$) parameters:

$$f_\theta(X) = \{\phi, v\} \quad (4.14)$$

The parameters $\{\phi, v\}$ are in-turn used to transform the input video $X \in [0, 1]^{T \times H \times W}$ to the stabilized output $\hat{X} \in [0, 1]^{T \times H \times W}$. In particular, the t -th frame of the input, X_t , is rotated about its center by $\phi[t]$, and translated by $\{v_x, v_y\} = v[t]$. The transfor-

mation parameter $v_x \in (-1, 1)$ translates the image by horizontally by $v_x \cdot W$ pixels, and the transformation parameter $v_y \in (-1, 1)$ translates the image vertically by $v_y \cdot H$ pixels. A tanh activation layer keeps the parameters within the required range. We use another network, $g_\psi : [0, 1]^{T \times H \times W} \rightarrow [-1, 1]^{H \times W}$, to construct a template frame $\hat{X} \in \mathbb{R}^{H \times W}$:

$$\hat{X} = \bar{X} + g_\psi(X) \quad (4.15)$$

by adding its output to the input mean \bar{X} for faster convergence.

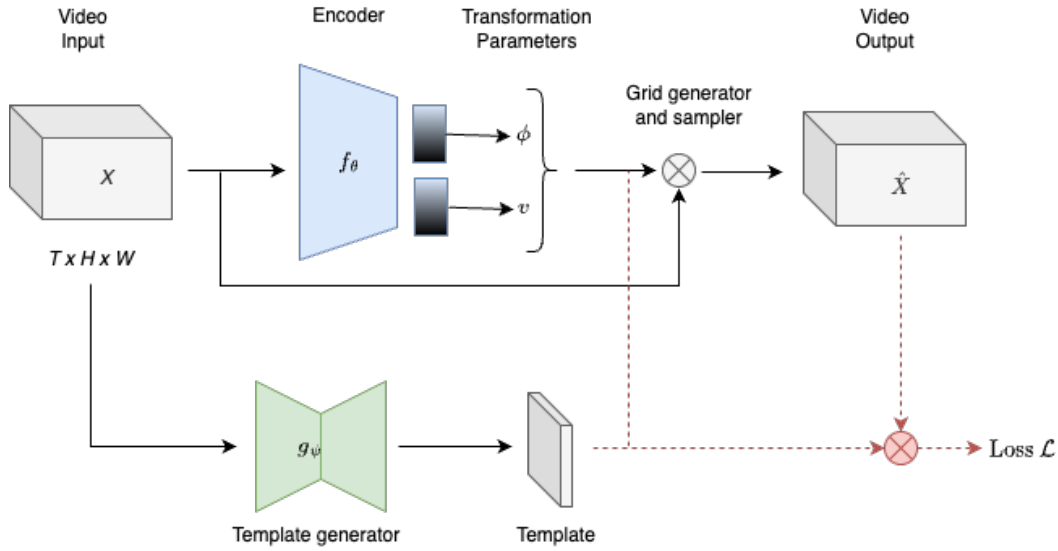


Figure 4-6: Proposed registration network architecture. Input video is passed through an encoder f_θ with two linear heads for rotation and translation parameters respectively. The video is then transformed by sampling from a transformation grid generated using the predicted parameters. Another network, g_ψ , independently predicts the template, which is used to evaluate the model performance in the loss function. Additionally, we apply a regularization penalty on the predicted transformation parameters.

Loss: The loss consists of a registration term and a regularization penalty:

$$\mathcal{L}(\hat{X}, \hat{\hat{X}}, \phi, v) = \mathcal{L}_{register}(\hat{X}, \hat{\hat{X}}) + \lambda \cdot \mathcal{L}_{penalty}(\phi, v) \quad (4.16)$$

Any metric such as MSE or cross-correlation may be employed for the registration

loss. For instance, the MSE registration loss is:

$$\mathcal{L}_{register}(\hat{X}, \hat{\hat{X}}) = \|\hat{X} - \hat{\hat{X}}\|_2^2 \quad (4.17)$$

We use an L_2 penalty:

$$\mathcal{L}_{penalty}(\phi, v) = \|\phi\|_2^2 + \|v\|_2^2 \quad (4.18)$$

We study the ablations on the choice of hyperparameter λ , registration loss metric $\mathcal{L}_{register}$, and the architecture of the template generator g_ψ ; and compare the performance of the registration network with ANTs-based registration approach 4.3.1.

Chapter 5

Experiments and Results

Here we describe the results of our methods (chapter 4) applied to the intravascular ultrasound data (chapter 3). Unless stated otherwise, all the recorded values are an average over 4 folds of cross-validation. We profile the folds in appendix B. Table B.1 shows the pullback IDs that were included in the folds and table B.2 shows the number of malapposed (label (+)), well-apposed (label (-)), and discarded frames in each of the folds.

5.1 Supervised malapposition detection

We report the results of the classification methods described in chapter 4. Unless otherwise stated, we use a ResNet-18 with a linear head of depth 2 as the classifier f_θ (refer to section 4.1 where this notation was introduced).

5.1.1 Loss functions

Weighted loss

Table 5.1 compares the accuracy (ACC), true positive rate (TPR or *sensitivity*), true negative rate (TNR or *specificity*), $F1$ score, and balanced accuracy (BA) for

two models, with and without using weighted loss. For instance, the first row of 5.1 shows that a baseline ResNet-18 model achieves an accuracy of 0.87, but most of it is attributed to its high true-negative rate. The second row of table 5.1 shows the classification metrics of a model with the same architecture trained using a 10-weighted loss ($w = 10$, section 4.1.1). Notice that the model achieves a much better TPR, and balanced accuracy (BA) at the cost of a reduction in accuracy and TNR. In what follows, we compare the models using the F1 score and balanced accuracy. For simplicity, we only report BA values moving forward. Unless stipulated otherwise, the loss is 1-weighted ($w = 1$).

	ACC	TPR	TNR	F1	BA
Baseline ($w = 1$)	0.87	0.23	0.93	0.24	0.58
Weighted ($w = 10$)	0.78	0.40	0.81	0.26	0.61

Table 5.1: The effect of using a weighted loss functions on various classification metrics. Notice that weighting the positive samples leads to higher TPR and BA, by increasing true and false positives and reducing false negatives.

Focal loss

We found the focusing parameter $\gamma = 3$ to work best for our experiments. In table 5.2, we set $\gamma = 3$. and experiment with three values of α . Note that $\alpha = 0.5$ corresponds to focusing with equal weights to the two classes. Notice how setting $\gamma = 3$ improves the balanced accuracy from 0.68 (table 5.3) to 0.72.

	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$
Balanced Accuracy	0.67	0.72	0.70

Table 5.2: The effect of using Focal Loss (section 4.1.1) on the task of malapposition detection. All the experiments were run with intensity and geometric augmentations (section 4.1.2).

