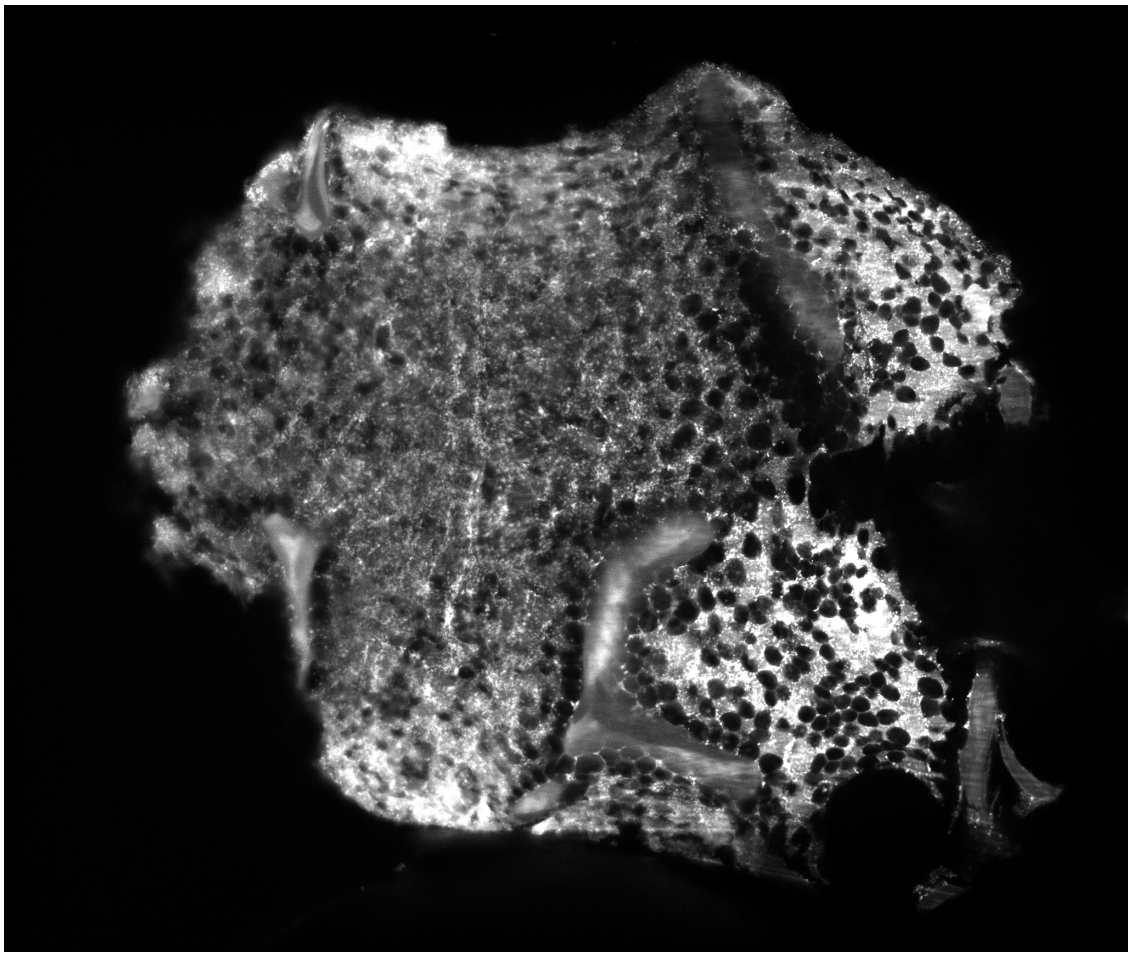

Prototype Based Segmentation of Bone Tissue Microscopy Images

Using Self-Supervised Vision Transformers and Feature Space Similarity

Master's thesis in Complex Adaptive Systems

MATILDA HELLSTRÖM

DEPARTMENT OF PHYSICS

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

Prototype Based Segmentation of Bone Tissue Microscopy Images

Using Self-Supervised Vision Transformers and Feature Space Similarity

MATILDA HELLSTRÖM

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Physics
Soft Matter Lab
Chalmers University of Technology
Gothenburg, Sweden 2026

Prototype Based Segmentation of Bone Tissue Microscopy Images
Using Self-Supervised Vision Transformers and Feature Space Similarity
MATILDA HELLSTRÖM

Supervisor: Mirja Granfors, Department of Physics
Examiner: Giovanni Volpe, Department of Physics

Master's Thesis 2026
Department of Physics
Soft Matter Lab
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 706 525653

Cover: Fluorescence microscopy slice showing an image from the bone tissue dataset used in this thesis.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Prototype Based Segmentation of Bone Tissue Microscopy Images
Using Self-Supervised Vision Transformers and Feature Space Similarity
Matilda Hellström
Department of Physics
Chalmers University of Technology

Abstract

Segmentation of microscopy images serves as a fundamental task within the field of biomedical research and clinical analysis. This thesis investigates whether pretrained self-supervised Vision Transformers, ViTs, can be used for prototype based similarity segmentation of unlabeled bone tissue microscopy images. The framework developed and presented utilizes pretrained DINOv2 backbones to extract feature embeddings from microscopy image patches. Positive and negative reference points are used to construct prototype embeddings, enabling similarity based segmentation within the learned feature space.

To evaluate how model capacity influences the learned feature space and segmentation performance, all available DINOv2 backbone sizes were included in the experiments. Feature space visualizations and prototype transfer experiments further enabled evaluation of representation quality as well as the robustness and generalization capabilities of the proposed framework. In addition, the DINO heatmaps were used as input to a U-Net to investigate whether they could improve segmentation quality in supervised learning.

The results show that pretrained ViTs extract feature representations in which tissue and background regions become partially separable within the learned feature space. PCA and UMAP visualizations indicate, together with clustering metrics, that structurally similar image patches tend to form clusters in the embedding space. The Giant backbone achieved the strongest segmentation performance with a mean dice score of 0.690 and an IoU of 0.534. Prototype transfer performed well within the same sample (mean dice score of 0.644), but performance decreased when transferring prototypes across samples (mean dice score of 0.542), indicating that the framework is sensitive to biological variability and domain shift. Providing a U-Net with the DINO output for refinement improved the dice scores while also reducing boundary alignment errors.

The study demonstrates that pretrained self-supervised Vision Transformers can be used for prototype based segmentation of bone tissue microscopy images. Despite being trained on natural RGB images rather than microscopy data, the evaluated DINOv2 backbones produced feature representations that enabled segmentation of bone structures without any task specific training.

Keywords: self-supervised representation learning, DINOv2, prototype based segmentation, microscopy image segmentation, Vision Transformers, feature space similarity, bone tissue analysis.

Acknowledgements

There are many people I would like to thank for their support and encouragement throughout my master thesis project.

First and foremost, I would like to thank my supervisor Mirja Granfors for all discussions and feedback throughout this project. Your support has been invaluable.

I would also like to thank my examiner Giovanni Volpe for helpful feedback and discussions throughout the project.

A special thank you to Andrei Chagin, Xin Liu and Nelson Tsz Long Chu at Sahlgrenska University Hospital for providing the dataset used in this study, sharing their knowledge and helping me understand the biological context of the project.

I would also like to thank everyone in the Soft Matter Lab for welcoming me into the group and for the many enjoyable lunch meetings and discussions I had the pleasure of being part of.

Finally, I would like to thank my friends and family for their encouragement and support throughout this year. Thank you for listening to my ideas, frustrations and countless thesis updates. Your support means more than you know.

Matilda Hellström, Gothenburg, June 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

DINO	Distillation with No Labels
DINOv2	Distillation with No Labels version 2
GT	Ground Truth
HD95	95th Percentile Hausdorff Distance
IoU	Intersection over Union
kNN	k-Nearest Neighbors
PCA	Principal Component Analysis
SSL	Self-Supervised Learning
UMAP	Uniform Manifold Approximation and Projection
ViT	Vision Transformer

Contents

List of Acronyms **ix**

List of Figures **xv**

List of Tables **xix**

1 Introduction **1**
1.1 Project Aim . . . 1
1.2 Research Questions . . . 2
1.3 Ethical Considerations . . . 3

2 Theory **5**
2.1 Self-Supervised Learning . . . 5
2.2 Vision Transformers . . . 5
2.3 The DINO Framework . . . 6
2.4 DINOv2 . . . 7
2.5 Cosine Similarity . . . 7
2.6 Prototype Based Segmentation . . . 8
2.7 Feature Space Visualization . . . 8
2.7.1 Principal Component Analysis . . . 9
2.7.2 Uniform Manifold Approximation and Projection . . . 9
2.8 Representation Quality Metrics . . . 9
2.8.1 Silhouette Score . . . 9
2.8.2 k-Nearest Neighbor Retrieval Accuracy . . . 9
2.9 U-Net . . . 10
2.10 Evaluation Metrics . . . 11
2.10.1 Dice . . . 11
2.10.2 Intersection Over Union . . . 11
2.10.3 Hausdorff Distance . . . 11

3 Methods **13**
3.1 Overview . . . 13
3.2 Dataset . . . 15
3.3 Data Preprocessing . . . 16
3.4 Feature Extraction Using DINO . . . 17
3.4.1 DINOv2 Backbone Variants . . . 17
3.5 Prototype Based Segmentation . . . 18

3.5.1 Sampling Reference Points . . . 18
3.5.2 Construction of Prototype Embeddings . . . 19
3.5.3 Producing Similarity Maps and Segmentation Masks . . . 20
3.6 Thresholding Strategy . . . 20
3.7 U-Net Refinement . . . 21
3.8 Experimental Setup . . . 22
3.8.1 Dataset Split and Training Approach . . . 22
3.8.2 Segmentation Setup . . . 22
3.8.3 U-Net Architecture and Training . . . 23
3.8.4 Evaluation Setup . . . 25
3.8.5 Generalization Across Slices and Samples . . . 26

4 Results **27**
4.1 Representation Quality Analysis . . . 27
4.1.1 Feature Space Visualization . . . 27
4.1.2 Clustering Representation Analysis . . . 27
4.2 Segmentation Performance . . . 28
4.2.1 Backbone Size Comparison . . . 28
4.2.2 Qualitative Segmentation Examples . . . 30
4.2.3 U-Net Refinement . . . 31
4.3 Robustness Analysis . . . 32
4.3.1 Point Sensitivity . . . 32
4.3.2 Threshold Sensitivity . . . 32
4.4 Generalization Analysis . . . 32
4.4.1 Same File Prototype Generalization . . . 33
4.4.2 Cross File Prototype Generalization . . . 35

5 Discussion **37**
5.1 Learned Feature Representations . . . 37
5.1.1 Feature Space Organization . . . 37
5.1.2 Relation Between Representation Quality and Segmentation Performance . . . 38
5.2 Prototype Based Segmentation Performance Analysis . . . 38
5.2.1 Influence of Backbone Capacity . . . 38
5.2.2 Dataset Dependent Behavior . . . 39
5.2.3 U-Net Refinement . . . 39
5.3 Generalization and Domain Shift . . . 39
5.3.1 Same Sample Prototype Transfer . . . 40
5.3.2 Cross Sample Prototype Transfer . . . 40
5.4 Limitations and Future Work . . . 40

6 Conclusion **43**
6.1 Summary of Findings . . . 43
6.2 Main Contributions . . . 43
6.3 Final Remarks . . . 44

A Appendix 1 **I**

A.1 Additional Segmentation Results I
A.2 Additional U-Net Refinement Results I

B Appendix 2 **V**
B.1 Additional Prototype Transfer Results V

List of Figures

2.1 *Overview of the DINO framework. An input image is transformed into multiple global and local crops using multi-crop augmentation. Global crops are processed by the teacher network, while the student network receives both global and local crops. The student is trained to match the teacher output distribution and is updated using stochastic gradient descent. The teacher parameters are updated as an exponential moving average of the student parameters, making it a stable target.* 6

2.2 *Overview of the U-Net architecture. The network consists of an encoder path (blue), a bottleneck layer (red) and a decoder path (green) connected through skip connections (gray arrows).* 10

3.1 *Overview of the proposed prototype based segmentation framework. a. Input image. b. Splitting the image into patches. c. Selection of positive and negative reference points. d. Localization of the image patches corresponding to the selected reference points. e. Extraction of patch embeddings using a pretrained DINOv2 Vision Transformer backbone. f. Similarity heatmap generated by computing cosine similarity scores for the prototype embeddings with all image patch embeddings. g. Final segmentation mask obtained through thresholding.* . 14

3.2 *Representative slices from the five bone samples used in this study. Visible differences in tissue morphology, image contrast and structural organization between the samples can be observed.* 15

3.3 *Representative example from the bone tissue dataset used in this study. a. Raw fluorescence microscopy image. b. Expert annotation used as the ground truth mask. c. Overlay of the ground truth mask on the microscopy image, highlighting the structures used for training and evaluation in red.* 16

3.4 *Example of positive and negative reference point sampling. Positive reference points (green) are sampled within the ground truth mask and represent the structure of interest, while negative reference points (red) are sampled from background regions. The sampled points are used to construct positive and negative prototype embeddings for the similarity based segmentation framework.* 19

3.5 *Examples of the inputs to the U-Net. From left to right: the raw microscopy image, the corresponding DINO similarity heatmap, the heatmap overlaid on the raw image and the thresholded DINO segmentation mask and heatmap overlaid on the raw image. These setups were evaluated as inputs to the U-Net.* 23

4.1 *Representative PCA and UMAP visualizations of patch embeddings extracted using the DINOv2 Small backbone. Structure and background patches are shown by orange and blue dots respectively, while positive and negative prototype embeddings are marked by green and red circles.* 28

4.2 *Representative PCA and UMAP visualizations of patch embeddings extracted using the DINOv2 Giant backbone. Structure and background patches are shown by orange and blue dots respectively, while positive and negative prototype embeddings are marked by green and red circles.* 28

4.3 *Mean dice score for each sample and DINOv2 backbone variant. Results are averaged across the evaluated slices for each sample.* 29

4.4 *Qualitative segmentation examples corresponding to the worst, median and best performing slices. The first column shows the microscopy images together with the sampled positive and negative reference points. The second column shows the generated similarity heatmaps, while the third column and forth columns visualizes the predicted segmentation (blue) and the ground truth annotation (red).* 30

4.5 *Representative comparison between the prototype based DINO segmentation framework and the evaluated U-Net refinements. The figure illustrates how different U-Net inputs influence the resulting segmentation masks shown in cyan relative to the ground truth annotation marked in magenta.* 31

4.6 *Point sensitivity plot illustrating the mean dice score as a function of the number of positive and negative reference points used for prototype construction.* 33

4.7 *Threshold sensitivity plot showing the mean dice score as a function of threshold value for the evaluated samples illustrating how segmentation performance changes as different threshold values are used for generating segmentation masks.* 33

4.8 *Representative example of prototype transfer within the same sample. Prototype embeddings extracted from one bone sample were applied to a slice from a the same sample. The figure content from top left to right: the microscopy image which the reference points were picked from, the target slice, the cosine similarity heatmap for the target image, the predicted mask and the corresponding ground truth mask.* 34

4.9 *Prototype transferability matrix showing the mean dice score when prototype embeddings extracted from a reference sample were used to segment images from a target sample. Rows correspond to the reference sample and columns correspond to the target sample.* 35