5.1.2 Encoding invariance in the classifier

Data augmentations

Table 5.3 shows the effect of adding intensity and geometric augmentations to our baseline ResNet-18 classifier. We notice that augmentations indeed help the model generalize better. In the following sections, we augment all experiments with intensity and geometric transforms.

Augmentation	Balanced Accuracy
None	0.58
Intensity	0.63
Intensity + Geometric	0.68

Table 5.3: The effect of data augmentations on classification accuracy. The numbers are average over four folds. Notice that both intensity and geometric transformations are useful for our task.

$E(2)$ -Equivariance

Table 5.4 shows the balanced accuracy obtained by using an $E(2)$ -equivariant network for the classification task. We notice that the equivariant network performs much better than baseline. However, once data augmentations are introduced, both the models have similar performance. Indeed, the geometric data augmentations used include continuous rotations and reflections (but not translations).

	Baseline	$E(2)$-Equivariant
No Augmentations	0.58	0.64
Augmentations	0.71	0.70

Table 5.4: The effect of using a $E(2)$ -equivariant steerable convolutional neural network for the malapposition classification task.

5.1.3 Representation choice: Cartesian vs. polar coordinates

We observe that transforming the Cartesian images into polar coordinates (section 4.1.3) usually leads to better performance compared to the Cartesian equivalents. Table 5.5 depicts this phenomenon.

	Cartesian	Polar
No Augmentations	0.58	0.63
Augmentations	0.71	0.72

Table 5.5: The effect of using polar coordinates for the malapposition classification task. Notice the jump in accuracy, and the similarity with table 5.4.

5.1.4 Including temporal information

Table 5.6 shows the effect of using clips of different lengths as input to the different channels of the ResNet encoder. We notice that using more than one frame helps only slightly. Furthermore, using very long clips decreases the performance. The maximum is achieved at $T_{clip} = 4$.

	$T_{clip} = 1$	$T_{clip} = 4$	$T_{clip} = 8$	$T_{clip} = 12$
Balanced Accuracy	0.737	0.742	0.738	0.710

Table 5.6: The effect of using clips of different lengths T_{clip} on the malapposition classification task. All the experiments are performed with One-Cycle cosine annealing learning rate schedule, and intensity and geometric augmentations.

5.1.5 Optimization choices

Table 5.7 shows the balanced accuracies obtained by using different learning rate schedules. We find that one-cycle cosine annealing achieves the best accuracy.

Learning Rate Schedule	Balanced Accuracy
No Decay	0.68
2-Step	0.70
4-Step	0.70
Cos-Anneal	0.73
One-Cycle	0.74

Table 5.7: The effect of using different popular learning rate schedulers for the malaposition classification task.

5.2 Unsupervised representation learning from unlabelled pullbacks

5.2.1 Contrastive Random Walks on Video

Performance on pre-training task: One way to measure performance on the pre-training task of finding correspondences in a temporal cycle is to measure the fraction of patches that are correctly mapped to themselves after the random walk. Borrowing the notation of section 4.2.1, for a cycle of half-length k , we call a patch $0 \leq j < N$ to be correctly mapped if the j -th row of the correspondence matrix assigns maximum probability to patch j itself, i.e., if

$$j = \arg \max_{0 \leq i < N} (\bar{A}_0^k \cdot \bar{A}_k^0)_{i,j} \quad (5.1)$$

Figure 5-1 plots the fraction of correctly mapped patches over different cycle-lengths, averaged over all out-of-sample clips. As the figure demonstrates, the model performs impressively well on the pre-training task of temporal consistency, although the performance declines as the task gets much harder (i.e., for longer cycles).

Label propagation using pre-trained encoder h_ϕ : Another way to evaluate temporal correspondence learnt by a model trained on the contrastive random walk objective is through using the pre-trained model for the closely related task of label propagation [26]. In label propagation, the input is a pair (X, L_0) of a clip X and

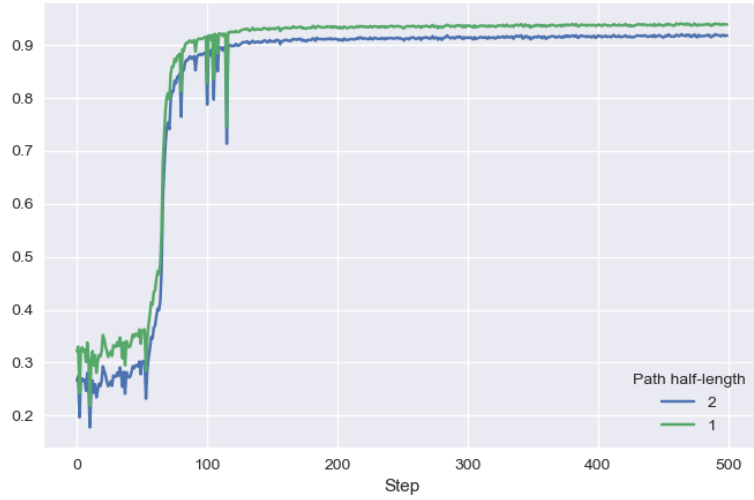
dense segmentation labels L_0 for the first frame X_0 of the clip. The encoder h_ϕ trained on the contrastive random walk task may subsequently be used to predict the labels $\hat{L}_1, \dots, \hat{L}_{T-1}$ for the rest of the frames X_1, \dots, X_{T-1} (we refer the reader to [26] for details). We use our small dataset of dense stent gap annotations (chapter 3) to generate segmentation labels \hat{L} for the subsequent frames. Since we lack ground truth annotations of the subsequent segmentations, we may only evaluate the predicted labels visually. Figure 5-2 shows the results of label propagation on one such clip. We notice that while the model can roughly localize the region of interest in the subsequent frames, it is unable to accurately propagate the segmentation.

Performance on downstream task: Table 5.8 shows the effect of using a pre-trained model for the classification task, under one setting of pretraining parameters. We notice that pre-training improves the classification accuracy when no augmentations are used. But once we add data augmentations (section 4.1.2), pre-training doesn't help anymore. Numerous settings of the clip-lengths, loss temperatures, and augmentations were tried during the pre-training, but all of them lead to similar results.

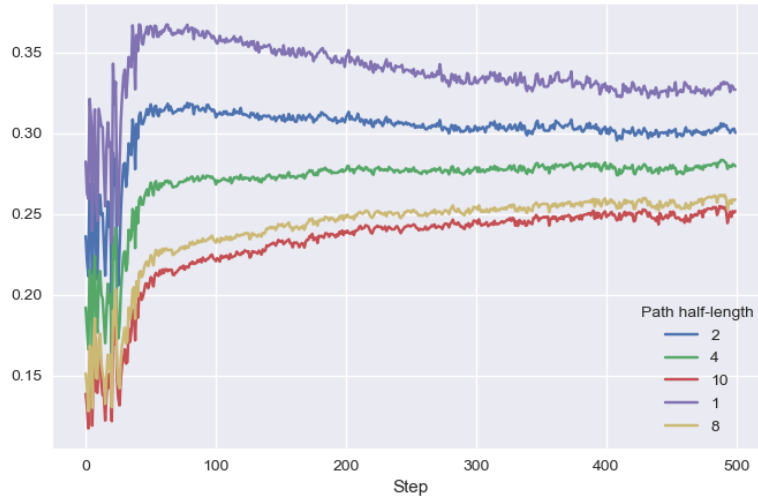
	Baseline	Pretrained
No Augmentations	0.58	0.61
Augmentations	0.71	0.65

Table 5.8: The effect of pre-training using contrastive random walks on videos on the classification accuracy. A ResNet-18 encoder was trained using the random walk objective [26], and a depth-2 MLP was added for the downstream classification fine-tuning. The pre-training was done on cycle-lengths ≤ 4 , and the patches and frames were randomly cropped and resized. The accuracy values are average over 4 folds. Notice that pre-training doesn't help once we add data augmentation.

Thus, while we can achieve a high accuracy on the self-supervised task of matching patches, this fails to translate to the downstream classification task. We conclude that contrastive random walk pre-training is not well-suited to the classification task at hand.



(a) Max cycle-length = 4



(b) Max cycle-length = 20

Figure 5-1: Out-of-sample performance on the pre-training task of finding correspondences in a cycle. We plot the average fraction of correctly mapped patches with training step across cycles of different lengths. The model in figure 5-1a was trained on clips of three frames, thus finding correspondences on cycles of length 2 and 4. On the other hand, the model in figure 5-1b was trained on the task of finding correspondences on cycles of length up to 20. For both the models, the input frames were divided into $N = 49$ overlapping patches using a 7×7 grid. Notice that the second task is much harder than the first, as also demonstrated by the low probabilities of correct return in figure 5-1b.

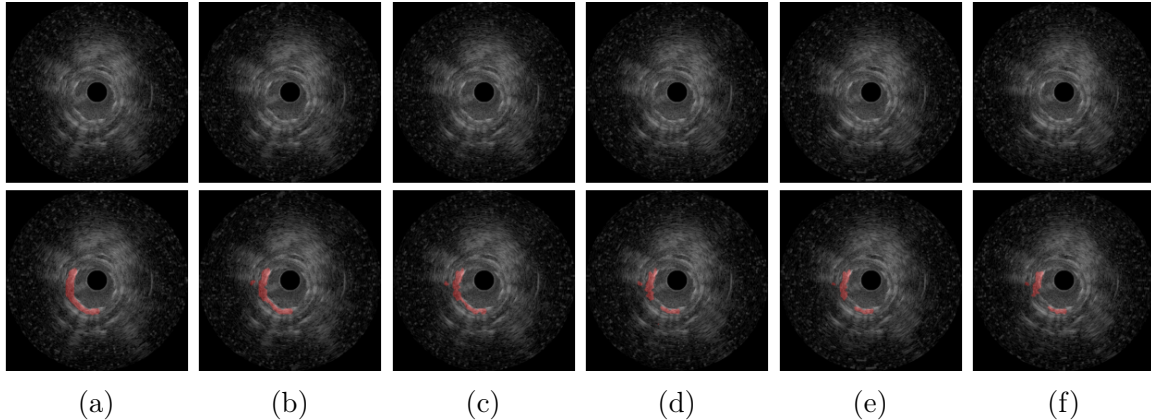


Figure 5-2: Frames X_i of an input clip (top) and their corresponding propagated labels \hat{L}_i (bottom). Here, column (a) is the annotated first frame of the clip (X_0, L_0). Columns (b)-(f) are the succeeding frames, and their propagated labels X_i, \hat{L}_i .

5.2.2 VideoMAE - Masked Autoencoders

Performance on pre-training task: We pre-train a VideoMAE model from scratch on the IVUS dataset. Figure 5-3 shows the results of this pre-training. We observe that the model does a great job in predicting the masked patches, and can reconstruct real imaging artifacts such as shadowing, reverberation etc.

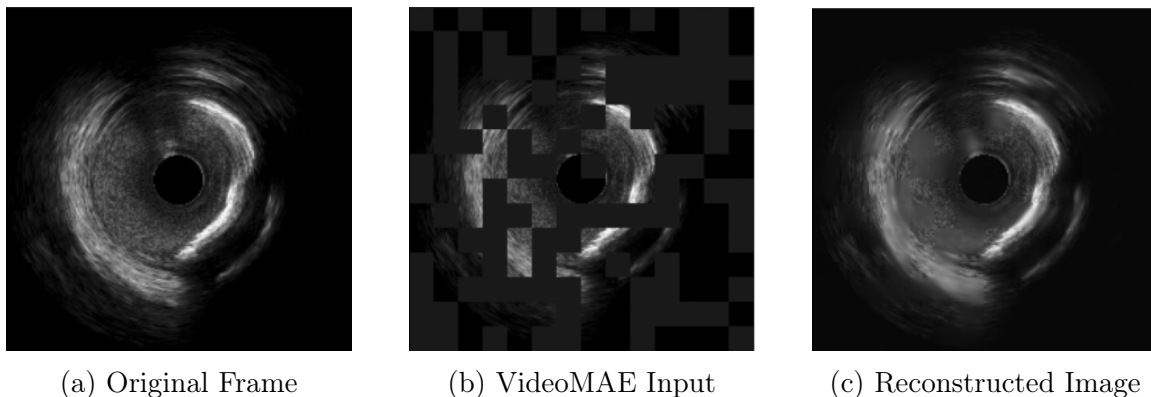


Figure 5-3: Pre-training using VideoMAE. The figure shows a sample output from the VideoMAE pre-training in-painting task applied on the IVUS images. Figure 5-3a shows the original frame, which is masked (figure 5-3b) and fed as an input to the model. Figure 5-3c shows the output of a trained model on the given input. Notice the checkerboard artifacts in the output. This particular model was trained over clips of length $T = 12$ with a masking ratio $\rho_{mask} = 50\%$.

Performance on downstream task: Table 5.9 shows the results of a trans-

former model pre-trained with the VideoMAE objective on the malapposition classification task. Once again, we observe that although pre-training improves classification accuracy in the absence of augmentations, it is no longer beneficial once augmentations are introduced.

	Baseline	Pretrained
No Augmentations	0.58	0.60
Augmentations	0.71	0.65

Table 5.9: The effect of pre-training using VideoMAE on the balanced accuracy for the malapposition task. The transformer encoder was pretrained using clips of length $T = 12$, with a masking ratio of $\rho_{mask} = 95\%$ using tubelets of length $t = 2$. During pre-training, clips were augmented with intensity and geometrical transforms.