4.10 *Representative example of cross file prototype transfer. Prototype embeddings extracted from one bone sample were used on a slice from a different sample. The figure shows the raw microscopy images, the generated similarity heatmap, the predicted segmentation mask and the corresponding ground truth mask.* 36

A.1 *Additional qualitative segmentation examples for all samples. The worst, median and best performing slices based on the dice score are shown for each sample. Predicted segmentation masks are marked in blue and ground truth masks in red.* II

A.2 *Qualitative U-Net refinement results for the worst, median and best performing slices from each sample based on the dice improvement. Blue dashed contours correspond to the original DINO prediction, lime denote the refined U-Net prediction and red corresponds to the ground truth mask. The shown values are the dice scores before and after refinement together with the dice improvement.* III

List of Tables

3.1 *Overview of the dataset used throughout this study. Due to computational limitations, 100 evenly distributed slices were selected from each sample for evaluation. The penultimate channels were used for conducting the similarity based segmentation* 15
3.2 *Overview of the evaluated pretrained DINOv2 Vision Transformer backbones [1].* 18
3.3 *Parameters and corresponding values used throughout the study.* . . . 23
3.4 *Overview of the U-Net architecture used in the refinement experiments.* 24
3.5 *Training settings for the U-Net experiments.* 25

4.1 *Feature space separability and retrieval accuracy for the evaluated backbone variants. The silhouette score quantifies the global separation between structure and background embeddings while the kNN retrieval accuracy measures the local feature space consistency. Values are reported as mean ± standard deviation across the used bone slices.* 29
4.2 *Segmentation performance for the evaluated pretrained DINOv2 backbone variants. Values are reported as mean ± standard deviation across the evaluated slices.* 29
4.3 *Segmentation performance for the evaluated U-Net input configurations. Values are reported as mean performance across all evaluated samples.* 31
4.4 *Prototype transfer performance for the evaluated transfer settings. Results are reported as mean ± standard deviation for prototype transfer within the same slice, within the same sample and across different bone samples.* 34

B.1 *Mean dice scores from extracting prototype embeddings from a reference sample and transferring them to a target sample. Values are reported as mean ± standard deviation.* V

1
Introduction

Segmentation of microscopy images constitutes a fundamental task within the field of medical image analysis. Accurate segmentation masks are often required for conducting analysis tasks such as distance measurements, volume estimation and surface analysis. In turn, these tasks form the foundation for a broad range of applications - ranging from disease analysis to academic research. Achieving accurate segmentation masks is therefore of central importance for microscopy image analysis.

However, the process of segmenting microscopy images generally relies on either manual annotation or supervised machine learning frameworks. As the latter requires large amounts of labeled data for training to be meaningful, both alternatives imply time consuming manual data annotation, which makes the segmentation process highly difficult to scale [2]. The need for finding alternative methods within the field of microscopy image segmentation is consequently pronounced.

Recent advances within the field of self-supervised learning have, however, demonstrated that meaningful representation learning can be achieved directly from unlabeled data [3]. Vision Transformers, ViTs, trained using the self-supervised machine learning framework DINO have showed strong representation learning capabilities without the need for annotated data or human intervention [4, 1].

To further explore the possibilities of self-supervised learning for segmentation tasks, this thesis investigates whether pretrained DINOv2 Vision Transformer backbones can be utilized for prototype based segmentation of bone tissue microscopy images.

The proposed framework combines feature extraction using pretrained self-supervised DINOv2 backbones with similarity based localization within the learned feature space using positive and negative prototype embeddings. The conducted experiments consequently enable evaluation of the representation quality and segmentation capabilities of pretrained self-supervised DINOv2 ViT backbones for microscopy bone tissue images.

1.1 Project Aim

The objective of this study is to develop and examine a similarity and prototype based segmentation pipeline for bone tissue microscopy images. The primary focus

is centered upon how pretrained DINOv2 backbones encode microscopy image information and how the learned representations can be utilized for segmentation by similarity comparsion of feature embeddings.

The proposed framework utilizes pretrained self-supervised DINOv2 ViT backbones to extract feature embeddings from image patches. In combination with specified reference points of what structures should be segmented and what should not, the prototype examples enable similarity comparsion between the prototype and the image patch embeddings. The resulting similarity scores are then used to generate similarity heatmaps and are thresholded to produce binary segmentation masks.

Using this methodology, the study aims to examine the potential of self-supervised machine learning frameworks for bone tissue image segmentation and investigate whether the developed framework can contribute towards more efficient analysis of bone tissue microscopy images.

The study further introduces a U-Net refinement experiment to assess how the self-supervised framework performs when compared to supervised alternatives. This stage will also help answer whether the self-supervised outputs can be enhanced using post-processing methods.

Additionally, the work aims to validate whether the proposed framework could be used for developing a software integration for more precise tissue analysis of bone samples at Sahlgrenska University Hospital, which requires analysis of the robustness and generalization capabilities of the developed framework.

1.2 Research Questions

This study seeks to answer whether pretrained DINOv2 vision transformer backbones can learn meaningful feature representations from unlabeled bone tissue microscopy images and whether these representations can be used to generate accurate segmentation masks with minimal manual intervention. To successfully do the latter, a framework utilizing self-supervised representations and a prototype based similarity segmentation approach was developed.

Summarized, the questions this study seeks to answer are the following:

1. To what extent can pretrained DINOv2 feature representations be utilized for segmenting bone tissue microscopy images?
2. What factors influence the segmentation quality, stability and robustness of the proposed framework?
3. How does the proposed self-supervised framework compare to a supervised refinement approach based on using a U-Net?
4. Can supervised refinement improve the segmentation masks generated by the proposed framework?

5. How well do the learned feature representations generalize across different microscopy slices and bone samples?

These questions are evaluated quantitatively using segmentation metrics and qualitatively through visual analysis.

1.3 Ethical Considerations

The dataset used for this study consists of human bone biopsy samples collected during femoral head replacement surgery. All data were provided by the Department of Internal Medicine and Clinical Nutrition at Sahlgrenska University Hospital and were used for research purpose. All data is therefore anonymized and the datasets are unpublished.

As the study aims to develop methods that could facilitate image analysis, the hope is to contribute to reduce the need for manual annotation of datasets. This could, in turn, help make microscopy image analysis more scalable. The results presented in this thesis should therefore be viewed as an investigation of the segmentation potential of self-supervised Vision Transformers for microscopy images and not for providing clinically accurate segmentation masks.

2
Theory

2.1 Self-Supervised Learning

Training of supervised machine learning frameworks commonly relies on large amounts of labeled data. However, annotation of microscopy image data is often time consuming, difficult to scale and, as a consequence, often impractical [5]. Finding alternatives reducing the need for annotated datasets have thereforew become increasingly important.

Self-supervised learning, SSL, is an umbrella term for machine learning frameworks that enable models to learn without the need for labels or annotations [3]. Unlike supervised machine learning approaches, self-supervised alternatives instead learn from structures and patterns within the unlabeled data [4]. The dependency on labeled datasets and manual intervention is consequently reduced significantly.

As the amount of labeled datasets often is scarce when working with medical microscopy images, self-supervised learning has become a promising alternative to supervised frameworks for segmentation of medical images [6].

2.2 Vision Transformers

When Vision Transformers were introduced by Alexey Dosovitskiy et al. at Google Research, a new paradigm of computer vision was initiated [7]. Instead of processing images using convolutions, ViTs divide images into patches and processes them as a sequence - comparable to how tokens are processed by transformers [8].

The ViTs transform the image patches into vector representations - referred to as feature embeddings. These embeddings contain information about the image patch and its visual contents. Self-attention is a core component of ViTs, allowing them to learn global and long ranging relationships within an image. Hence, the embeddings hold information not only about the what is present in the individual patch, but also about its relation to the rest of the image [7, 8].

The embeddings together form a feature space. The organization of the learned feature space also makes localization of similar looking patches possible - if two patches contain visually similar structures, their vector embeddings should be alike.

By localizing a patch within a cluster in the feature space, it is therefore possible to find vectors corresponding to patches containing similar structures [3].

2.3 The DINO Framework

DINO is a self-supervised learning framework based on a teacher-student architecture developed by META [4]. The core idea of the framework is to train two neural networks, referred to as a student and a teacher network, to output the same probability distributions using data without labels or annotations. An overview of the framework and its main components is illustrated in Figure 2.1.

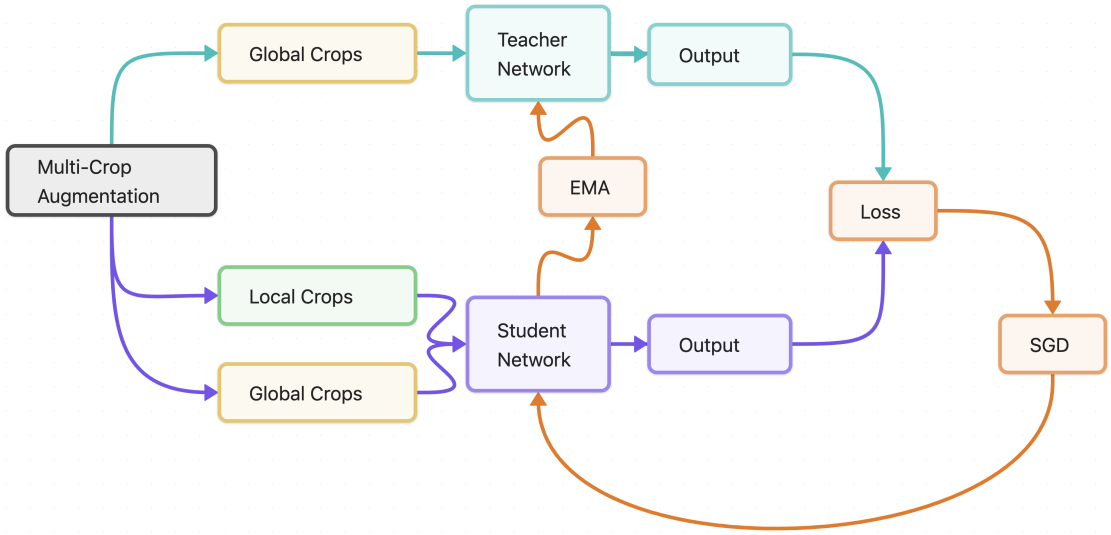

Figure 2.1: *Overview of the DINO framework. An input image is transformed into multiple global and local crops using multi-crop augmentation. Global crops are processed by the teacher network, while the student network receives both global and local crops. The student is trained to match the teacher output distribution and is updated using stochastic gradient descent. The teacher parameters are updated as an exponential moving average of the student parameters, making it a stable target.*

To minimize the loss between the teacher and student outputs, the student weights are updated using stochastic gradient descent. The teacher weights are computed as the exponential moving average of the student weights, making the teacher a stable target which the student can learn from.

Moreover, the teacher and the student network are not provided the same images. Instead, multi-crop augmentation is used for producing global and local crops, which forces the networks to learn position and scale independent features.

The global crops, which contain a large proportion of the original image, are fed to the teacher network. Local crops, on the other hand, contain smaller portions of the image. The student network is provided both global and local crops with the goal

of matching the output of the teacher network.

To make the outputs match, the difference in the output distributions from the student and teacher networks is minimized using cross-entropy loss, see Equation 2.1 below, where $P_t(x)$ and $P_s(x)$ denote the output probability distributions from the teacher and student networks.

$$\min_{\theta_s} -P_t(x) \log P_s(x) \tag{2.1}$$

Normalization of the teacher output, referred to as centering, is also applied to stabilize training and reduce the risk of overfitting or collapsing. Moreover, temperature scaling constitutes a core component for making the training stable as it helps control the sharpness of the output distribution [4].

Although the DINO framework can be used to train several types of neural network architectures, it has proven to be particularly effective when combined with Vision Transformers [4, 9]. The learned feature spaces produced by ViTs trained using DINO tend to be highly organized, making the pretrained backbones suitable as frozen feature extractors. This enables direct and efficient use of the models for both classification, retrieval and segmentation tasks [1].

2.4 DINOv2

DINOv2 is an enhanced version of the original DINO framework developed by META that is trained on larger and more diverse image datasets using improved training strategies. DINOv2 has demonstrated strong transferability across several downstream tasks, allowing pretrained backbones to be used directly as frozen extractors without requiring any task specific training [1].

2.5 Cosine Similarity

Cosine similarity is a commonly used similarity measure for comparing feature embeddings within a learned feature space. The metric quantifies how similar two vectors are by measuring the angle between them [3].

For two vectors $\mathbf{a}$ and $\mathbf{b}$, the cosine similarity, S, is defined as

$$S = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}, \tag{2.2}$$

where $\mathbf{a} \cdot \mathbf{b}$ denote the dot product between the vectors and $\|\mathbf{a}\|$ and $\|\mathbf{b}\|$ are their Euclidean norms.

The cosine similarity values ranges between -1 and 1, where values close to 1 indicate highly similar vectors, values close to 0 correspond to orthogonal vectors and values close to -1 is obtained when two vectors are highly dissimilar. Image patches containing visually similar objects should produce embeddings with similar directions in the learned feature space [4]. Cosine similarity is therefore a suitable metric for identifying image regions corresponding to similar tissue structures and forms a central component of the prototype based segmentation framework used throughout this thesis.

2.6 Prototype Based Segmentation

Prototype based segmentation is a segmentation approach which is based on utilizing the organization of a learned feature space for similarity comparison. The framework relies on how visually similar patches of an image should obtain similar feature embeddings, resulting in cluster formation within the learned feature space. Consequently, patches containing the same tissue types are expected to occupy the same regions of the feature space [10].

Segmentation of a certain tissue type therefore becomes a task of analyzing cluster formation in the learned feature space. By specifying feature embeddings - prototype embeddings - representing the structure to segment, the framework does not require any task specific training or annotated data.

Comparison between image patch embeddings and prototype embeddings is performed using a similarity metric - such as cosine similarity. The resulting scores can then be combined into a similarity map describing how well different image regions match the selected prototypes. Regions with high similarity to the structure prototype and low similarity to background prototypes are therefore likely target structure and vice versa.

Binary segmentation masks can be obtained by applying thresholding to the generated similarity masks. Hence, prototype based segmentation becomes a task of retrieving information about the learned feature embeddings and how these compare to the specified prototype embeddings. This is an advantage compared to conventional segmentation frameworks as it enables training free segmentation [11]. The proposed framework can therefore be considered a training free zero-shot segmentation approach.

2.7 Feature Space Visualization

To enable interpretation and analysis of high dimensional feature representations, dimensionality reduction techniques can be used to project the data to a lower dimensional space. This section describes the two techniques used in this thesis.

2.7.1 Principal Component Analysis

Principal Component Analysis, PCA, is a linear dimensionality reduction technique that projects high dimensional data into a lower dimensional space while preserving as much of the data variance as possible [12]. PCA is widely used for analysis and visualization of high dimensional feature representations.

2.7.2 Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection, UMAP, is a non-linear dimensionality reduction technique designed to preserve local neighborhood relationships within high dimensional data [13]. Like PCA, UMAP is a commonly used tool for visualizing and analyze learned feature spaces.

2.8 Representation Quality Metrics

The quality of learned feature representations can be evaluated by analyzing how well semantically similar samples get organized within the feature space [3]. In this thesis, representation quality is evaluated using silhouette score and k-nearest neighbor retrieval accuracy.

2.8.1 Silhouette Score

The silhouette score is a widely used clustering metric that quantifies how well samples belonging to different groups are separated within a feature space [14]. The silhouette score for a sample is defined as

$$s(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))}, \tag{2.3}$$

where a(x) denotes the average distance to samples belonging to the same group and b(x) denotes the average distance to samples belonging to the nearest neighboring group.

The silhouette score ranges from -1 to 1. Values close to 1 indicate highly separated clusters, 0 correspond to overlapping clusters and negative values are obtained when the feature space is highly unorganized and samples belong to the wrong clusters.

2.8.2 k-Nearest Neighbor Retrieval Accuracy

k-Nearest Neighbor, kNN, retrieval accuracy measures the local properties of a feature space. The metric is computed by specifying the k nearest neighbors for each sample. The fraction of neighboring samples that belong to the same class as the

considered sample is what forms the retrieval accuracy. High retrieval accuracy is thus obtained when the learned feature space is well organized with semantically similar samples located close to each other [4, 1].

2.9 U-Net

A U-Net is a convolutional neural network architecture developed by Ronneberger et al. for biomedical image segmentation [15]. The architecture was designed to enable accurate segmentation while requiring limited amounts of annotated training data, which makes it particularly suitable for medical imaging applications.

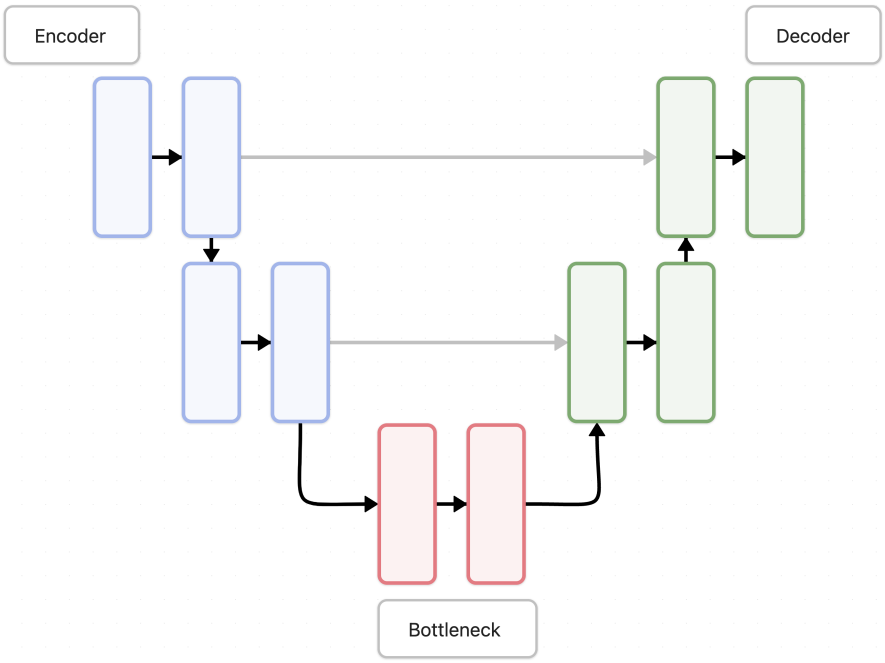

Figure 2.2: *Overview of the U-Net architecture. The network consists of an encoder path (blue), a bottleneck layer (red) and a decoder path (green) connected through skip connections (gray arrows).*

As shown in Figure 2.2, the network consists of three main components - an encoder, a decoder and skip connections between them - as illustrated in Figure 2.2. The encoder downsamples an input image using convolution and pooling operations, which enables the network to learn feature representations. The decoder works the opposite way by upsampling, allowing the network to produce segmentation masks [15].

The skip connections are central for transferring information directly between the encoder and decoder. This allows the decoder to combine high level information with finer details, resulting in more precise segmentation. This is especially important for microscopy image segmentation due to the need for accurate boundary localization. This is what motivates the wide usage of U-Nets within the field of

biomedical image segmentation.

2.10 Evaluation Metrics

The segmentation performance of the proposed framework is evaluated using three main metrics - dice, intersection over union and the 95th percentile Hausdorff distance. These metrics provide information about how well the predicted and the ground truth masks overlap, as well as how well the mask boundaries match.

2.10.1 Dice

Dice has been used as the main metric for evaluation of mask overlap. It is a standard metric commonly used for segmentation and is defined as

$$\text{Dice} = \frac{2TP}{2TP + FP + FN}, \tag{2.4}$$

where TP, FP and FN denote the true positive, the false positive and the false negative pixels when comparing the predicted and ground truth masks. If a perfect overlap occurs, the dice score will equal 1, while 0 is obtained when the predicted and ground truth masks don't overlap at all [16].

2.10.2 Intersection Over Union

Intersection over union or IoU, also known as Jaccard index, has been used in addition to dice for measuring the overlap between the predicted and ground truth masks. The metric is defined as

$$\text{IoU} = \frac{TP}{TP + FP + FN}. \tag{2.5}$$

Hence, IoU is a more strict metric compared to dice. The values of IoU also range from 0 to 1, where 0 is a result of no overlap, whereas 1 is obtained for perfect alignment [16].

2.10.3 Hausdorff Distance

The metric Hausdorff Distance, HD, has been used for evaluation of mask boundary accuracy. This metric is defined according to Equation 2.6 and provides information about how well the boundaries align by measuring the largest distance between the mask borders. Hence, low HD values correspond to better aligned masks [16].

The metric is defined as

$$HD(X,Y) = \max\left\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(y,x)\right\}, \tag{2.6}$$

where X and Y are sets of boundary points from the predicted and the ground truth mask respectively and d(x,y) denote the Euclidean distance between two points x and y. To reduce the impact of noise and outliers, the 95th percentile of the Hausdorff distance has been used.

3
Methods

This chapter describes the methodology and experimental design used for the project. The proposed pipeline framework is presented together with data preparation and preprocessing, the feature extraction method, the prototype based similarity segmentation, the thresholding strategies considered, used evaluation metrics and the U-Net refinement experiment used for segmentation enhancement and comparative analysis.

The chapter is structured to provide sufficient detail for reproducing both the segmentation pipeline and the conducted experiments. To enhance clarity and motivate method and design choices, certain methodological discussions are integrated into the chapter. However, broader analysis and discussion of the method choices and limitations are covered in the chapter 5 Discussion.

3.1 Overview

The segmentation pipeline developed in this thesis project performs similarity based segmentation of bone tissue microscopy images using pretrained self-supervised Vision Transformers. The framework is centered around the usage of pretrained DINOv2 ViT backbones for extracting feature embeddings from microscopy images. The foundation of the framework relies on how visually similar image patches should generate similar feature embeddings within the learned feature space, as described in section 2.6 Prototype Based Segmentation.

By specifying reference points that correspond to what structures should be segmented and what should not, patches that represent the structure and background respectively are specified. The embeddings corresponding to these patches are in turn what generates the prototypes that are compared against all image patch embeddings.

More specifically, the prototype embeddings are used for computing a cosine similarity score for each image patch. The computed similarity scores can thereafter be used for generating a similarity map indicating what regions are similar to the structure prototype embedding and dissimilar to the background prototype embedding. Applying thresholding enables for generation of binary segmentation masks containing the structure of interest.

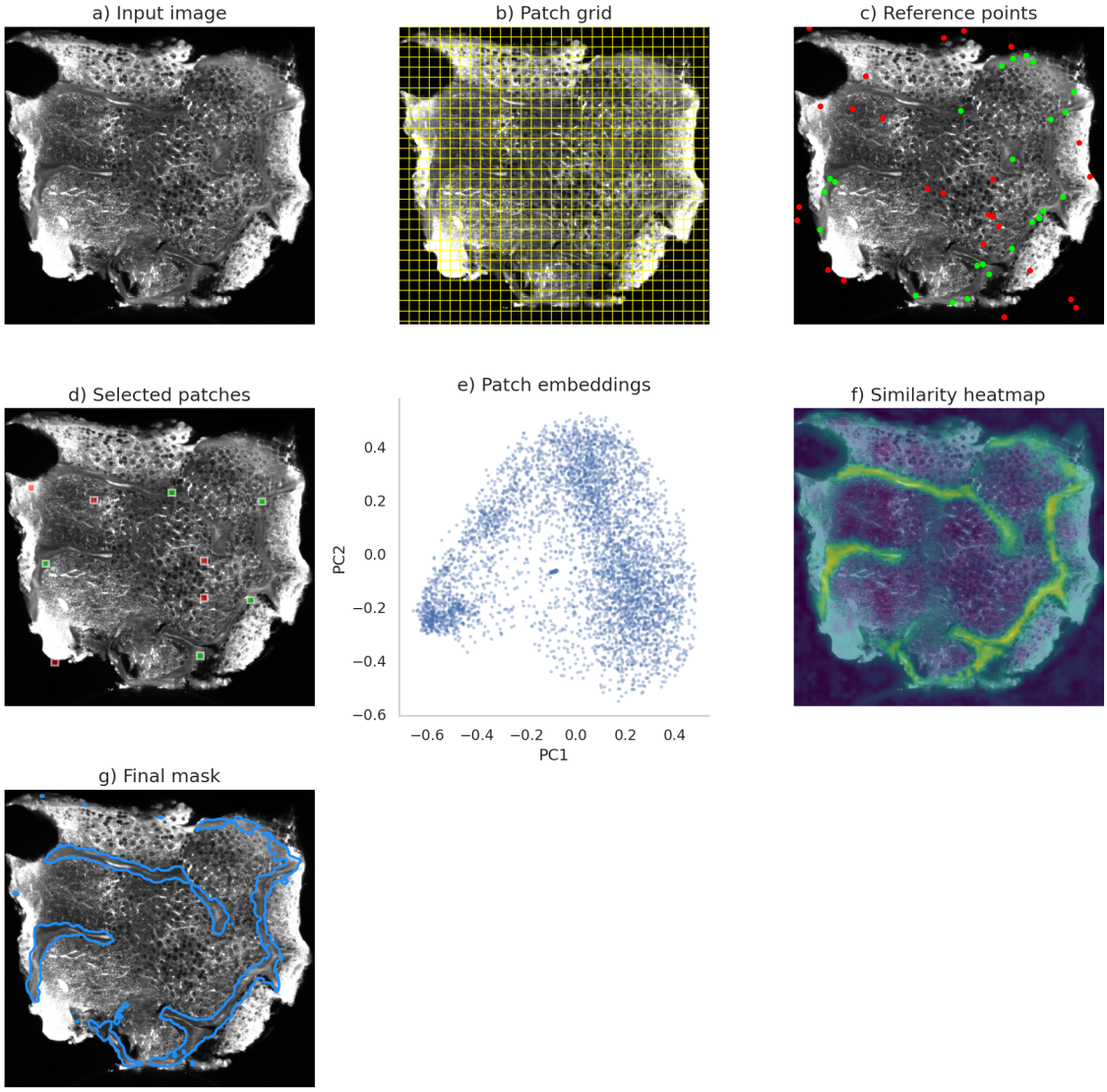

Figure 3.1: *Overview of the proposed prototype based segmentation framework. a. Input image. b. Splitting the image into patches. c. Selection of positive and negative reference points. d. Localization of the image patches corresponding to the selected reference points. e. Extraction of patch embeddings using a pretrained DINOv2 Vision Transformer backbone. f. Similarity heatmap generated by computing cosine similarity scores for the prototype embeddings with all image patch embeddings. g. Final segmentation mask obtained through thresholding.*

The pipeline consists more precisely of the following steps:

- Extraction of patch embeddings using a pretrained DINOv2 ViT backbone.
- Sampling or selecting positive and negative reference points corresponding to the structure and background respectively.
- Construction of prototype embeddings by passing the image patches containing the reference points through the pretrained model.
- Computation of a cosine similarity heatmap using prototype and image patch

embeddings.
- Thresholding to obtain a binary segmentation mask.

Additional experiments were constructed to compare the proposed self-supervised framework with supervised segmentation alternatives. A shallow U-Net was therefore introduced, both for comparison and investigation of whether the generated segmentation masks could be enhanced using supervised post-processing.

3.2 Dataset

The dataset used in this thesis project was provided by the Department of Internal Medicine and Clinical Nutrition at Sahlgrenska University Hospital. The dataset consisted of five human bone biopsy samples obtained from patients undergoing femoral head replacement surgery in connection with hip fracture treatment. The samples are referred to as files A, C, D, E and F throughout the report. Representative slices from the five bone samples are shown in Figure 3.2.

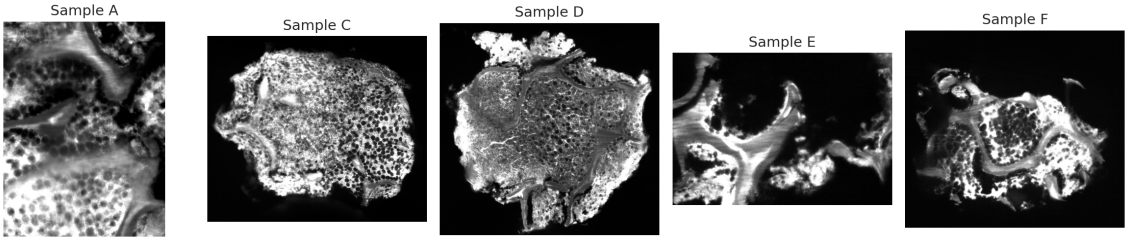

Figure 3.2: *Representative slices from the five bone samples used in this study. Visible differences in tissue morphology, image contrast and structural organization between the samples can be observed.*

The samples were analyzed using fluorescence microscopy and stored as Imaris IMS files, containing both image data for each sample together with manually annotated ground truth segmentation masks produced by an expert in the field.

Table 3.1: *Overview of the dataset used throughout this study. Due to computational limitations, 100 evenly distributed slices were selected from each sample for evaluation. The penultimate channels were used for conducting the similarity based segmentation*

File	Z-slices	Resolution (pixels)	Channels	Evaluated slices
A	253	1026 × 775	5	100
C	632	1725 × 2048	6	100
D	478	2933 × 3050	6	100
E	487	847 × 1228	6	100
F	291	1632 × 1825	6	100

As shown in Table 3.1, substantial variation exists between the samples in terms of both image resolution and the number of available z-slices. The largest sample, file D, contains images with a resolution of 2933 × 3050 pixels, whereas file E contains images with size 847 × 1228 pixels. Such differences may influence both representation learning and the transferability of prototype embeddings between samples.

Due to computational limitations, subsets of 100 slices per sample have been used for evaluation. To capture local differences within a sample, the slices have been picked evenly across each sample. The first and last 40 slices were excluded from each file as these slices often contained limited image content or less reliable ground truth annotations.