5.3 Jitter removal from pullbacks

We evaluate our registration approaches and compare the performance of ANTs-based (section 4.3.1) and deep-learning based (section 4.3.2) methods.

5.3.1 Registration using ANTs

To begin our exploration, we run algorithm 1 on short clips consisting of $T_{clip} = 30$ frames. In order to get a strong baseline, we run a multi-scale ANTs registration with smoothing for minimizing mean squares distance between the frames and template (see listing 5.1).

```
antsRegistration -d 2 \
    -m MeanSquares[template.jpg , frame.jpg , 1] \
    -t Rigid[0.1] \
    -c 100x100x100 \
    -s 32x16x1 \
    --float \
    -f 4x2x1
```

Listing 5.1: ANTs registration command

Table 5.10 compares the results of algorithm 1 with and without using multi-scale optimization over all the training dataset.

	ΔMSE
Single-scale	6.9×10^{-4}
Multi-scale	7.8×10^{-4}

Table 5.10: Results of clip stabilization using ANTs (algorithm 1). Notice the benefit of using multi-scale registration (listing 5.1).

Next, we run algorithm 2 on a longer video of $T = 600$ frames, using a window half-size of $w = 24$ frames, which is approximately the duration of one cardiac cycle (chapter 3). Figure 5-4 shows the lateral view of the input (figure 5-4a) and output (figure 5-4b) videos. Notice the reduction in the saw-teeth between the input and the output videos.

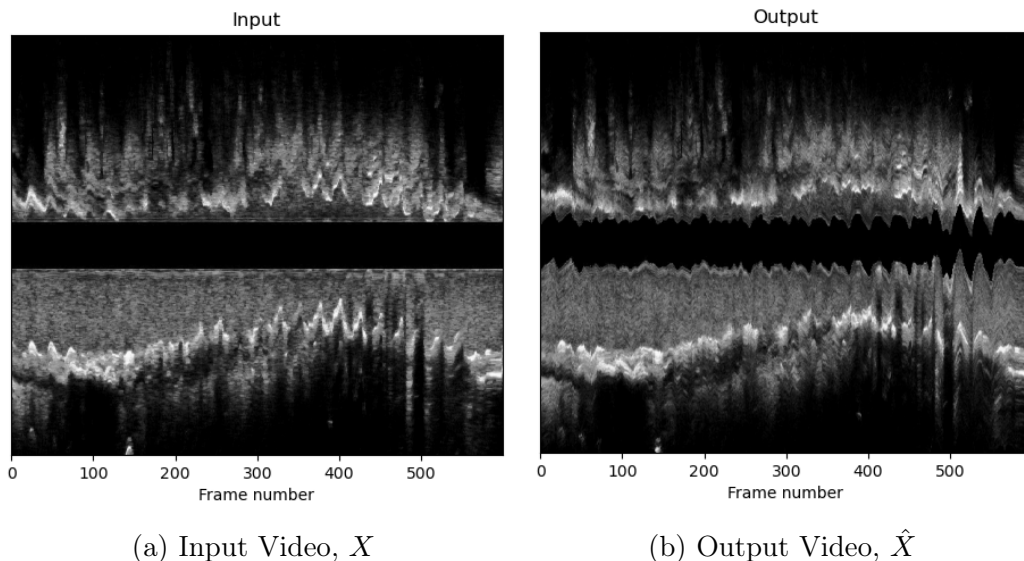
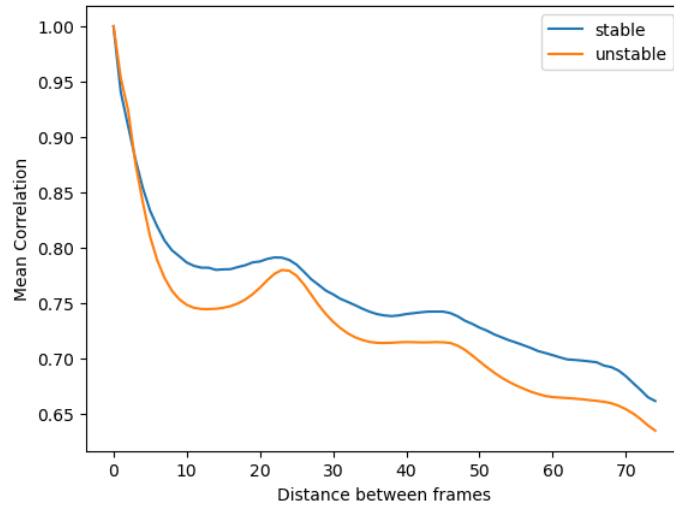


Figure 5-4: Results of applying procedure in algorithm 2 to an intravascular ultrasound video of length $T = 600$. Notice how the algorithm is able to reduce the saw-tooth pattern in figure 5-4a to a much smoother vessel boundary in figure 5-4b, thus de-noising the jitters caused by the cardiac cycle.

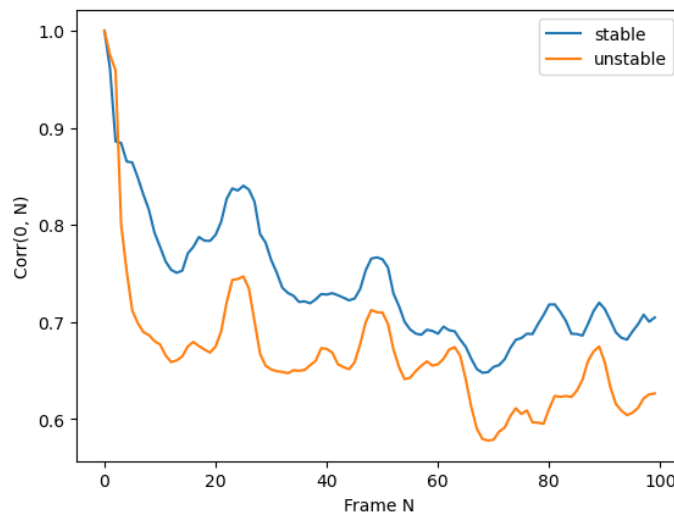
We may further quantitatively look at the motion reduction by looking at the increase in correlation between frames in the input (unstable) and the output (stable)

videos (chapter 3). The mean correlation at a distance d is defined by:

$$\text{MeanCorrelation}(d; X) = \frac{1}{T-d} \sum_{i=0}^{T-1-d} \text{Corr}(X[i], X[i+d])$$



(a)



(b)

Figure 5-5: Results of applying procedure in algorithm 2 to an intravascular ultrasound video of length $T = 600$. Notice how the algorithm is able to significantly increase the correlation between the frames of the video. Notice also that the procedure is able to significantly reduce but not eliminate the periodic peaks caused by the cardiac cycle.