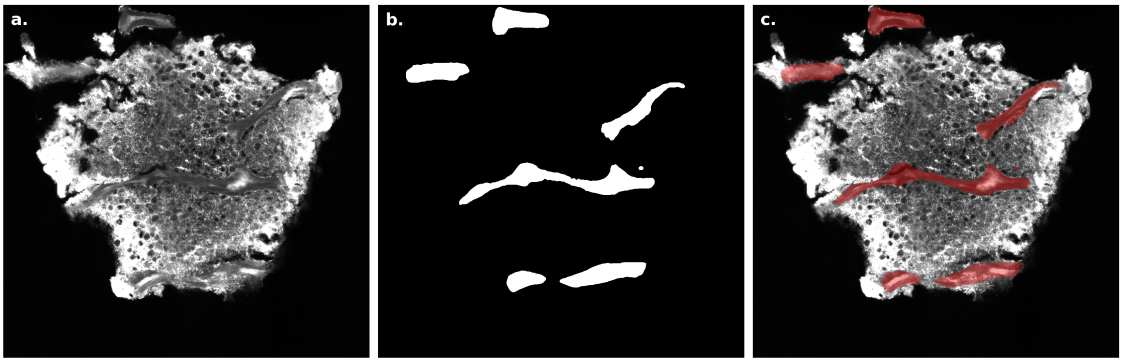

Figure 3.3: *Representative example from the bone tissue dataset used in this study. a. Raw fluorescence microscopy image. b. Expert annotation used as the ground truth mask. c. Overlay of the ground truth mask on the microscopy image, highlighting the structures used for training and evaluation in red.*

The last channel in each file contained the ground truth mask, which was utilized to evaluate the proposed segmentation framework and for training of the U-Net in the additional supervised experiments. Figure 3.3 illustrates a representative microscopy slice together with its corresponding ground truth annotation and overlay.

3.3 Data Preprocessing

As the bone tissue images are grayscale, it was necessary to replicate the single channel three times to match the RGB images from ImageNet which the DINOv2 backbones have been trained on [17]. The images were thereafter normalized using the ImageNet mean and standard deviation.

All images have been normalized using percentile clipping to avoid including extremely bright or dark outliers. Moreover, clipping 5 percent of the brightest and darkest pixels proved to stabilise the results, which motivated this method choise.

All images were resized to a resolution of 980 × 980 pixels to ensure that the images can be split into an integer of patches with a fixed size of 14 × 14 pixels. Using a resolution of 980 × 980 pixels therefore makes padding or other fixes redundant while still keeping the patches comparable.

3.4 Feature Extraction Using DINO

Frozen Vision Transformers trained on the ImageNet dataset using the DINO framework have been used as feature extractors in the segmentation framework. Each image passed through the DINO backbone is split into 14 × 14 pixel patches. The feature extractor then outputs a feature embedding for each patch containing semantic and structural information about the patch. Using this information to find the patches that contain a specific structure lays the foundation for the proposed similarity based segmentation framework.

The embeddings extracted using the pretrained DINO ViT backbones are, in combination with selected reference points, used to compute similarity scores for all image patches. This enables the generation of similarity heatmaps, which can be utilized to find semantically similar features within an image.

This approach is consequently based on using pretrained self-supervised ViTs to produce feature representations where patch embeddings corresponding to patches containing the same objects form clusters in the learned feature space.

3.4.1 DINOv2 Backbone Variants

To investigate how model capacity affect the learned representations and segmentation performance, four pretrained DINOv2 Vision Transformer backbone variants were evaluated throughout this study - Small, Base, Large and Giant. These variants differs in both parameter count, the embedding dimensionality and transformer depth as described in Table 3.2 [1].

This yields different model capacities, which could enable larger versions to learn richer feature representations. As it therefore is plausible that richer representations also result in an overall higher segmentation accuracy when used in the proposed pipeline, conducting these experiments was motivated.

All evaluated DINOv2 variants use the same patch size to ensure a fixed resolution. The processed patches are thus equal for all backbone variants and the experiments run for each setting. To ensure fair comparison of the segmentation performance for the different architectures, the same seeding for sampling the reference points were used.

Table 3.2: *Overview of the evaluated pretrained DINOv2 Vision Transformer backbones [1].*

Backbone	Parameters	Embedding Dimension	Transformer Layers
Small	21 M	384	12
Base	86 M	768	12
Large	300 M	1024	24
Giant	1100 M	1536	40

3.5 Prototype Based Segmentation

A prototype based similarity segmentation framework was developed for segmentation of certain structures within the bone tissue samples. Inspiration was taken from Oxford Instrument Imaris software in combination with previous work on similarity based feature recognition using pretrained ViT backbones [11, 18].

The proposed framework is based on using positive and negative reference points that represent the structure of interest and the background or other structures respectively. The reference points are then used to locate patches containing structure and background. Passing these patches through the feature extractor outputs the prototype embeddings.

However, using randomly sampled reference points to mimic user point selection proved to be necessary to develop a robust pipeline with reproducible results and enable systematic evaluation. In order for sampling to be possible, only the samples in the dataset with corresponding ground truth masks were included in the experiments, motivating why it was necessary to exclude file B.

3.5.1 Sampling Reference Points

For each image, positive points were sampled from inside the ground truth mask, while the negative points were sampled from regions corresponding to the background. Sampling was performed randomly, but remained reproducible due to the use of seeding. Figure 3.4 gives an example of a slice and its sampled reference points.

As points too close to the boundaries caused patches not clearly belonging to the object or the background to get selected, erosion and dilation have been used. This allows for setting a limit on how close the points can be sampled to the ground truth mask boundaries. As a result, ambiguous patches containing both structure and background were selected less frequently when producing the prototype embeddings.

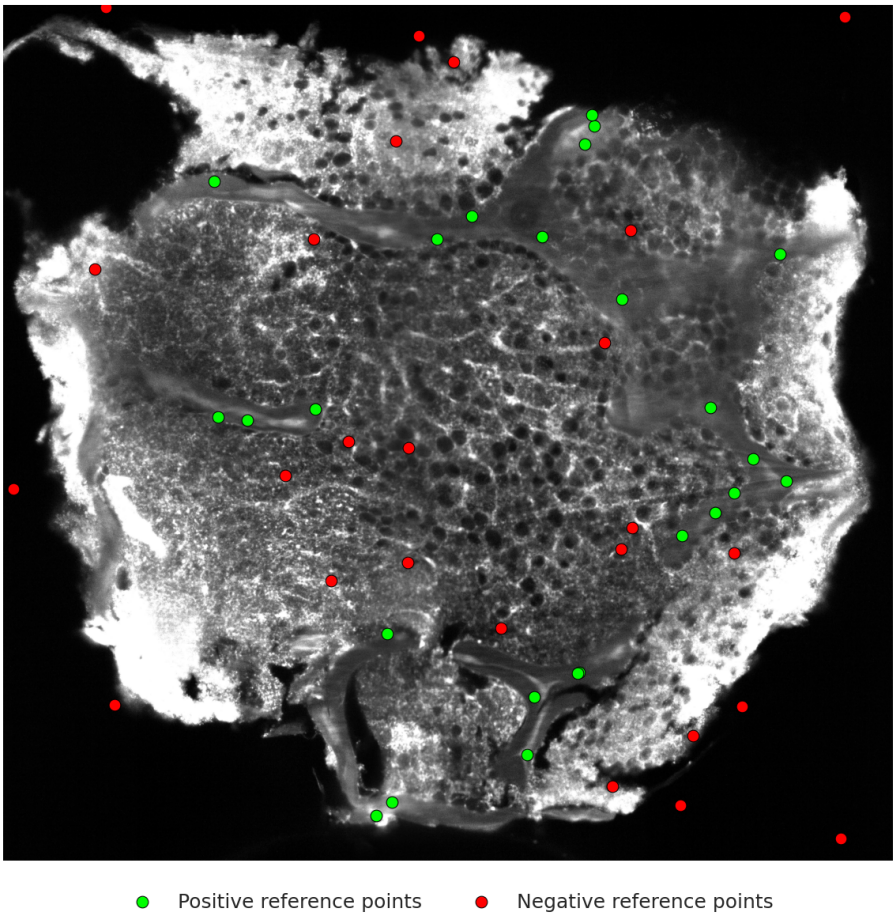

Figure 3.4: *Example of positive and negative reference point sampling. Positive reference points (green) are sampled within the ground truth mask and represent the structure of interest, while negative reference points (red) are sampled from background regions. The sampled points are used to construct positive and negative prototype embeddings for the similarity based segmentation framework.*

3.5.2 Construction of Prototype Embeddings

The sampled reference points are used to find the patches that do and does not contain the structure to segment. By mapping the reference points to the corresponding patch in the image patch grid, the reference patches and thus the prototype embeddings can be specified.

Using multiple reference points reduced sensitivity to and improved segmentation stability. Instead of relying on one single positive or negative, using several points contributed to decrease the effect of noise and poorly picked patches.

Using the mean of the selected embeddings for the positive and negative reference point patches proved to take a greater amount of the structure characteristics into account when compared to relying on individual patches to form a representative patch embedding. Mean aggregation was therefore a necessary choice for construct-

ing the positive and negative prototype embeddings respectively.

Normalization of the embeddings also constituted an important preprocessing step. The normalization ensured that the cosine computation remained meaningful by making the direction rather than magnitude of the embeddings contribute to the similarity score.

3.5.3 Producing Similarity Maps and Segmentation Masks

Utilization of how visually similar patches should generate similar feature embeddings organized in the vicinity of each other in the learned feature space lays the foundation for the developed framework. It is, more specifically, what makes the task of localizing a certain object structure turn into a task of similarity retrieval within the learned feature space.

Similarity maps for the microscopy images have been produced by computing the cosine similarity between the prototype embeddings and all corresponding image patch embeddings. High similarity to the positive prototype patches increases the overall similarity score, S, according to Equation 3.1:

$$S = S_p - \lambda S_n, \tag{3.1}$$

where S_p and S_n denote the cosine similarity between the image patch embedding and the positive protoype embedding and the negative protoype embedding respectively.

The constant λ was swept and set to 0.8, resulting in suppression of embeddings that attain high similarity to the negative prototype embedding. The combined score, denoted S in Equation 3.1, have been used to generate the similarity heatmaps. When applying thresholding in accordance to what section 3.6 Thresholding Stategy describes, binary segmentation masks were obtained from the similarity heatmaps.

3.6 Thresholding Strategy

Thresholding has been applied to obtain binary segmentation masks based on the computed similarity scores for all patches. By comparing the predicted masks with the ground truth masks, it is possible to evaluate how well the segmented regions match the structure areas. As the similarity scores vary between the images due to sample differences, using a fixed threshold was found to yield poor segmentation performance.

Instead, applying a top percentage thresholding strategy proved to be a more suitable option. Chosing a fixed percentage value to select the most similar pixels within

the heatmaps worked comparatively well for all evaluated samples. Hence, a certain percentage of the most similar patches are chosen to form the segmentation mask.

There are several advantages of using the top percentage thresholding strategy. First of all, using a top percentage value rather than a fixed value makes the pipeline less sensitive to variations in similarity score ranges for different bone samples, which contributes to increased stability in segmentation performance across samples. Furthermore, the top percentage approach proved to be robust across samples without requiring manually selected global thresholds or tuning of parameters.

However, the segmentation quality is still sensitive to the chosen top percentage value. Selecting a too high percentage can result in inclusion of false positives, while a too low percentage will exclude patches containing the structure that should be segmented. The threshold value selection will subsequently be a parameter that must be chosen carefully to ensure balanced segmentation, which is discussed more in depth in chapter 5 Discussion.

To investigate the framework senstivity to the selected percentage value, experiments examining the segmentation performance as a function of the chosen threshold have been conducted. Additionally, multiple alternative thresholding strategies were also examined during earlier stages of the framework development. As these methods failed in achieving stable performance, the top percentage thresholding strategy, which generated the most consistent segmentation performance across all evaluated samples, was chosen and used throughout this study.

3.7 U-Net Refinement

Inspired by previous work in combining learned representations using a self-supervised framework with a supervised refinement, a U-Net was introduced to seek an answer to whether the DINO segmentation masks could be improved using supervised post-processing methods [19, 20].

Introducing a U-Net implies that the framework no longer is self-supervised. Thus, the conducted experiments should be treated as an addition to the existing pipeline with the purpose of investigating how the DINO output can be improved using post-processing.

The similarity based DINO segmentation pipeline outputs both similarity heatmaps and segmentation masks. Although the thresholding strategy is optimized to work across a wide range of samples, the resulting masks generally capture most of the target structures. However, the boundary precision and the recognition of fine objects could, on the other hand, often be improved.

The purpose of the additional U-Net experiments was therefore to explore whether the information contained in the raw microscopy images together with the infor-

mation provided by the DINO heatmaps could be utilized to improve segmentation masks and boundary accuracy.

Beyond evaluating the proposed framework itself, the U-Net experiments also investigate the potential role of DINO heatmaps as complementary inputs to supervised segmentation networks. This is particularly relevant in settings where a trained U-Net already is available as the heatmaps may provide useful complementary information.

3.8 Experimental Setup

This section contains the experimental and evaluation setups that have been used for the experiments. The experiments have been constructed to enable evaluation of both the effectiveness, robustness and generalization possibilities of the proposed framework as well as the additonal U-Net experiment.

3.8.1 Dataset Split and Training Approach

The dataset provided by Sahlgrenska University Hospital consisted of six bone samples stored in IMS format. One sample, referred to as file B, was excluded from all experiments due to absence of a complete ground truth segmentation mask. Consequently, five samples (A, C, D, E and F) were included in the study.

For each sample, 100 slices were selected for evaluation, resulting in a total of 500 evaluated slices across the dataset. The selected slices were distributed evenly along the z-range in order to capture variations within each sample.

For the U-Net refinement experiments, one sample was held out to be used for validation while the remaining four samples were used for training. This procedure was repeated for all five samples, ensuring that the refinement model was evaluated on previously unseen data.

Experiments involving stochastic sampling were repeated using multiple seeds. This ensured reproducibility and enabled statistical analysis of the results.

3.8.2 Segmentation Setup

The segmentation experiments were conducted using a fixed set of parameters, which can be found in Table 3.3. The parameter referred to as λ contributes to regulate the suppressing effect of the negative prototype embeddings. By sweeping this parameter, it was found that 0.8 yielded good segmentation performance, why it was fixed at this value throughout the experiments.

The described top percentage thresholding strategy was used for constructing binary segmentation masks. Parameter sweeps for all files resulted in the usage of 10 % as the top percentage. Investigation of how the number of chosen reference points affects the segmentation performance proved that accuracy plateaus at around 25 points. 25 positive and negative reference points respectively has therefore been used for the experiments unless something else is stated.

To ensure compatibility and fair comparison for all images passed through the pre-trained ViT, the images were resized before feature extraction. 980 × 980 pixels was chosen as image resolution for all images fed to the ViT.

Table 3.3: *Parameters and corresponding values used throughout the study.*

Parameter	Value
Similarity score parameter, λ	0.8
Top percent threshold	10%
Number of positive reference points	25
Number of negative reference points	25
ViT evaluation size	980

3.8.3 U-Net Architecture and Training

To investigate how the different U-Net inputs affects the refinement, a shallow U-Net was trained using the following four setups:

- Raw image only
- DINO similarity heatmap only
- Raw image and DINO heatmap
- Raw image, DINO heatmap and the binary DINO segmentation mask

Figure 3.5 visualizes the input settings.

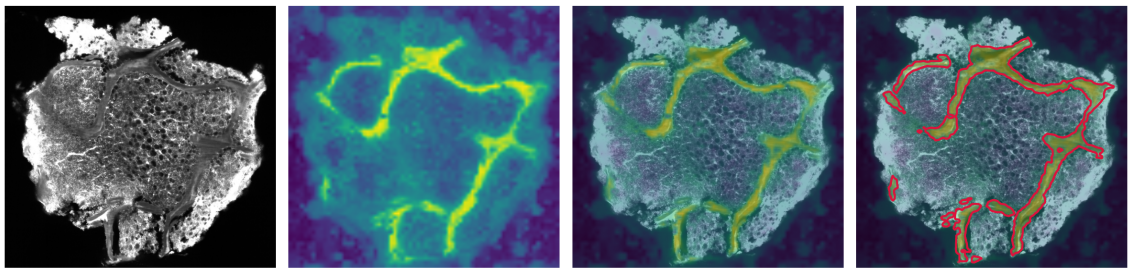

Figure 3.5: *Examples of the inputs to the U-Net. From left to right: the raw microscopy image, the corresponding DINO similarity heatmap, the heatmap overlaid on the raw image and the thresholded DINO segmentation mask and heatmap overlaid on the raw image. These setups were evaluated as inputs to the U-Net.*

As a baseline, the U-Net was trained using the raw images. However, the information about what regions belong to the structure of interest can be found in the DINO

heatmap or in the segmentation masks. Hence, they are likely to provide additional information about the segmentation masks that may help guide the U-Net training.

The shallow U-Net used for this experiment consisted of two encoder downsampling blocks and two decoder upsampling blocks with skip connections. Due to the different input configurations, the number of input channels was alternated between the runs. Table 3.4 contains the relevant information about the U-Net architecture.

Table 3.4: *Overview of the U-Net architecture used in the refinement experiments.*

Component	**Configuration**
Encoder blocks	2
Decoder blocks	2
Convolutions per block	2
Skip connections	Yes
Normalization	Batch normalization
Kernel size	3×3
Activation	ReLU
Pooling	MaxPool2d(2)
Upsampling	ConvTranspose2d(2,2)
Final convolution	1×1
Output	Logits

Inspired by previous work, a loss function combining binary cross-entropy, BCE, and dice loss, referred to as Combo Loss, was used during training [21]. BCE evaluates the pixel wise classification accuracy by compairing predicted probabilities to the ground truth labels while dice loss optimizes the overlap between the predicted and ground truth segmentation masks. Combining the two loss terms enables the network to take both local pixel classification errors and global segmentation quality into account.

This combination is particularly useful for biomedical image segmentation. While BCE help prodice accurate classification of individual pixels, the dice loss reduces the impact of class imbalance and promotes accurate segmentation of the target structures. The total loss can be expressed as

$$L = \alpha L_{\text{BCE}} + (1 - \alpha) L_{\text{Dice}}, \tag{3.2}$$

where L_{BCE} and L_{Dice} correspond to the binary cross-entropy and dice loss respectively and the constant α was set to 0.5 for equal contributions from each term.

To ensure that the resolution remained consistent for all images, the images were resized to 980×980 pixels before training. The predicted segmentation masks were then rescaled back for evaluation. Additional information about the U-Net training setup can be found in Table 3.5.

Table 3.5: *Training settings for the U-Net experiments.*

Parameter	Value
DINO backbone for heatmaps	DINOv2 Giant
Validation strategy	Leave one sample out
Training samples per split	4 samples
Training slices per sample	25
Validation samples per split	1 sample
Validation slices per sample	25
Number of seeds	3
Batch size	1
Input resolution	980×980
Learning rate	0.001
Optimizer	Adam
Maximum epochs	15
Early stopping patience	5 epochs
Threshold candidates	0.2, 0.4, 0.6, 0.8
Loss function	Combo Loss

3.8.4 Evaluation Setup

The proposed segmentation framework has been evaluated both quantitatively and qualitatively, focusing on comparing segmentation overlap and boundary accuracy, as well as the robustness and generalization capabilities across slices and bone samples. Dice, IoU, HD95, silhouette score and kNN retrieval accuracy were used as the main evaluation metrics in accordance with what is described in Chapter 2.10.

To reduce the stochastic effects that follow the sampling of reference points, the experiments were conducted using multiple seeds. The reported performance values therefore represent averaged results across several runs which enables a more accurate evaluation of the framework.

All four pretrained DINOv2 backbone variants were evaluated using the same experimental setup. In addition, experiments investigating the effect of thresholding strategy, number of reference points and prototype transfer across slices and samples were performed.

The U-Net experiments were evaluated using the same approach as the prototype based framework to enable fair comparison between the self-supervised and supervised approaches. Comparisons were conducted both quantitatively using the segmentation metrics and qualitatively through visual inspection of the resulting segmentation masks and boundary precision.

The learned feature space was evaluated qualitatively using principal component analysis, PCA, and uniform manifold approximation and projection, UMAP and quantitatively using silhouette score and kNN retrieval accuracy.

3.8.5 Generalization Across Slices and Samples

To investigate the robustness and generalization capabilities of the proposed framework, experiments for evaluating prototype transfer across slices and bone samples were conducted.

Instead of generating prototype embeddings and use these to segment only the same slice, the prototype embeddings were also used and evaluated on different slices and samples. This enabled analysis of how transferable the learned feature representations are across samples.

The motivation behind these experiments is based on how biologically similar structures should generate similar feature embeddings despite variation between slices and samples. If the learned representations are robust, prototype embeddings generated from one slice should therefore remain useful for localizing similar structures in other slices and samples.

Three experimental configurations were evaluated:

- Same Slice: Reference points and target image originate from the same microscopy slice.
- Same File: Reference points are sampled from one slice while segmentation is performed on different slices from the same bone sample.
- Cross File: Reference points are sampled from one bone sample while segmentation is performed on slices from another sample.

These experiments enable evaluation of how well the learned feature representations generalize across different slices and biological samples.

4
Results

This chapter presents the results of the proposed framework and is divided into four parts - representation quality analysis, segmentation performance evaluation, robustness analysis and generalization analysis. Both quantitative metrics and qualitative visualizations are presented for investigation of the learned feature space, the segmentation performance of the framework and the transferability of the learned representations across different microscopy slices and bone samples.

4.1 Representation Quality Analysis

This section presents the analysis of the learned feature representations produced by the pretrained DINOv2 backbones.

4.1.1 Feature Space Visualization

Figures 4.1 and 4.2 present representative PCA and UMAP visualizations of the organization of patch embeddings extracted using the DINOv2 Small and Giant backbones respectively. Structure and background patches are shown using orange and blue dots, while positive and negative prototype embeddings are indicated by green and red markers.

Structure and background patches are distributed across different regions of the feature space for each backbone variant. The prototype embeddings are positioned near the patch groups they represent. Note the differences in cluster compactness and overlap that can be seen between the two backbones.

4.1.2 Clustering Representation Analysis

Feature separability was quantified using clustering metrics together with k-nearest neighbor retrieval accuracy. Table 4.1 contains the silhouette scores and retrieval accuracy for the DINOv2 backbone variants.

The Large and Giant backbones achieved nearly identical silhouette scores of 0.100 while the retrieval accuracy ranged from 0.958 to 0.960.

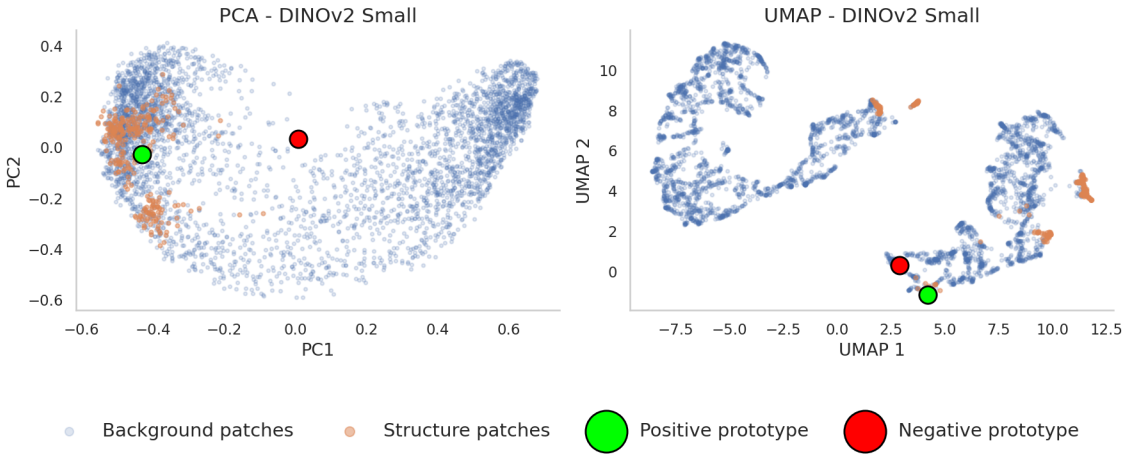

Figure 4.1: *Representative PCA and UMAP visualizations of patch embeddings extracted using the DINOv2 Small backbone. Structure and background patches are shown by orange and blue dots respectively, while positive and negative prototype embeddings are marked by green and red circles.*

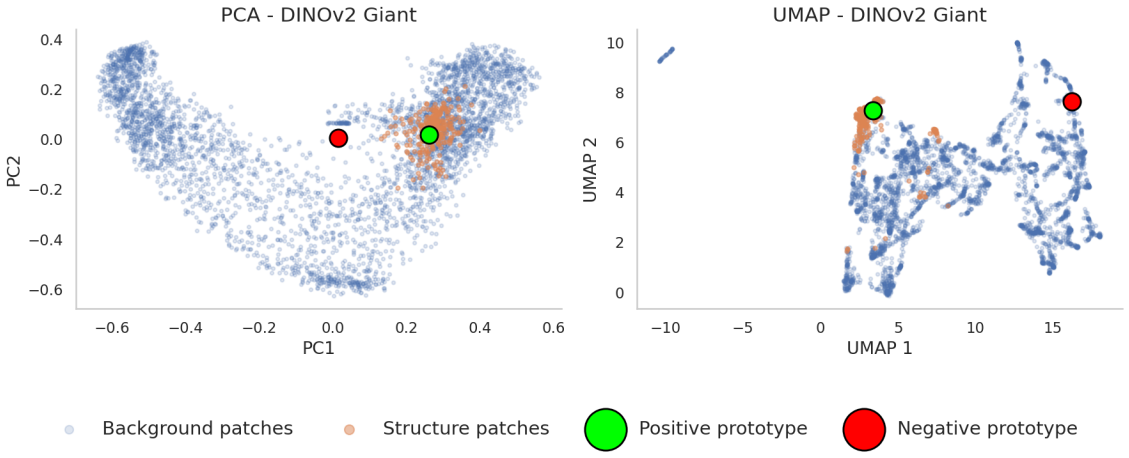

Figure 4.2: *Representative PCA and UMAP visualizations of patch embeddings extracted using the DINOv2 Giant backbone. Structure and background patches are shown by orange and blue dots respectively, while positive and negative prototype embeddings are marked by green and red circles.*

4.2 Segmentation Performance

This section presents the segmentation performance of the proposed framework.

4.2.1 Backbone Size Comparison

Table 4.2 summarizes the segmentation performance for all DINOv2 backbones in terms of dice, IoU and HD95 across all microscopy slices. In general, the larger backbone variants achieved higher dice and IoU scores together with lower HD95 values.

Table 4.1: *Feature space separability and retrieval accuracy for the evaluated backbone variants. The silhouette score quantifies the global separation between structure and background embeddings while the kNN retrieval accuracy measures the local feature space consistency. Values are reported as mean ± standard deviation across the used bone slices.*

Backbone	Silhouette	kNN Retrieval Accuracy (k=10)
Small	0.070 ± 0.072	0.959 ± 0.013
Base	0.066 ± 0.060	**0.960 ± 0.013**
Large	**0.100 ± 0.042**	**0.960 ± 0.013**
Giant	**0.100 ± 0.032**	0.958 ± 0.014

Table 4.2: *Segmentation performance for the evaluated pretrained DINOv2 backbone variants. Values are reported as mean ± standard deviation across the evaluated slices.*

Backbone	Dice	IoU	HD95
Small	0.664 ± 0.085	0.503 ± 0.092	180.37 ± 129.67
Base	0.664 ± 0.080	0.502 ± 0.086	150.30 ± 103.09
Large	0.685 ± 0.090	0.527 ± 0.101	129.28 ± 106.45
Giant	**0.690 ± 0.093**	**0.534 ± 0.105**	**108.62 ± 105.94**

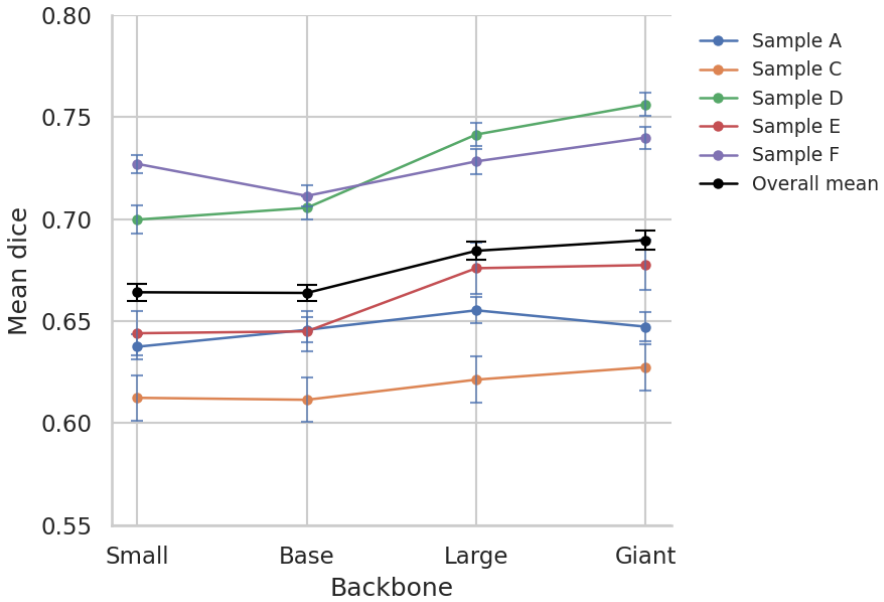

Figure 4.3: *Mean dice score for each sample and DINOv2 backbone variant. Results are averaged across the evaluated slices for each sample.*

The Giant backbone achieved the highest dice score, 0.690 ± 0.093, and the high-

est IoU score, 0.534 ± 0.105. It also achieved the lowest HD95 value, 108.62 ± 105.94, indicating the best overall segmentation performance among the evaluated DINOv2 backbone variants. The Large backbone achieved similar overlap performance, closely followed by the Small and Base backbones.

Figure 4.3 shows the mean dice score for each sample and backbone variant. Performance differences between backbone variants varied across samples. The Giant backbone achieved the highest mean dice score in several of the evaluated samples, while smaller differences were observed for others.

4.2.2 Qualitative Segmentation Examples

Representative examples corresponding to the worst, median and best performing slices for all samples can be found in Figure 4.4. Additional qualitative figures have been provided in Appendix A.