As shown in figure 5-5, algorithm 2 is able to significantly increase pairwise correlation and reduce jitters.

5.3.2 Registration using deep-learning

In this section, we evaluate the proposed registration network (section 4.3.2). For the encoder f_θ , we use a ResNet-18 architecture with two linear heads, each of depth 2. For the template generating network g_ψ , we use a U-Net architecture with 3 channels.

Experiments using MNIST: To begin, we test the network on a simpler alignment task of aligning ‘ones’ in the MNIST dataset [13]. The input $X \in [0, 1]^{30 \times 28 \times 28}$ consists of a stack of 30 randomly chosen pictures of the digit ‘1’ from the MNIST dataset. Figure 5-6 shows the results. We observe that the model achieves good alignment results for MNIST images. Specifically, we find that the model performs equally well across various loss functions (including local cross correlation and MSE), initialization methods, and even when the U-Net is disabled and replaced by the input or output sample mean as the reference template.

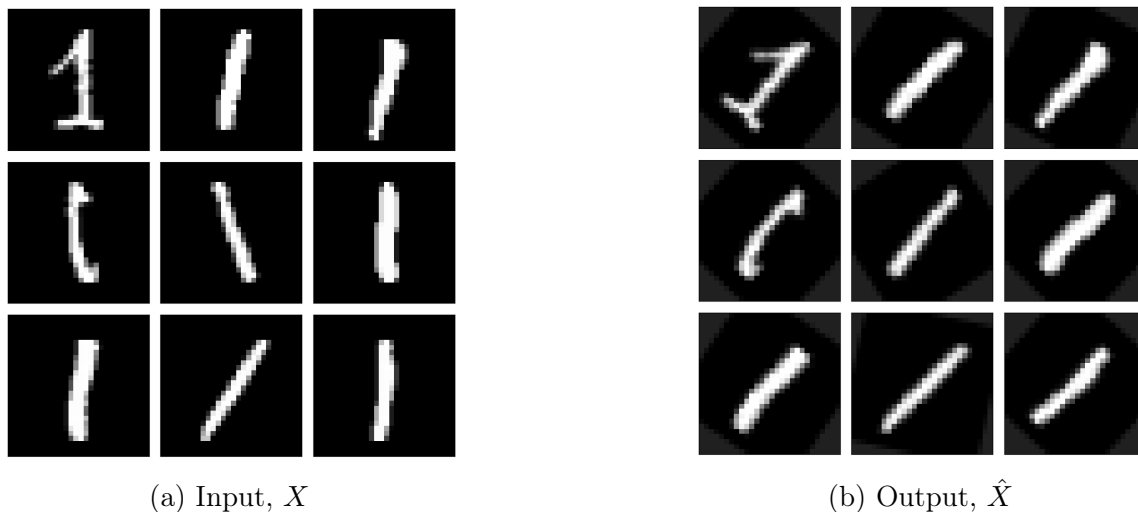


Figure 5-6: Results of using registration network for the task of aligning digits. A ResNet-18 encoder was used with a rotation head of depth 2 and a U-Net with 3 channels. The loss consisted of the MSE registration loss with respect to the U-Net output template and an l_2 penalty loss with $\lambda = 0.1$. The clip length was chosen to be $T = 30$.

Stabilization of only one IVUS clip: Next, we check our model on the IVUS dataset by fitting it on a single IVUS clip input $X \in [0, 1]^{30 \times 256 \times 256}$ and comparing the results with that achieved by ANTs (algorithm 1). Table 5.11 shows the results. We notice that the Registration network performs comparably to the multi-scale ANTs approach on a single input.

Method	Zero-Initialization	MSE(\hat{X})	Visual Quality
None	n/a	2.2×10^{-3}	Very Jittery
ANTs : single-scale	n/a	1.6×10^{-3}	Jittery
ANTs : multi-scale	n/a	1.3×10^{-3}	Quite Stable
Registration Net	✗	1.2×10^{-3}	Quite Stable
Registration Net	✓	1.2×10^{-3}	Quite Stable

Table 5.11: Results of stabilization of a single clip using ANTs and the registration network. The first row shows the statistics of the original clip (without any stabilization). In the last two rows, the registration network is trained on MSE loss with no penalty ($\lambda = 0$).

Stabilization using all clips: Next, we train the registration model on all the clips constructed from the pullbacks present in the training fold. Table 5.10 shows the results of such an approach. We notice that the first row has the highest Δ MSE; however, a closer look at the output parameters reveals that the model in this case learns to send the clip outside the frame (translation to infinity). In order to mitigate this, we add a penalty to the translation parameters. A simple l_2 penalty seems to over-penalize translation, resulting in a lower Δ MSE. Therefore, we use a U-shaped penalty, in the third row of table 5.12, which performs at par with single-scale ANTs registration. Thus the registration network offers a differentiable stabilization framework that performs as well as classical approaches.

Table 5.13 shows the performance on out-of-sample data. A comparison with table 5.10 shows that the performance drops slightly on out-of-sample data. However, as table 5.14 shows, unlike slower traditional approaches, the trained model is able to correct for motion in real-time.

$\mathcal{L}_{penalty}$	ΔMSE
0	25×10^{-4}
$\ v\ _2^2$	6.2×10^{-4}
$\ \max(0, v - 0.1)\ _2^2$	7.0×10^{-4}

Table 5.12: Results of clip stabilization using Registration network for different penalties on the predicted parameters. Each row reports the results from best of three values of the regularization hyperparameter λ : 0.01, 0.1, 1. Compare this to the ANTs results, table 5.10.

Method	ΔMSE
ANTs: Single-scale	7.0×10^{-4}
ANTs: Multi-scale	8.0×10^{-4}
Registration Network	6.0×10^{-4}

Table 5.13: A comparison of performance on unseen data. Note that while ANTs performs a new optimization on every input, the registration network simply performs one forward pass in order to predict the transformation parameters.

	ANTs: multi-scale	ANTs: single-scale	Registration Net
Time per clip	29.3s	10.2s	0.5s

Table 5.14: Run-time for the three stabilization approaches. Reported values are an average over 3000 clips of length $T = 30$.

Chapter 6

Discussion and Conclusion

6.1 Discussion

In this work, we have explored various ablations, pre-training methods, and registration approaches. Our results show that in cases of severe data imbalance, a meticulously selected loss function has a significant effect, and demonstrate the benefit of using focal loss in such cases (section 5.1.1). We also observe that a model trained on polar IVUS images works better than one trained on Cartesian images (section 5.1.3). The intuitive explanation is that since CNNs are approximately translational equivariant one can hope to achieve approximate rotational equivariance by using convolutions on polar images; and equivariance under rotations is more relevant than that under translations for malapposition classification. Indeed, models trained on polar images perform similar to the $E(2)$ -equivariant models (section 5.1.2), both with and without augmentations. Furthermore, the advantage is lost when data augmentations are introduced.