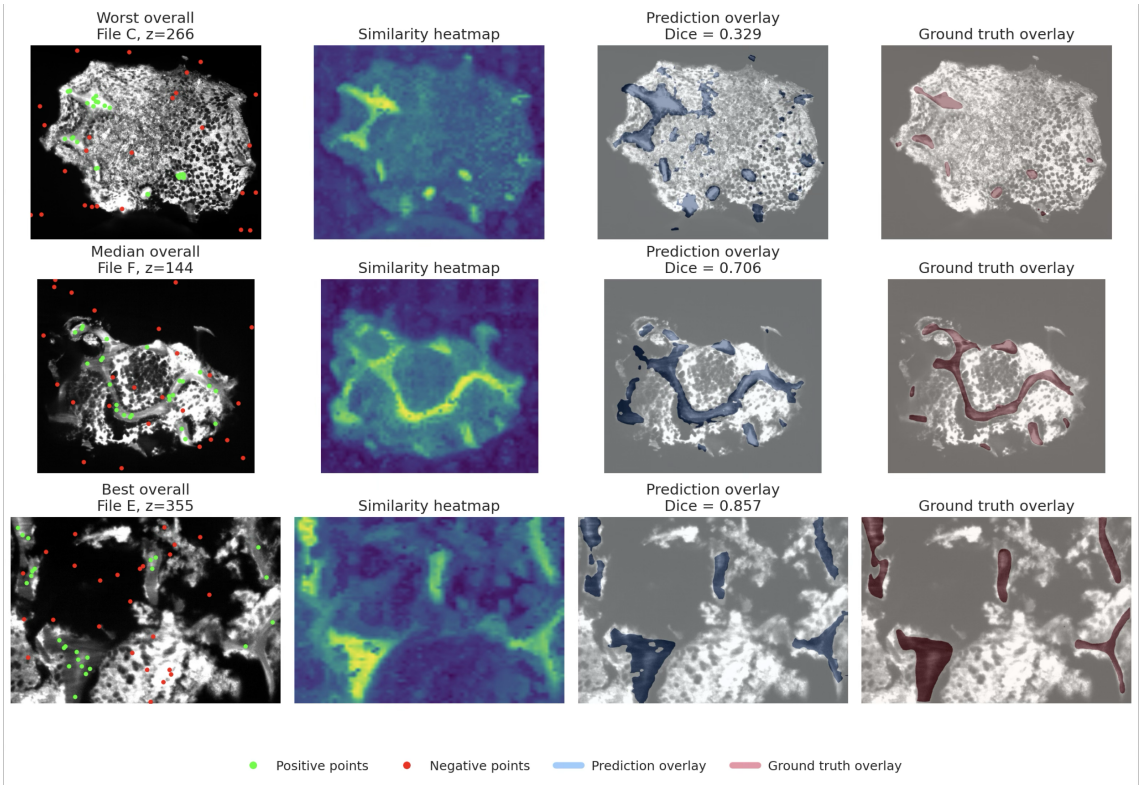

Figure 4.4: *Qualitative segmentation examples corresponding to the worst, median and best performing slices. The first column shows the microscopy images together with the sampled positive and negative reference points. The second column shows the generated similarity heatmaps, while the third column and forth columns visualizes the predicted segmentation (blue) and the ground truth annotation (red).*

4.2.3 U-Net Refinement

The U-Net experiment included four different input configurations. The evaluated configurations and their corresponding segmentation performance can be found in Table 4.3. Note that the Small backbone has been used due to computational limitations.

Table 4.3: *Segmentation performance for the evaluated U-Net input configurations. Values are reported as mean performance across all evaluated samples.*

Input Configuration	Dice	HD95
DINOv2 Small	0.627	252.5
Raw image	0.602	213.5
Heatmap	0.757	**143.3**
Raw image + Heatmap	**0.772**	154.6
Raw image + Heatmap + Mask	0.752	156.4

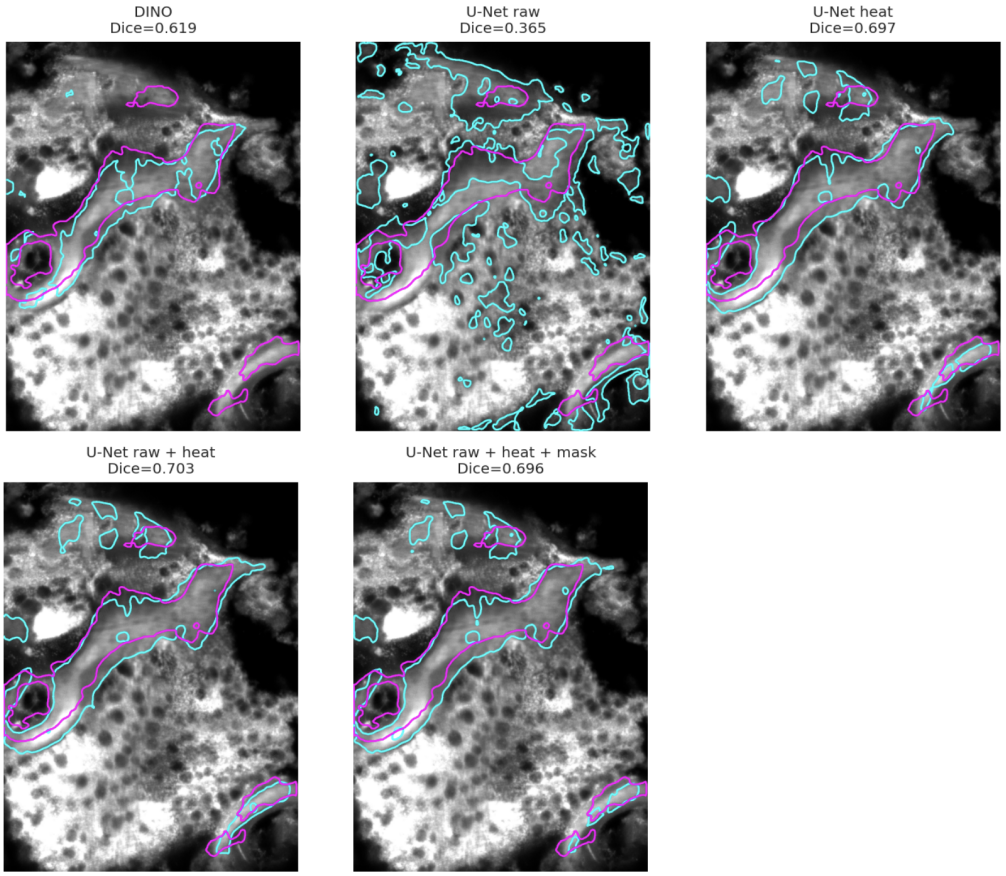

Figure 4.5: *Representative comparison between the prototype based DINO segmentation framework and the evaluated U-Net refinements. The figure illustrates how different U-Net inputs influence the resulting segmentation masks shown in cyan relative to the ground truth annotation marked in magenta.*

As seen in Table 4.3, the highest dice score was obtained using the raw image together with the DINO heatmap as U-Net input, while the lowest HD95 value was achieved using the heatmap configuration. The raw image as the input alone achieved lower segmentation performance across all evaluated metrics.

Examples of the resulting segmentation masks for the evaluated configurations are shown in Figure 4.5. Additional qualitative examples of the U-Net refinement results are provided in Appendix B, where the worst, median and best performing slices for each sample are shown.

4.3 Robustness Analysis

This section presents the framework sensitivity to the reference point sampling and threshold selection.

4.3.1 Point Sensitivity

Figure 4.6 presents the distribution of dice scores across different numbers of sampled points. As can be seen in the figure, dice increased with the number of sampled reference points before approaching a plateau. The largest performance changes were observed at lower numbers of sampled points, while smaller changes were observed beyond approximately 20–25 points.

4.3.2 Threshold Sensitivity

Threshold sensitivity was analyzed by measuring the resulting dice for different threshold values used on the heatmaps. Figure 4.7 shows the dice scores as a function of threshold value for the evaluated samples.

The highest mean dice score was obtained using a threshold of 10%. Similar performance was observed for thresholds between 8% and 14%, which indicates that the framework is relatively robust to small variations in threshold selection. Substantially lower thresholds did, however, result in decreased segmentation performance as can be seen in Figure 4.7.

4.4 Generalization Analysis

This section presents the results from the experiments investigating the transferability and generalization capabilities of the learned feature representations. Detailed pairwise transfer results between all microscopy samples are provided in Appendix B.

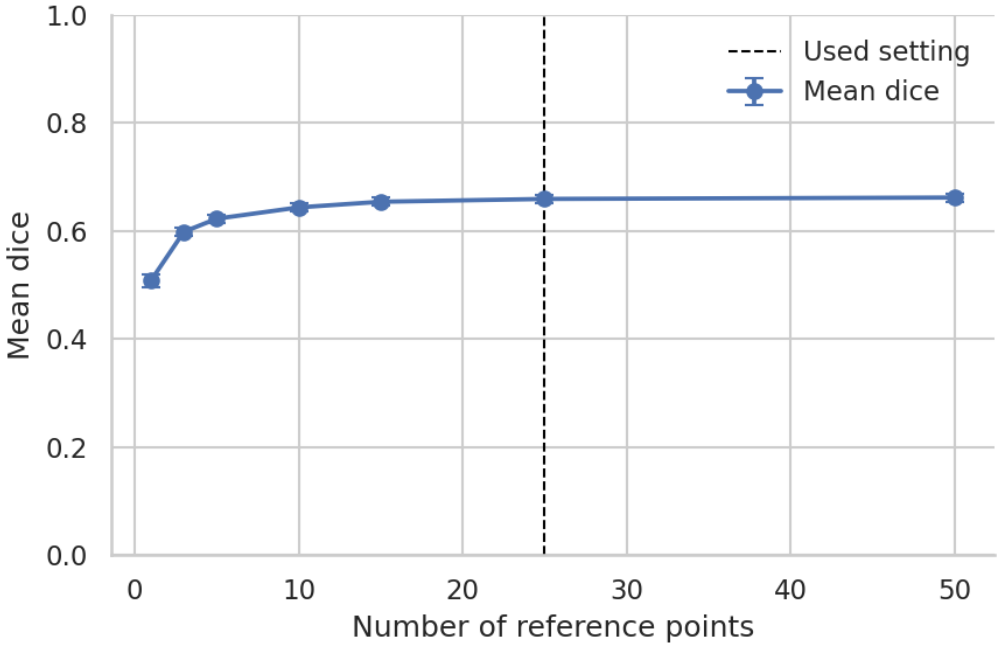

Figure 4.6: *Point sensitivity plot illustrating the mean dice score as a function of the number of positive and negative reference points used for prototype construction.*

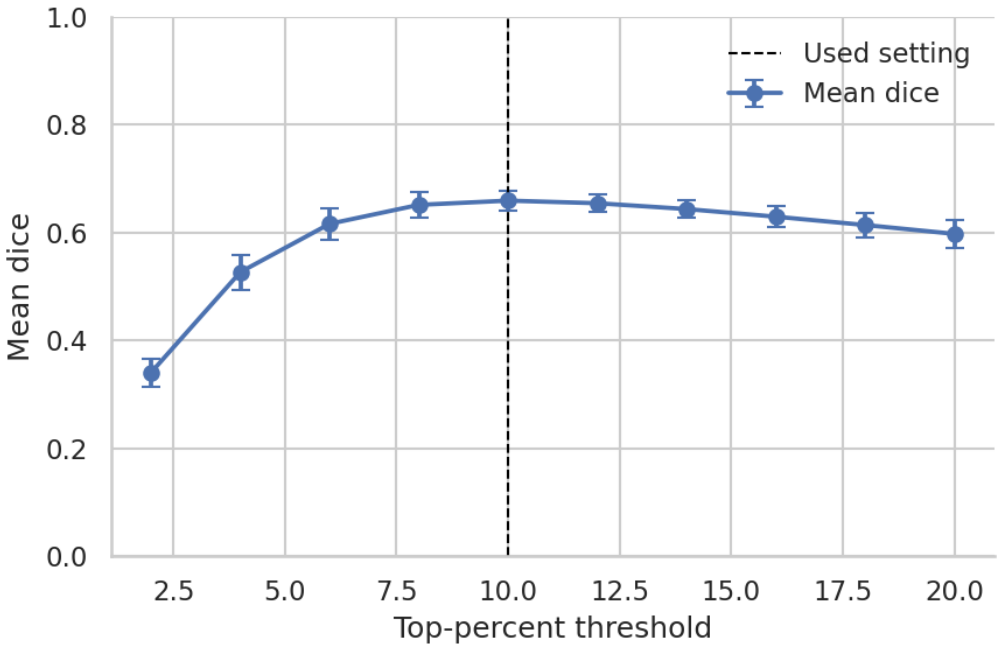

Figure 4.7: *Threshold sensitivity plot showing the mean dice score as a function of threshold value for the evaluated samples illustrating how segmentation performance changes as different threshold values are used for generating segmentation masks.*

4.4.1 Same File Prototype Generalization

The first generalization experiment included extracting prototype embeddings from one slice to then apply them to other slices from the same file.

Figure 4.8 presents a representative example of prototype transfer between slices within the same bone sample, while Table 4.4 summarizes the segmentation performance for the evaluated transfer settings.

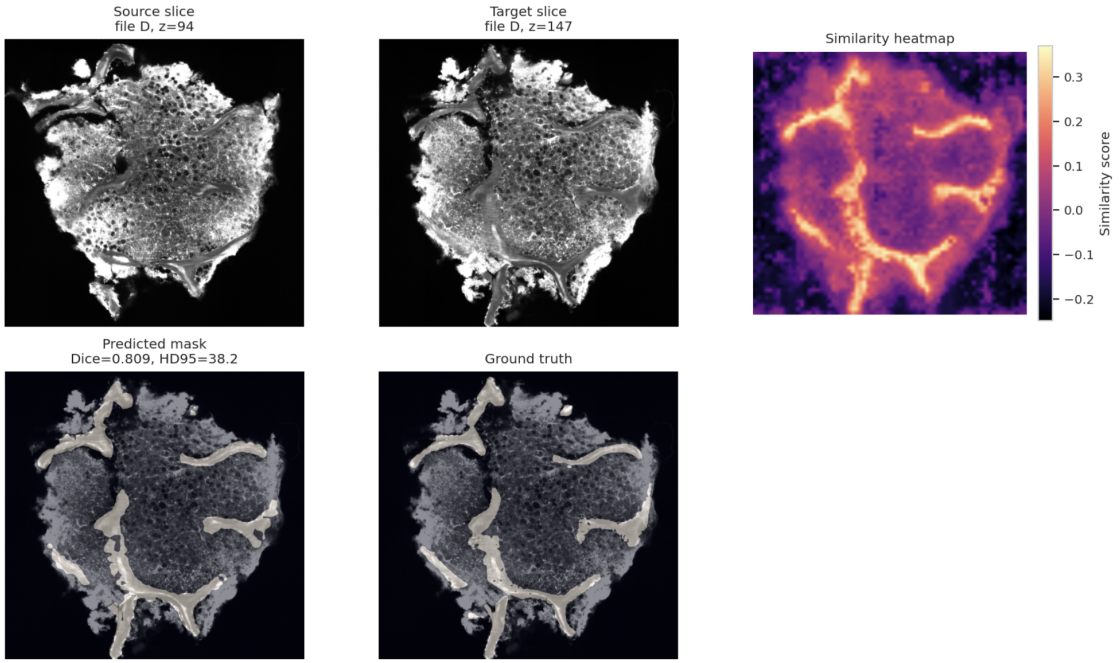

Figure 4.8: *Representative example of prototype transfer within the same sample. Prototype embeddings extracted from one bone sample were applied to a slice from a the same sample. The figure content from top left to right: the microscopy image which the reference points were picked from, the target slice, the cosine similarity heatmap for the target image, the predicted mask and the corresponding ground truth mask.*

Table 4.4: *Prototype transfer performance for the evaluated transfer settings. Results are reported as mean ± standard deviation for prototype transfer within the same slice, within the same sample and across different bone samples.*

Transfer Setting	**Dice**	**IoU**
Same slice	**0.647 ± 0.091**	**0.485 ± 0.097**
Same sample	0.644 ± 0.100	0.483 ± 0.109
Different sample	0.542 ± 0.141	0.384 ± 0.129

As can be seen in Table 4.4, prototype transfer within the same sample resulted in segmentation performance comparable to what was obtained when prototypes were taken from the evaluated slice. The largest performance decrease was observed for prototype transfer between different samples.

4.4.2 Cross File Prototype Generalization

The second generalization experiment included extracting prototype embeddings from one slice and then applying them to slices from different bone samples.

Compared to prototype transfer within the same sample, cross sample transfer resulted in lower dice and IoU scores together with increased performance variability, as shown in Table 4.4.

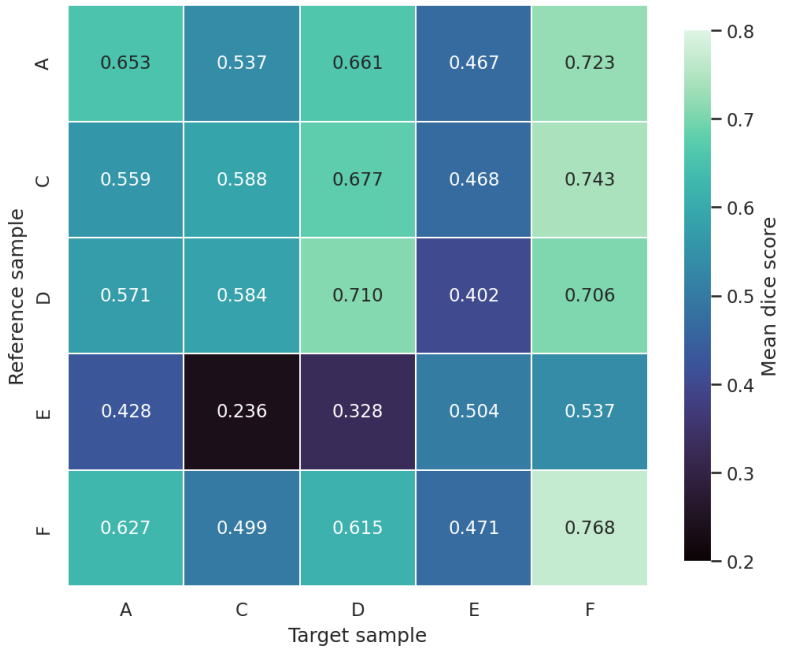

Figure 4.9: *Prototype transferability matrix showing the mean dice score when prototype embeddings extracted from a reference sample were used to segment images from a target sample. Rows correspond to the reference sample and columns correspond to the target sample.*

The transferability matrix shown in Figure 4.9 provides a more detailed overview of the cross sample generalization performance. The highest dice scores are generally observed along the diagonal, which corresponds to transfers within the same sample. However, several non diagonal entries also achieve high accuracy.

A representative example of cross file prototype transfer is shown in Figure 4.10. Although meaningful segmentation still can be achieved, the resulting segmentation mask deviates more from the ground truth compared to a within sample transfer.

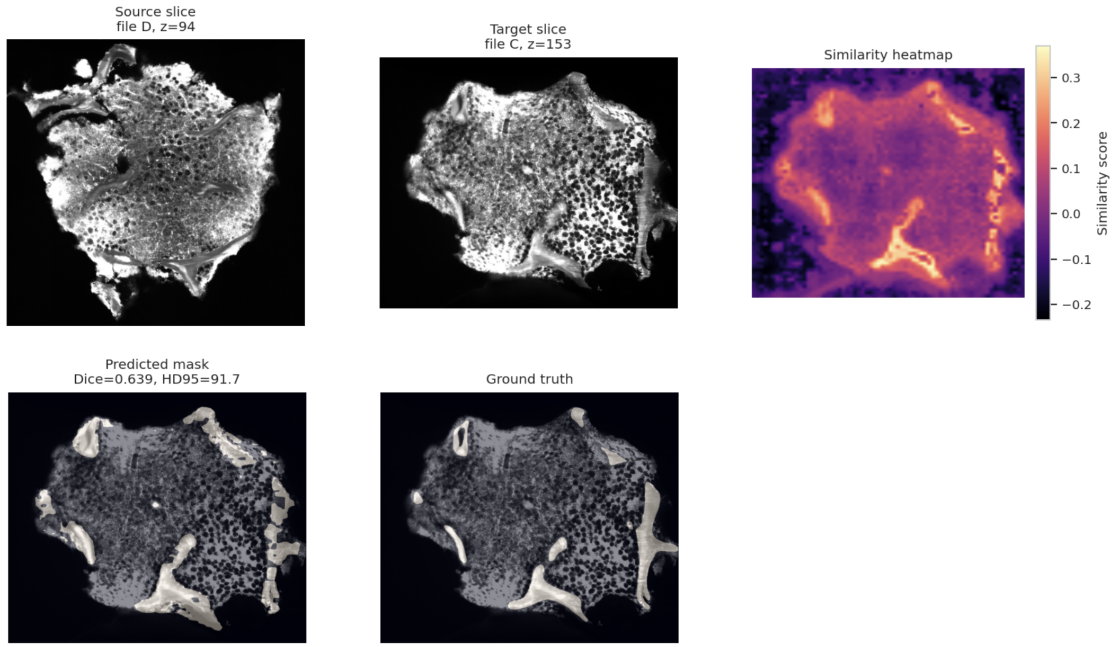

Figure 4.10: *Representative example of cross file prototype transfer. Prototype embeddings extracted from one bone sample were used on a slice from a different sample. The figure shows the raw microscopy images, the generated similarity heatmap, the predicted segmentation mask and the corresponding ground truth mask.*

5

Discussion

5.1 Learned Feature Representations

This section discusses the findings related to the organization of the learned feature space and their relation to the segmentation performance.

5.1.1 Feature Space Organization

PCA and UMAP visualizations show that structure and background patches tend to form partially separable clusters within the learned feature space. The positive and negative prototype embeddings are observed to be placed within the corresponding regions indicating that the sampled prototypes capture representative features of each patch type.

The qualitative observations are also supported by the quantitative evaluation. Positive silhouette scores were obtained for all evaluated backbone variants, ranging from 0.066 for the Base backbone to 0.100 for the Large backbone. Retrieval accuracy remained consistently high and ranged from 0.958 to 0.960 across all evaluated backbones, which indicates that structure and background embeddings are partially separated within the latent space and how neighboring embeddings tend to contain similar structures, suggesting an accurate representation organization.

Even though the structure and background embeddings are not completely separable, the learned feature space organization appears to be sufficient for similarity based segmentation. Partial overlap between structure and background clusters is, in fact, expected as a result of biological tissue being heterogeneous. Many patches will contain several object types and as structural boundaries can be gradual, many transitions between structures will be ambiguous even for an expert. This is further supported by the relatively moderate silhouette values observed across all backbone variants.

The Large and Giant backbones achieved the highest silhouette scores among the evaluated models. Qualitatively, the PCA and UMAP visualizations suggest that the Giant backbone produces more compact clusters and less visible overlap between structure and background patches than the Small backbone.

5.1.2 Relation Between Representation Quality and Segmentation Performance

Backbone variants with higher silhouette scores generally achieved higher dice and IoU scores. For example, the Large and Giant backbones achieved silhouette scores of 0.100 respectively, while also producing the highest dice scores of 0.685 and 0.690, suggesting that improved feature space separability could contribute to improved segmentation performance.

The segmentation performance remained relatively strong despite the moderate silhouette scores. This suggests that complete separation of the structure and background within the learned feature space is not required for obtaining successful prototype based segmentation. Instead, partial semantic organization and local feature consistency appear to be sufficient for producing meaningful similarity based segmentation masks.

5.2 Prototype Based Segmentation Performance Analysis

This section discusses the segmentation performance obtained using the proposed prototype based segmentation framework and the factors influencing the resulting segmentation quality.

5.2.1 Influence of Backbone Capacity

Segmentation performance generally improved with increasing backbone capacity. The Giant backbone achieved the highest dice, IoU and the lowest HD95 and performed best overall when compared to the other backbones.

The improvement was, however, not equal across evaluation metrics. While the dice score increased from 0.664 for the Small backbone to 0.690 for the Giant backbone - corresponding to an improvement of approximately 3.9 % - the reduction in HD95 was larger and decreased from 179.3 to 109.7, which corresponds to a reduction of approximately 39%. This suggests that increasing backbone capacity primarily improves boundary localization rather than overlap accuracy and may indicate that useful feature representations are already learned by the smaller DINOv2 variants.

A possible explanation is that larger backbones are capable of capturing more complex relationships between image patches and therefore produce more informative feature representations. This is further supported by the representation quality analysis where the Large and Giant backbones achieved the highest silhouette scores and most organized feature spaces.

5.2.2 Dataset Dependent Behavior

Segmentation performance varied across the evaluated microscopy datasets, as shown in Figure 4.3. Although larger DINOv2 backbone variants generally achieved improved segmentation performance, the magnitude of the improvement depended on the individual dataset.

A trend where increasing backbone size improved the segmentation performance could be observed for datasets C, D and F. Dataset A showed a somewhat different behavior. While the Large backbone achieved the highest dice score, the Giant backbone produced the lowest HD95 value, indicating that the effect of increasing backbone size depended on the properties of the dataset.

Dataset E showed the smallest performance differences between the backbone variants. The similar performance across all backbone variants suggests that factors other than representation quality have contributed to limit the segmentation performance for this dataset. Potential explanations include low image contrast, more heterogeneous tissue or less distinct separation of structure and background - which are the main differences that can be distinguished visually between sample E and the rest of the data files.

5.2.3 U-Net Refinement

The results indicate that U-Net input using the DINOv2 Small backbone outputs consistently outperformed the raw image input configuration. The best performance was obtained using the raw image together with the heatmap as the input, which increased the mean dice score from 0.627 to 0.772 while reducing the mean HD95 from 252.5 to 154.6.

Comparing the raw image U-Net input with the configurations using DINO outputs suggests that much of the performance gain originates from the information within the DINO representations rather than from the U-Net architecture itself. By providing an initial localization of the relevant structures the heatmaps appear to guide the U-Net towards more accurate segmentation and improved boundary localization.

5.3 Generalization and Domain Shift

This section discusses the transferability of the learned feature representations within and between microscopy samples.

5.3.1 Same Sample Prototype Transfer

The results demonstrate that prototype embeddings taken from one microscopy slice successfully can be transferred to neighboring slices within the same bone sample while maintaining segmentation performance comparable to what was achieved when prototypes were extracted directly from the evaluated slice. As seen in Table 4.4, mean dice score decreased only marginally from 0.647 for same slice evaluation to 0.644 when prototypes were transferred to other slices from the same sample.

This strong transfer performance suggests that the proposed framework is not dependent on the embeddings from the slice the prototypes are extracted from. The prototype embeddings instead seem to capture more general characteristics of the structures which remain consistent across slices within a sample. As seen from a practical perspective, this means that prototype embeddings could be reused across a whole sample. This could in turn enable semi-automatic segmentation of an entire sample after selecting reference points from only one slice.

5.3.2 Cross Sample Prototype Transfer

The segmentation performance decreased significantly when prototype embeddings were transferred between different bone samples. The mean dice score decreased from 0.644 for transfer within a sample to 0.542 for cross sample transfer, while the standard deviation increased from 0.100 to 0.141. Although the generated similarity heatmaps often remained capable of partially localizing the target structures, the resulting segmentation masks included more false positive areas as well as less accurate boundaries.

These results suggest that the learned feature representations remain sensitive to variations between microscopy samples. Differences in tissue morphology, image contrast and intensity may contribute to make transitions between structure and background regions more gradual and thereby affecting the segmentation performance. Hence, good performance within a sample does not guarantee equally strong performance when the same prototype embeddings are applied to another sample.

5.4 Limitations and Future Work

The proposed framework has several limitations that need to be acknowledged. First, the segmentation pipeline relies on manually selected positive and negative reference points together with a thresholding strategy optimized on only five samples. Even though the robustness experiments showed relatively stable performance across a range of threshold values and number of reference points, the segmentation accuracy remained sensitive to prototype selection and threshold choice.

The top-percentage strategy of generating masks comes with a particularly important limitation. Since a fixed percentage of the most similar patches always get

selected, the predicted mask size becomes dependent on the chosen threshold value. While this provided stable segmentation performance across the evaluated samples it is likely to become problematic when the image proportions of the target structure varies. In such cases, the framework will over or underestimate the segmented area. This will result in limited applicability for measurements such as tissue area or volume estimation. Future work should therefore investigate adaptive thresholding strategies that can better account for variations in structure size and image content.

Another limitation became evident in the cross sample transfer experiments where the performance decreased substantially. While prototype transfer within the same microscopy sample showed stable performance, transferring prototypes between different samples resulted in a reduced segmentation accuracy. These findings suggests that the learned representations are sensitive to variations in tissue morphology as well as intensity and contrast distributions. The current framework is thus not completely able to handle domain shifts between samples, which limits the transferability of prototype embeddings across different samples.

An additional limitation originates from the image preprocessing procedure. All microscopy images were resized to 980×980 pixels before feature extraction. While this ensured a consistent patch grid across all experiments it also implies that pixels and the DINO patches correspond to different tissue areas in the samples due to the different original image resolutions. Structures with similar biological size may therefore be represented at different scales for the different samples, which could contribute to reducing the performance in the cross sample transfer experiments. This should be investigated in future work and scale aware preprocessing or a preserved image resolution could be considered to enhance the robustness of prototype transfer across samples.

All experiments were conducted independently on two-dimensional microscopy slices. Spatial relationships between neighboring slices and three-dimensional structures were therefore not taken into account. Utilizing the volumetric information could potentially improve both segmentation consistency and robustness should be considered in future work.

Furthermore, the evaluation was conducted on a limited number of tissue samples. Additional evaluation on larger and more diverse datasets is needed to determine how well the framework generalizes beyond the datasets used in this study.

The U-Net refinement experiments also introduce a supervised component to the framework. While the refinement improved segmentation performance, it requires ground truth annotations during training and therefore reduces one of the primary advantages of the proposed self-supervised approach. The refinement results should therefore be viewed as an additional experiment rather than a part of the self-supervised segmentation pipeline.