We explored two pre-training methods, and observed that in each case, a pre-trained model performs better than the baseline. However, when augmentations are introduced, this advantage is lost (section 5.2). We conclude that malapposition detection is a hard task for which both temporal correspondence and video in-painting

are not suitable pre-training objectives.

Our results also demonstrate that cardiac motion can be quantified from IVUS (section 3.2) and used in order to make informed choices for the stabilization algorithm (algorithm 2). We propose a registration network that works well on MNIST images (figure 5-6) and exhibits comparable performance to the multi-scale ANTs based algorithm (algorithm 1) for fitting single clips (table 5.11) during training. However, while fitting on multiple clips, the registration networks performs slightly worse than multi-scale ANTs approach, but comparably to the single-scale approach (table 5.12). The intuitive explanation is that the network sees the images only at one scale, and performs only one pass, instead of multiple blurred and scaled inputs as in multi-scale ANTs. Therefore, we anticipate that a multi-scale version of the network formed by cascading a few layers of the original network, taking inputs at different scales, and composing the resulting transforms, will perform comparably to the multi-scale version of ANTs. Finally, we find that a trained network performs slightly worse on unseen data (table 5.13). This may be partly remedied by introducing additional regularization.

6.1.1 Future Directions

Overall, we find that constructing good representations of IVUS images from a small dataset is a hard task owing to their low signal-to-noise ratio. We have already seen that methods like augmentations or $E(2)$ -equivariant networks that make the representations equivariant to different noise mechanisms (intensity, blur, translation, rotation) are a powerful tool to improve results on downstream tasks. Therefore, a promising future direction is to make the encoding invariant or equivariant to other artifacts such as shadowing, echoing, phantom walls, etc. (section 2.1.2). One way to achieve this is to construct additional augmentations specific to intravascular ultrasound imaging.

Another way to construct better representations is to pre-process the inputs by

explicitly reducing noise such as motion artifacts. Therefore, another promising future direction is to use the registration methods developed in this work to pre-process the input clips before feeding them into the encoders.

6.2 Conclusion

This work explores learning useful representations of intravascular ultrasound images from a small dataset by taking advantage of well established techniques such as augmentations and pre-training. A careful application of these techniques increases the balanced accuracy of malapposition classification from 58% to 74%. Furthermore, the work introduces two methods of mitigating motion artifacts in the IVUS pullbacks. The registration network presents a differentiable method for image registration, and is able to stabilize unseen videos with a speed ~ 20 times faster than the classical approaches, while compromising only slightly in the quality.

Appendix A

Classification Metrics

We provide definitions of some common binary classification performance metrics. The true binary labels and the labels predicted by a binary classifier together partition the data into four disjoint sets. Accordingly, we define true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as follows:

- **True Positives (TP):** the number of instances that are actually positive and are correctly classified as positive by the model.
- **True Negatives (TN):** the number of instances that are actually negative and are correctly classified as negative by the model.
- **False Positives (FP):** the number of instances that are actually negative but are incorrectly classified as positive by the model.
- **False Negatives (FN):** the number of instances that are actually positive but are incorrectly classified as negative by the model.

Then, we may define the following classification performance metrics:

1. **Accuracy:** Accuracy reflects the degree to which a predicted value matches the actual value. It is defined as the ratio of the number of correct predictions

to the total number of predictions made. Mathematically,

$$\text{Accuracy (ACC)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (\text{A.1})$$

2. **Precision:** Precision reflects the degree to which a predicted positive value is actually positive. Mathematically,

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (\text{A.2})$$

3. **True Positive Rate:** True positive rate measures the fraction of positive instances that are correctly classified. It is also called *recall* or *sensitivity*. Mathematically,

$$\text{Recall (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (\text{A.3})$$

4. **F1 score:** The F1 score is a statistical measure that combines both precision and recall to provide a single metric for evaluating the performance of a classification model. It is defined as the harmonic mean of precision and recall. Mathematically,

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (\text{A.4})$$

5. **True Negative Rate:** Similarly to the true positive rate, true negative rate measures the fraction of negative instances that are correctly classified. Its also called *specificity*. Mathematically,

$$\text{Specificity (TNR)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (\text{A.5})$$

6. **Balanced Accuracy:** Balanced accuracy is a statistical measure that is commonly used in classification tasks where the class distribution is imbalanced. It is defined as the average of the true positive rate (TPR) and true negative rate

(TNR). Mathematically,

$$\text{BalancedAccuracy (BA)} = \frac{\text{TNR} + \text{TPR}}{2} \quad (\text{A.6})$$

Appendix B

Fold-wise Breakdown of the Dataset

For cross validation, the 28 labelled IVUS pullbacks were divided into four folds, each containing seven runs. Table B.1 shows the pullback IDs that were included for each of the four folds. Table B.2 shows the frame-wise breakdown of the folds.

Fold#0	Fold#1	Fold#2	Fold#3
PD1F8661	PDGXX83G	PD2ZWGIA	PDZZC35W
PDBYVJ49	PD4AI4VP	PDUELC5Q	PD8BV7LL
PD7X95ZF	PD9BD8NQ	PD21GP2E	PDN7CHZM
PDJB5MCG	PDBZAINH	PDV81ZQO	PD4ZZ87V
PDUSG6PY	PDRXEGJ1	PDLKZB3R	PDLRSMLG
PD7COUKT	PD1H8QJE	PDIA5KDD	PDG1NTSU
PDP1M8D8	PD46JH8G	PDU9MNR7	PD5WUZGE

Table B.1: The patient IDs included in the four folds.

Fold	Malapposed (+)	Healthy (-)	Discarded	Total
Fold#0	1329	7552	2309	11190
Fold#1	773	11358	2478	14609
Fold#2	1028	9827	2408	13263
Fold#3	565	8522	1477	10564
Total	3695	37259	8672	49626

Table B.2: Fold-wise breakdown of malapposed (label (+)), well-apposed (label (-)), and discarded frames. The labels were assigned based on the strategy depicted in figure 3-5.

Bibliography

- [1] *What's new in cardiovascular imaging?* Developments in Cardiovascular Medicine ; Volume 204. Springer-Science+Business Media, B.V., Dordrecht, 1st ed. 1998. edition, 1998.
- [2] Automatic border detection in intravascular ultrasound images for quantitative measurements of the vessel, lumen and stent parameters. *International Congress Series*, 1230:916–922, 2001. Computer Assisted Radiology and Surgery.
- [3] Labeeb Mohsin Abdullah, Nooritawati Md Tahir, and Mustaffa Samad. Video stabilization based on point feature matching technique. In *2012 IEEE Control and System Graduate Research Colloquium*, pages 303–307, 2012.
- [4] Youngoh Bae, Soo-Jin Kang, Geena Kim, June-Goo Lee, Hyun-Seok Min, Hyungjoo Cho, Do-Yoon Kang, Pil Hyung Lee, Jung-Min Ahn, Duk-Woo Park, Seung-Whan Lee, Young-Hak Kim, Cheol Whan Lee, Seong-Wook Park, and Seung-Jung Park. Prediction of coronary thin-cap fibroatheroma by intravascular ultrasound-based machine learning. *Atherosclerosis*, 288:168–174, 2019.
- [5] Retesh Bajaj, Xingru Huang, Yakup Kilic, Anantharaman Ramasamy, Ajay Jain, Mick Ozkor, Vincenzo Tufaro, Hannah Safi, Emrah Erdogan, Patrick W. Serruys, James Moon, Francesca Pugliese, Anthony Mathur, Ryo Torii, Andreas Baumbach, Jouke Dijkstra, Qianni Zhang, and Christos V. Bourantas. Advanced deep learning methodology for accurate, real-time segmentation of high-resolution intravascular ultrasound images. *International Journal of Cardiology*, 339:185–191, 2021.
- [6] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. VoxelMorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, Aug 2019.
- [7] Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise, 2017.
- [8] Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E. Newby, Rohan Dharmakumar, and Sotirios A. Tsfataris. Disentangled representation learning in cardiac image analysis. *Medical Image Analysis*, 58:101535, 2019.

- [9] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58:101539, 2019.
- [10] Hyungjoo Cho, Soo-Jin Kang, Hyun-Seok Min, June-Goo Lee, Won-Jang Kim, Se Hun Kang, Do-Yoon Kang, Pil Hyung Lee, Jung-Min Ahn, Duk-Woo Park, Seung-Whan Lee, Young-Hak Kim, Cheol Whan Lee, Seong-Wook Park, and Seung-Jung Park. Intravascular ultrasound-based deep learning for plaque characterization in coronary artery disease. *Atherosclerosis*, 324:69–75, 2021.
- [11] F. Ciompi, S. Balocco, J. Rigla, X. Carrillo, J. Mauri, and P. Radeva. Computer-aided detection of intracoronary stent in intravascular ultrasound sequences. *Med Phys Oct*, 43:10, 2016.
- [12] Francesco Ciompi, Oriol Pujol, Carlo Gatta, Marina Alberti, Simone Balocco, Xavier Carrillo, Josepa Mauri-Ferre, and Petia Radeva. Holimab: A holistic approach for media–adventitia border detection in intravascular ultrasound. *Medical Image Analysis*, 16(6):1361–8415, 2012.
- [13] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.
- [14] Clifton W Callaway Alanna M Chamberlain Alexander R Chang Susan Cheng Stephanie E Chiuve Mary Cushman Francesca N Delling Rajat Deo Emelia J Benjamin, Salim S Virani. Heart disease and stroke statistics—2018 update: a report from the american heart association. 137(12):e67–e492, 2018.
- [15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019.
- [16] G. Finet, E. Maurincomme, A. Tabib, R.J. Crowley, I. Magnin, R. Roriz, J. Beaune, and M. Amiel. Artifacts in intravascular ultrasound imaging: Analyses and implications. *Ultrasound in Medicine and Biology*, 19(7):533–547, 1993.
- [17] Yabo Fu, Yang Lei, Tonghe Wang, Walter J Curran, Tian Liu, and Xiaofeng Yang. Deep learning in medical image registration: a review. *Physics in Medicine and Biology*, 65(20):20TR01, oct 2020.
- [18] Carlo Gatta, Oriol Pujol, Oriol Rodríguez-Leor, Josepa Ferre, and Petia Radeva. Fast rigid registration of vascular structures in ivus sequences. *Information Technology in Biomedicine, IEEE Transactions on*, 13:1006 – 1011, 12 2009.
- [19] Florin C. Ghesu, Bogdan Georgescu, Awais Mansoor, Youngjin Yoo, Dominik Neumann, Pragneshkumar Patel, R. S. Vishwanath, James M. Balter, Yue Cao, Sasa Grbic, and Dorin Comaniciu. Self-supervised learning from 100 million medical images, 2022.

- [20] Ali Gholipour, Judy A. Estroff, and Simon K. Warfield. Robust super-resolution volume reconstruction from slice acquisitions: Application to fetal brain mri. *IEEE Transactions on Medical Imaging*, 29(10):1739–1758, 2010.
- [21] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018.
- [22] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fr"und, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thureau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *IEEE/CVF International Conference on Computer Vision*, 2017.
- [23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [24] Malte Hoffmann, Andrew Hoopes, Douglas N. Greve, Bruce Fischl, and Adrian V. Dalca. Anatomy-aware and acquisition-agnostic joint registration with synthmorph, 2023.
- [25] Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3922–3931, 2021.
- [26] Allan Jabri, Andrew Owens, and Alexei A. Efros. Space-time correspondence as a contrastive random walk. <https://arxiv.org/abs/2006.14613>, 2020.
- [27] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning, 2021.
- [28] Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Self-supervised learning for spinal mris, 2017.
- [29] Jianbo Jiao, Richard Droste, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. Self-supervised representation learning for ultrasound video, 2020.
- [30] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey, 2019.
- [31] Bernhard Kainz, Markus Steinberger, Wolfgang Wein, Maria Kuklisova-Murgasova, Christina Malamateniou, Kevin Keraudren, Thomas Torsney-Weir, Mary Rutherford, Paul Aljabar, Joseph V. Hajnal, and Daniel Rueckert. Fast volume reconstruction from motion corrupted stacks of 2d slices. *IEEE Transactions on Medical Imaging*, 34(9):1901–1913, 2015.

- [32] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *IEEE/CVF International Conference on Computer Vision*, 2011.
- [33] June-Goo Lee, Jiyuon Ko, Hyeonyong Hae, Soo-Jin Kang, Do-Yoon Kang, Pil Hyung Lee, Jung-Min Ahn, Duk-Woo Park, Seung-Whan Lee, Young-Hak Kim, Cheol Whan Lee, Seong-Wook Park, and Seung-Jung Park. Intravascular ultrasound-based machine learning for predicting fractional flow reserve in intermediate coronary artery lesions. *Atherosclerosis*, 292:171–177, 2020.
- [34] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. In *Advances in Neural Information Processing Systems*, pages 317–327, 2019.
- [35] Philip R. Liebson and Lloyd W. Klein. intravascular ultrasound in coronary atherosclerosis: A new approach to clinical assessment. *American Heart Journal*, 123(6):1643–1660, 1992.
- [36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8417976>, 2018.
- [37] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. <https://arxiv.org/abs/1608.03983>, 2017.
- [38] Aravindh Mahendran, James Thewlis, and Andrea Vedaldi. Cross pixel optical flow similarity for self-supervised learning, 2018.
- [39] Osama Makansi, Eddy Ilg, and Thomas Brox. End-to-end learning of video super-resolution with motion compensation, 2017.
- [40] Hyun-Seok Min, Dongmin Ryu, Soo-Jin Kang, June-Goo Lee, Ji Hyeong Yoo, Hyungjoo Cho, Do-Yoon Kang, Pil Hyung Lee, Jung-Min Ahn, Duk-Woo Park, Seung-Whan Lee, Young-Hak Kim, Cheol Whan Lee, Seong-Wook Park, and Seung-Jung Park. Prediction of coronary stent underexpansion by pre-procedural intravascular ultrasound-based deep learning. *JACC: Cardiovascular Interventions*, 14(9):1021–1029, 2021.
- [41] Gary S. Mintz. Clinical utility of intravascular imaging and physiology in coronary artery disease. *Journal of the American College of Cardiology*, 64(2):735–1097, 2014.
- [42] Seyed Sadegh Mohseni Salehi, Shadab Khan, Deniz Erdogmus, and Ali Gholipour. Real-time deep pose estimation with geodesic loss for image-to-template rigid registration. *IEEE Transactions on Medical Imaging*, 38(2):470–481, 2019.

- [43] Takeshi Nishi, Rikiya Yamashita, Shinji Imura, Kazuya Tateishi, Hideki Kitahara, Yoshio Kobayashi, Paul G. Yock, Peter J. Fitzgerald, and Yasuhiro Honda. Deep learning-based intravascular ultrasound segmentation for the assessment of coronary artery disease. *International Journal of Cardiology*, 333:55–59, 2021.
- [44] Steven Nissen and Paul Yock. Intravascular ultrasound : Novel pathophysiological insights and current clinical applications. *Circulation.*, 10(1161):604–16, 2001.
- [45] Max L. Olender, Lambros S. Athanasiou, Lampros K. Michalis, Dimitris I. Fotiadis, and Elazer R. Edelman. A domain enriched deep learning approach to classify atherosclerosis using intravascular ultrasound imaging. *IEEE Journal of Selected Topics in Signal Processing*, 14(6):1210–1220, 2020.
- [46] Viorica Patraucean, Ankur Handa, and Roberto Cipolla. Spatio-temporal video autoencoder with differentiable memory. 2015.
- [47] Mohammad Reza Rezaei-Dastjerdehei, Amirmohammad Mijani, and Emad Fatemizadeh. Addressing imbalance in multi-label classification using weighted cross entropy loss function. In *2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME)*, pages 333–338, 2020.
- [48] Misael Rosales, Petia Radeva, Oriol Rodriguez-Leor, and Debora Gil. Modelling of image-catheter motion for 3-d ivus. *Medical Image Analysis*, 13(1):91–104, 2009. Includes Special Section on Medical Image Analysis on the 2006 Workshop Microscopic Image Analysis with Applications in Biology.
- [49] Arthur Shiyovich and Ran Kornowski. Chapter 19 - dedicated thrombus-containing stent platforms. In On Topaz, editor, *Cardiovascular Thrombus*, pages 285–302. Academic Press, 2018.
- [50] Leslie N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay, 2018.
- [51] Hae-Geun Song, Soo-Jin Kang, Jung-Min Ahn, Won-Jang Kim, Jong-Young Lee, Duk-Woo Park, Seung-Whan Lee, Young-Hak Kim, Cheol Lee, Seong-Wook Park, and Seung-Jung Park. Intravascular ultrasound assessment of optimal stent area to prevent in-stent restenosis after zotarolimus-, everolimus-, and sirolimus-eluting stent implantation. *Catheterization and cardiovascular interventions : official journal of the Society for Cardiac Angiography Interventions*, 83, 05 2014.
- [52] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint, 2012.
- [53] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. In *International Conference on Machine Learning*, 2015.

- [54] Zheng Sun and Jiejie Du. Suppression of motion artifacts in intravascular photoacoustic image sequences. *Biomed. Opt. Express*, 12(11):6909–6927, Nov 2021.
- [55] Yao Yue Meng Qi Sun Huifeng Sun Zheng, Du Jiejie. Deep learning method for motion artifact correction in intravascular photoacoustic image sequence. *IEEE Trans Med Imaging.*, 42(1):66–78, 2023.
- [56] Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. Vimpac: Video pre-training via masked token prediction and contrastive learning. <https://arxiv.org/abs/2106.11250>, 2021.
- [57] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- [58] Alexander G. Truesdell, Mirvat A. Alasnag, Prashant Kaul, Syed Tanveer Rab, Robert F. Riley, Michael N. Young, Wayne B. Batchelor, Akiko Maehara, Frederick G. Welt, Ajay J. Kirtane, and null null. Intravascular imaging during percutaneous coronary intervention. *Journal of the American College of Cardiology*, 81(6):590–605, 2023.
- [59] Nicholas J. Tustison, Philip A. Cook, Arno Klein, Gang Song, Sandhitsu R. Das, Jeffrey T. Duda, Benjamin M. Kandel, Niels van Strien, James R. Stone, James C. Gee, and Brian B. Avants. Large-scale evaluation of ants and freesurfer cortical thickness measurements. *NeuroImage*, 99:166–179, 2014.
- [60] Ning Wang, Yibing Song, Chao Ma, Wengang Zhou, Wei Liu, and Houqiang Li. Unsupervised deep tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1308–1317, 2019.
- [61] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [62] Xiaolong Wang, Kaiming He, and Abhinav Gupta. Transitive invariance for self-supervised visual representation learning, 2017.
- [63] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. *Learning correspondence from the cycle consistency of time*. CVPR, 2019.
- [64] Maurice Weiler and Gabriele Cesa. General $e(2)$ -equivariant steerable cnns. <https://doi.org/10.48550/arXiv.1911.08251>, 2021.
- [65] Wikipedia. Precision and recall. https://en.wikipedia.org/wiki/Precision_and_recall. Accessed: April 19, 2023.

- [66] Ji Yang, Lin Tong, Mehdi Faraji, and Anup Basu. Ivus-net: An intravascular ultrasound segmentation network. In Anup Basu and Stefano Berretti, editors, *Smart Multimedia*, pages 367–377, Cham, 2018. Springer International Publishing.
- [67] L. Zhang, A. Wahle, Z. Chen, J. J. Lopez, T. Kovarnik, and M. Sonka. Predicting locations of high-risk plaques in coronary arteries in patients receiving statin therapy. *in IEEE Transactions on Medical Imaging*, 37(1):151–161, January 2018.
- [68] Shengyu Zhao, Tingfung Lau, Ji Luo, Eric I-Chao Chang, and Yan Xu. Unsupervised 3d end-to-end medical image registration with volume tweening network. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1394–1404, may 2020.
- [69] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection, 2017.