A deeper U-Net architecture could potentially improve refinement performance by

increasing the model capacity. However, as the annotated samples are limited, a deeper model might also increase the risk of overfitting - which should be investgated in future work.

Automated prototype selection and adaptive thresholding are also examples of directions for further work. This could reduce the needed user interaction while domain adaptation techniques could help improve robustness across different samples. Since the DINO heatmaps proved useful for U-Net refinement, exploring more advanced refinement architectures could also be a next step.

Despite the mentioned limitations, the results indicate that self-supervised Vision Transformers pretrained using the DINOv2 framework can be utilized for microscopy image segmentation with minimal user interaction - despite being trained on natural RGB images rather than microscopy data. These findings therefore highlight the potential of prototype based segmentation using pretrained self-supervised representations for biomedical imaging applications.

6
Conclusion

6.1 Summary of Findings

This thesis investigated the use of pretrained self-supervised DINOv2 Vision Transformer backbones for prototype based segmentation of bone tissue microscopy images. The proposed framework utilizes similarity between image patch embeddings and prototype embeddings to generate segmentation masks without requiring supervised end-to-end segmentation training.

The results demonstrated that the pretrained DINOv2 backbones learn feature embeddings in which tissue structures become partially separable within the learned embedding space. PCA and UMAP visualizations also showed, together with clustering metrics, that structure and background patches tend to occupy different regions of the feature space. This lays the foundation for the framework and enables the similarity based segmentation.

The segmentation experiments further showed that the proposed prototype based framework was able to localize structures using only a small number of sampled reference points. The larger DINOv2 backbone variants generally achieved better segmentation performance, both in terms of overlap metrics and boundary accuracy.

The similarity heatmaps were found to contain information that could be utilized by a supervised refinement model to improve the segmentation masks. Combining the heatmaps with the raw images as input to a U-Net refinement step improved segmentation accuracy - particularly for the most challenging samples.

Prototype embeddings could be reused across slices within the same sample while maintaining similar segmentation performance. However, reduced performance was observed during cross sample transfers which also highlights the framework sensitivity to domain shift and sample variations.

6.2 Main Contributions

The main contributions of this thesis can be summarized as follows:

- A prototype based segmentation framework utilizing pretrained self-supervised DINOv2 Vision Transformer backbones for bone tissue microscopy image seg-

mentation was developed and evaluated.
- The learned DINOv2 feature spaces were analyzed using PCA, UMAP and clustering metrics which showed that tissue and background patches tend to occupy different regions of the feature space.
- A comparative evaluation of the four DINOv2 backbone variants was conducted to investigate the effect of model capacity on representation quality and segmentation performance.
- A supervised U-Net refinement step using the DINO outputs was introduced and shown to improve segmentation accuracy and boundary localization.
- Prototype transfer was evaluated both within the same sample and across different samples to investigate generalization and sensitivity to domain shift.

6.3 Final Remarks

The results of this thesis demonstrate that self-supervised Vision Transformers pretrained using the DINOv2 framework successfully can be used for segmenting microscopy images requiring minimal manual intervention. Utilizing feature space similarity in combination with prototype embeddings enabled meaningful segmentation of the bone tissue images, although the models were trained using RBG images rather than microscopy data.

Bibliography

[1] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," 2024. [Online]. Available: https://arxiv.org/abs/2304.07193

[2] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," 2020. [Online]. Available: https://arxiv.org/abs/1908.10454

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," 2014. [Online]. Available: https://arxiv.org/abs/1206.5538

[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised Vision Transformers," 2021. [Online]. Available: https://arxiv.org/abs/2104.14294

[5] V. Bundele, K. Sarıtaş, B. Kargi, O. A. Çal, K. Tezören, Z. Ghaderi, and H. Lensch, "Evaluating self-supervised learning in medical imaging: A benchmark for robustness, generalizability, and multi-domain impact," 2025. [Online]. Available: https://arxiv.org/abs/2412.19124

[6] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, and M. Norouzi, "Big self-supervised models advance medical image classification," 2021. [Online]. Available: https://arxiv.org/abs/2101.05224

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[9] B. Kundu, B. Khanal, R. Simon, and C. A. Linte, "Assessing the performance of the DINOv2 self-supervised learning Vision Transformer model for the segmentation of the left atrium from mri images," 2024. [Online]. Available: https://arxiv.org/abs/2411.09598

[10] N. Cavagnero, G. Rosi, C. Cuttano, F. Pistilli, M. Ciccone, G. Averta, and F. Cermelli, "Pem: Prototype-based efficient maskformer for image segmentation," 2024. [Online]. Available: https://arxiv.org/abs/2402.19422

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023. [Online]. Available: https://arxiv.org/abs/2304.02643

[12] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[13] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[14] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[16] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, pp. 1–28, 2015.

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

[18] A. González-Marfil, E. Gómez-de Mariscal, and I. Arganda-Carreras, "DINOSim: Zero-shot object detection and semantic segmentation on microscopy images," *bioRxiv*, 2025.

[19] Y. Lu, Y. Wu, R. Kateb, and A. Chaddad, "Semi-supervised medical image segmentation via dual networks," 2025. [Online]. Available: https://arxiv.org/abs/2505.17690

[20] H. Khan, W. Chen, and M. K. K. Niazi, "Weakly supervised teacher-student framework with progressive pseudo-mask refinement for gland segmentation," 2026. [Online]. Available: https://arxiv.org/abs/2603.08605

[21] S. A. Taghanaki, Y. Zheng, S. K. Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh, "Combo loss: Handling input and output imbalance in multi-organ segmentation," 2021. [Online]. Available: https://arxiv.org/abs/1805.02798

A

Appendix 1

The examples presented in this appendix complement the quantitative results reported in the main text.

A.1 Additional Segmentation Results

Figure A.1 presents the worst, median and best performing evaluated slices from each sample ranked according to dice score.

A.2 Additional U-Net Refinement Results

The worst, median and best U-Net refinement cases from each sample are shown in Figure A.2. The ranking was based on the U-Net dice improvement compared to the DINO baseline segmentation. Each example shows the original DINO prediction, the refined U-Net prediction and the corresponding ground truth mask.

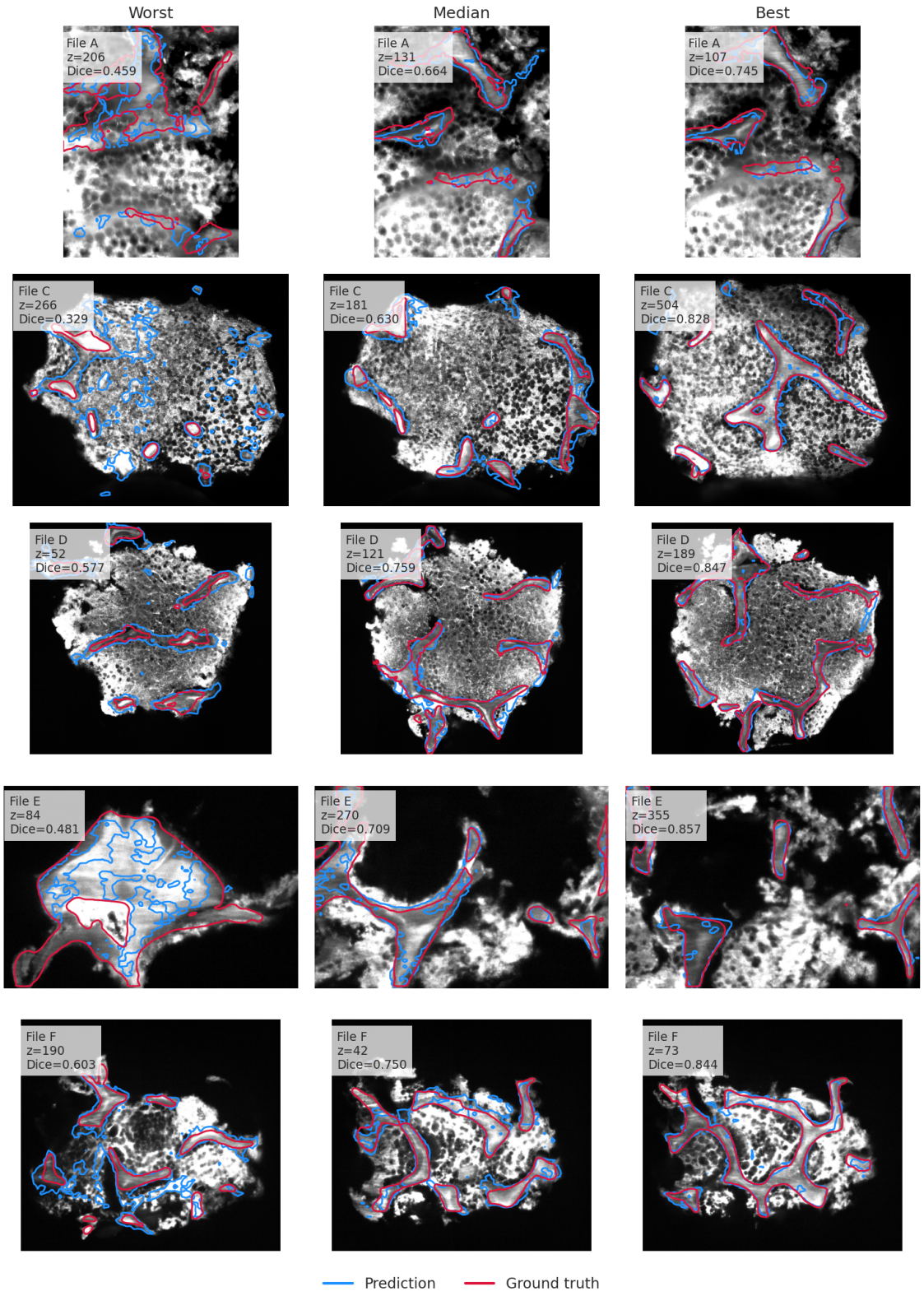

Figure A.1: *Additional qualitative segmentation examples for all samples. The worst, median and best performing slices based on the dice score are shown for each sample. Predicted segmentation masks are marked in blue and ground truth masks in red.*

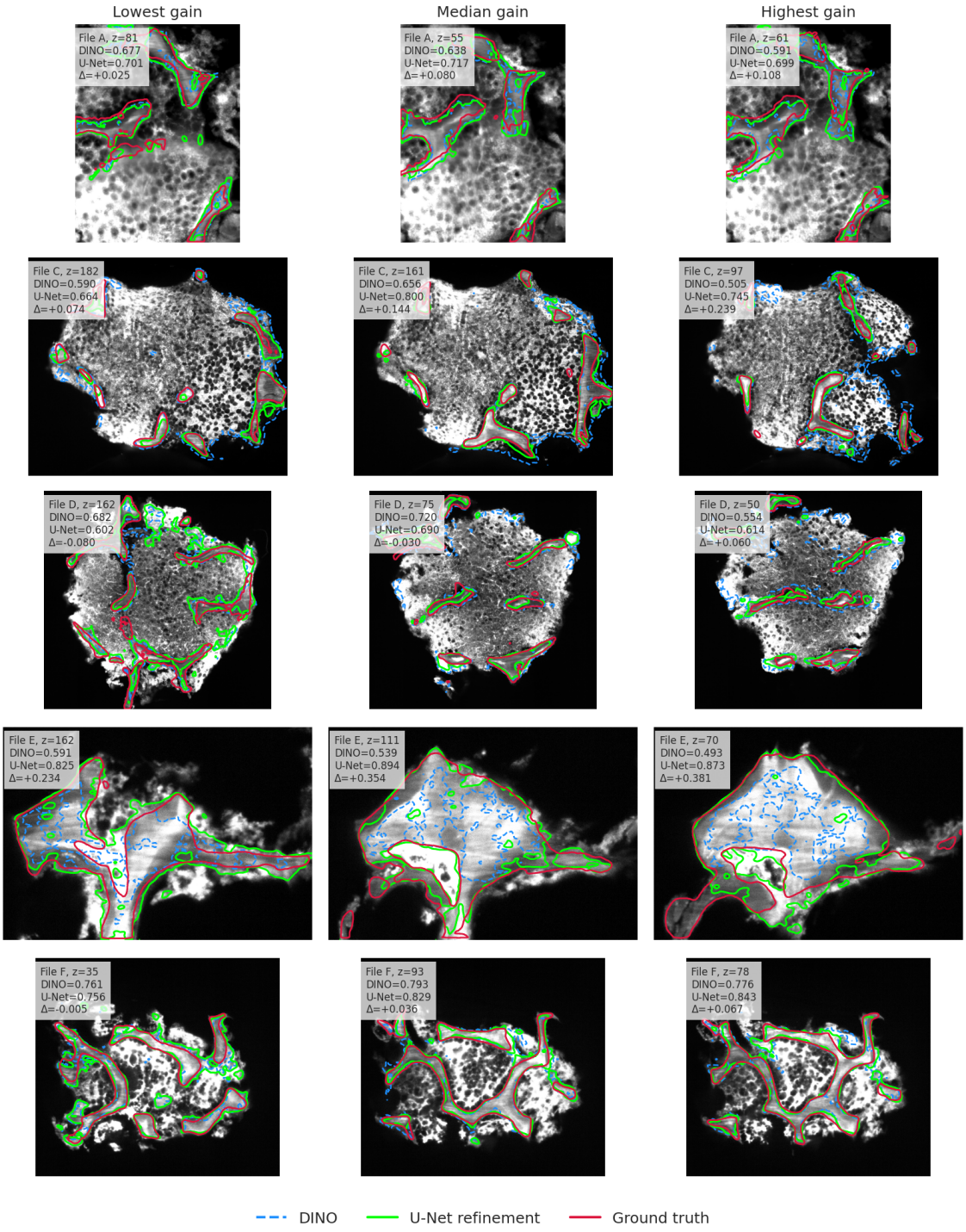

Figure A.2: *Qualitative U-Net refinement results for the worst, median and best performing slices from each sample based on the dice improvement. Blue dashed contours correspond to the original DINO prediction, lime denote the refined U-Net prediction and red corresponds to the ground truth mask. The shown values are the dice scores before and after refinement together with the dice improvement.*

B

Appendix 2

B.1 Additional Prototype Transfer Results

This appendix provides the complete prototype transfer results underlying the transferability analysis presented in Section 4.4.

Table B.1: *Mean dice scores from extracting prototype embeddings from a reference sample and transferring them to a target sample. Values are reported as mean ± standard deviation.*

Reference	A	C	D	E	F
A	0.653 ± 0.03	0.537 ± 0.07	0.661 ± 0.05	0.467 ± 0.03	0.723 ± 0.02
C	0.559 ± 0.04	0.588 ± 0.06	0.677 ± 0.04	0.468 ± 0.03	0.743 ± 0.03
D	0.571 ± 0.02	0.584 ± 0.07	0.710 ± 0.05	0.402 ± 0.03	0.706 ± 0.03
E	0.428 ± 0.06	0.236 ± 0.10	0.328 ± 0.11	0.504 ± 0.02	0.537 ± 0.07
F	0.627 ± 0.02	0.499 ± 0.07	0.615 ± 0.04	0.471 ± 0.02	0.768 ± 0.02

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY