



UNIVERSITY OF PADOVA

DEPARTMENT OF MATHEMATICS "TULLIO LEVI-CIVITA"

MASTER THESIS IN DATA SCIENCE

SELF-SUPERVISED LEARNING FOR COLONOSCOPY: A STUDY ON POLYP COUNTING AND MODEL GENERALIZATION

SUPERVISOR

PROF. LAMBERTO BALLAN
UNIVERSITY OF PADOVA

CO-SUPERVISOR

DR. LUCA PAROLARI
UNIVERSITY OF PADOVA

MASTER CANDIDATE

ENRICA BONGIOVANNI

STUDENT ID

2104581

ACADEMIC YEAR

2024-2025

Abstract

All endoscopy procedures are now guided by a video stream, which allows for a less invasive operation, reduced risks, and faster recovery. However, the endoscopy datasets available today represent only a very small fraction of all colonoscopies performed. For this reason, there has been a growing need to leverage all available endoscopy data, including unannotated data, which is relatively easier to collect and significantly cheaper than data manually annotated by expert endoscopists. The recent success of self-supervised pre-training strategies has not gone unnoticed in the medical computer vision community, which has quickly adopted them to address the problem of data and annotation scarcity. However, unlike self-supervised models on natural images—which can draw on hundreds of millions, if not billions, of examples—in colonoscopy the numbers usually reach only hundreds of thousands, at best a few million. This naturally raises an important question: are these general-purpose models robust enough to be applied to tasks different from those illustrated in the pre-training data? Can they also generalize to new datasets? The aim of this thesis is to analyze and compare this type of models on the task of polyp counting, which requires the ability to determine the number of polyps encountered throughout the video. Beyond this comparative study, this thesis will explore how representation learning with different objectives can affect downstream tasks.

Contents

ABSTRACT	v
LIST OF FIGURES	x
LIST OF TABLES	xv
LISTING OF ACRONYMS	xix
1 INTRODUCTION	1
2 BACKGROUND	5
2.1 Foundations of Deep Learning	5
2.1.1 Artificial Neural Networks	6
2.1.2 Convolutional Neural Networks	7
2.1.3 Transformers	8
2.2 Self-supervised Learning	10
2.2.1 Contrastive Learning Frameworks	11
2.2.2 DINO framework	12
2.2.3 Masked Autoencoding Methods	13
2.3 Polyp Re-Identification	15
3 DATASET	23

3.1	REAL-Colon	23
3.2	Dataset Split	25
4	MODELS	27
4.1	Endo-FM	27
4.1.1	Architecture	27
4.1.2	Loss	29
4.1.3	Setup	30
4.2	EndoFM-LV	31
4.2.1	Architecture and loss	31
4.2.2	Pre-training setup	32
4.3	EndoViT	33
4.3.1	Architecture	34
4.3.2	Pre-training setup	34
4.4	SurgeNet	36
5	RESULTS	39
5.1	Performance Metrics	40
5.1.1	ReID	40
5.1.2	Retrieval	40
5.1.3	Counting	41
5.1.4	Tracking	42
5.2	Performance Analysis – Baseline Evaluation	43
5.3	Fine-Tuning and Re-evaluation of EndoFM	55
5.3.1	Losses	56
5.3.2	Setup	58

5.3.3 Results	58
6 CONCLUSION	69
A TABLES	75
REFERENCES	81
ACKNOWLEDGMENTS	89

Listing of figures

2.1	Illustration of a fully connected feed-forward neural network. Input features propagate through multiple hidden layers, where learned weights transform the representation before producing the final output.	6
2.2	Overview of a CNN. The network extracts visual features through convolution and pooling layers, followed by fully connected layers that perform the final classification.	8
2.3	Overview of the Transformer architecture. The encoder processes the input sequence through repeated layers of self-attention and feed-forward networks, while the decoder attends both to previous outputs (via masked self-attention) and to the encoder representations to produce the final output distribution.	10
2.4	Different examples of data augmentation.	12
2.5	MoCo contrastive learning pipeline. The online encoder extracts a representation for the query image, whereas the momentum encoder produces representations for a large set of keys stored in a queue. The contrastive loss encourages the query to be close to its corresponding key and far from all others.	13

2.6	Illustration of a teacher–student self-distillation framework. The student network learns to match the teacher’s centered softmax outputs using a cross-entropy loss, while the teacher weights are updated through exponential moving average (EMA).	14
2.7	Different examples of masking in time.	14
2.8	Multi-view transformer encoder.	16
2.9	Visual encoder and clustering module.	19
4.1	Overview of the EndoFM structure. The model is trained via distillation objectives computed between student and teacher outputs.	28
4.2	Diagram of EndoFM-LV. Global and local views are extracted from long endoscopy videos, patch-tokenized, and processed by a student transformer with random masking. A teacher transformer, updated via EMA, provides target embeddings. Coarse and nuanced contrastive losses enforce consistency across masked, global, and local representations, enabling long-range temporal modeling.	31
4.3	Pretraining and finetuning pipeline for EndoViT. During pretraining, masked video patches are reconstructed using an encoder–decoder architecture trained with an MSE loss. The pretrained encoder is then finetuned on downstream tasks by attaching a task-specific head.	35
4.4	Overview of the SurgeNetXL self-supervised pretraining pipeline. The model is trained on a large collection of surgical videos sourced from YouTube, private and public datasets, using a DINO-style self-distillation objective.	36
5.1	False Positive Rate across different models.	50

5.2	Fragmentation Rate across different models.	50
5.3	Precision across different models.	51
5.4	Recall across different models.	51
5.5	Association Accuracy across different models.	53
5.6	Association Precision across different models.	53
5.7	Association Recall across different models.	54
5.8	IDF ₁ across different models.	54
5.9	Validation loss comparison for the same model trained with different objectives. When using 2 views per polyp, the two training strategies exhibit identical validation loss trends.	59
5.10	Validation loss trends for the three loss objectives with $k = 4$ views per identity.	60

Listing of tables

5.1	ReID results across crop factors on REAL-Colon for EndoFM and EndoFM-LV with variable length temporal fragments.	44
5.2	ReID results across crop factors on REAL-Colon for EndoViT and SurgeNet with variable length temporal fragments.	45
5.3	ReID results across crop factors on REAL-Colon for EndoFM when fragment length is fixed to 8.	46
5.4	ReID results across crop factors on REAL-Colon for EndoFM-LV with fragment length fixed to 8 and 32.	46
5.5	ReID results across crop factors on REAL-Colon for EndoViT and SurgeNet with fragment length fixed to 8.	47
5.6	Retrieval results across crop factors on REAL-Colon with variable length temporal fragments.	47
5.7	Retrieval results across crop factors on REAL-Colon with variable length temporal fragments.	48
5.8	Retrieval results across crop factors on REAL-Colon with fragment length fixed to 8.	48
5.9	Retrieval results across crop factors on REAL-Colon with fragment length fixed to 8 and 32.	49

5.10	Retrieval results across crop factors on REAL-Colon with fragment length fixed to 8.	49
5.11	Performance comparison of EndoFM finetuning with $k = 4$ views per identity. Metrics are reported as Top-1 and Top-5 classification accuracy on validation and test sets.	61
5.12	Polyp re-identification performance of EndoFM under different loss objectives. We report AUPR and AUROC for full-frame images and for polyp-centred crops obtained with five different scaling factors. The pretrained EndoFM encoder (no finetuning) is included as a reference baseline.	61
5.13	Polyp retrieval performance of EndoFM under different loss objectives. We report mAP and Hit Rate at rank 1 and 5 (HR@1, HR@5) for full-frame images and for polyp-centred crops obtained with five different scaling factors. The pretrained EndoFM encoder (no finetuning) is included as a reference baseline.	63
5.14	Polyp counting metrics using T-AP as clustering algorithm across the analyzed models.	64
5.15	Polyp counting metrics using Threshold-based method as clustering algorithm across the analyzed models.	64
5.16	Polyp counting metrics using AP as clustering algorithm across the analyzed models.	64
5.17	Polyp counting metrics using DBSCAN as clustering algorithm across the analyzed models.	65
5.18	Polyp counting metrics using HDBSCAN as clustering algorithm across the analyzed models.	65

5.19 Polyp tracking metrics using T-AP as clustering algorithm across the analyzed models.	66
5.20 Polyp tracking metrics using Threshold-based method as clustering algorithm across the analyzed models.	66
5.21 Polyp tracking metrics using AP as clustering algorithm across the analyzed models.	66
5.22 Polyp tracking metrics using DBSCAN as clustering algorithm across the analyzed models.	67
5.23 Polyp tracking metrics using HDBSCAN as clustering algorithm across the analyzed models.	67
A.1 Polyp counting metrics using T-AP as clustering algorithm across the analyzed models.	75
A.2 Polyp counting metrics using Threshold-based method as clustering algorithm across the analyzed models.	75
A.3 Polyp counting metrics using Affinity Propagation as clustering algorithm across the analyzed models.	76
A.4 Polyp counting metrics using DBSCAN as clustering algorithm across the analyzed models.	76
A.5 Polyp counting metrics using HDBSCAN as clustering algorithm across the analyzed models.	76
A.6 Polyp tracking metrics using T-AP as clustering algorithm across the analyzed models.	76
A.7 Polyp tracking metrics using Threshold-based methods as clustering algorithm across the analyzed models.	77

A.8 Polyp tracking metrics using AP as clustering algorithm across the analyzed models.	77
A.9 Polyp tracking metrics using DBSCAN as clustering algorithm across the analyzed models.	77
A.10 Polyp tracking metrics using HDBSCAN as clustering algorithm across the analyzed models.	77
A.11 ReID rankings.	78
A.12 ReID rankings.	78
A.13 ReID rankings based on AUPR.	78
A.14 Retrieval ranking.	78
A.15 Retrieval ranking.	79
A.16 Retrieval ranking.	79
A.17 Polyp counting rankings using T-AP based on the FR.	79
A.18 Polyp tracking rankings using T-AP based on the IDF ₁	79
A.19 ReID rankings based on AUPR.	80
A.20 Retrieval ranking.	80

Listing of acronyms

SSL	Self-Supervised Learning
ANN	Artificial Neural Networks
SGD	Stochastic Gradient Descent
CNN	Convolutional Neural Networks
DINO	Distillation with No Labels
MAE	Masked Autoencoders
ADR	Adenoma Detection Rate
PPC	Polyps Per Colonoscopy
SFE	Single-Frame Encoder
MVE	Multi-View Encoder
ROC	Receiver Operating Characteristic
AUPR	Area Under the Precision-Recall curve
mAP	Mean Average Precision
HR@K	Hit Rate at K

LOOCV Leave-One-Out Cross-Validation

ViT Vision Transformer

MHSA Multi-Head Self-Attention

MTM Masked Token Modeling

MSE Mean Squared Error

SWA Stochastic Weight Averaging

HOTA Higher Order Tracking Accuracy

T-AP Temporal Affinity Propagation

AP Affinity Propagation

LN Layer Normalization

MLP Multi-Layer Perceptron

FPR False Positive Rate

TPR True Positive Rate

FR Fragmentation Rate

ReID Re-Identification

AssA Association Accuracy

AssPr Association Precision

AssRe Association Recall

IDF_I Identity F_I Score

1

Introduction

The analysis of endoscopic video data plays a crucial role in modern gastroenterology, supporting clinical workflows such as lesion detection and localization. Despite the growing availability of large-scale video recordings, the development of effective deep learning models in this domain remains hindered by a fundamental limitation: the scarcity of annotated data. Labeling endoscopic frames requires expert medical supervision, is time-consuming and becomes impractical when full-length procedures must be exhaustively annotated across multiple views and temporal segments. Consequently, there is a strong need for annotation-efficient methods capable of learning meaningful and transferable visual representations without relying on extensive supervision.

To address this challenge, recent research has turned toward **self-supervised learning (SSL)**, which enable networks to learn visual semantics directly from unlabeled videos. These models are designed to capture generalizable representations that

can be adapted to a wide range of downstream tasks with minimal fine-tuning. In the field of endoscopic video analysis, several self-supervised foundation models have been proposed, each introducing complementary innovations. Endo-FM represents one of the earliest attempts to establish a large-scale endoscopic foundation model through the self-distillation paradigm, while its variant Endo-FM-LV extends this framework to longer videos while integrating masked token modeling. Endo-ViT leverages pre-training on the Endo700k dataset to learn spatially rich and domain-specific features. More recently, SurgeNet and SurgeNetXL have pushed the scale of surgical representation learning further, exploiting hybrid convolution attention architectures (e.g., CAFormer) and self-distillation strategies to capture complex surgical and anatomical patterns.

We approach the problem by reasoning over *tracklets*, short temporal sequences of consecutive frames belonging to the same polyp. Tracklets provide a natural mid-level representation between frames and complete procedures, allowing us to analyze each lesion’s appearance across time. Starting from these tracklets, we begin with the task of **polyp re-identification (ReID)**, where the goal is to determine whether two tracklets describe the same physical polyp. We represent each tracklet through an embedding obtained by the models described above and compare pairs using a similarity matrix. We further evaluate the learned embeddings through **polyp retrieval**, where we assess how well a query tracklet is matched with all other views of the same polyp across the procedure. This task allows us to measure the quality of the learned representation and whether all instances of the same entity can be consistently retrieved despite variations in viewpoint, illumination, and tissue appearance. While ReID and retrieval focus on pairwise similarity, **polyp counting** aggregates embeddings of same polyps to estimate the total number of distinct lesions present in a procedure. We treat counting as a clustering problem, where tracklets are grouped according to their similarity in the embedding space. Each resulting cluster represents one unique polyp and the total number of

clusters represents the predicted count. This formulation directly links the quality of the learned representation to a clinically meaningful outcome: fragmentation of clusters reflects the model’s ability to discriminate among polyps with variable appearance. Multiple clustering algorithms are tested and evaluated, considering different metrics such as the **False Positive Rate (FPR)** and **Fragmentation Rate (FR)**. Finally, we extend our evaluation to **polyp tracking**, where we assess whether the model can maintain identity consistency across consecutive frames. We measure performance using adapted multi-object tracking metrics including **Association Accuracy (ASSA)** or **Identity F1 Score (IDF1)**. All tasks are evaluated on the validation and test splits of the **REAL-Colon** dataset, which provides full-procedure colonoscopy videos with polyp identity annotations, through bounding boxes, making it ideal for studying tracklet-based re-association and counting. In this way we ensure a consistent and clinically representative benchmark across models.

To further investigate how representation learning influences these downstream tasks, we re-train EndoFM on REAL-Colon. Our fine-tuning strategy explores three different loss formulations. The first is a *supervised contrastive loss*, where embeddings from the same polyp are pulled together while those from different polyps are pushed apart, encouraging clear inter-class separation. The second is a **temporally-aware contrastive loss**, which extends this framework by incorporating temporal proximity. The third objective, the **intra-inter class loss**, explicitly balances intra-class compactness and inter-class dispersion. We fine-tune EndoFM with each of these objectives independently, maintaining identical training configurations to ensure fair comparison. Once trained, we evaluate the resulting models across the four tasks using the same REAL-Colon splits. To contextualize the results, we also compare our fine-tuned models against the original pre-trained EndoFM, quantifying how different loss designs influence performance when trained on another dataset. Through this comparative study, we aim to determine which

learning formulation best preserves polyp identity across the diverse visual and temporal conditions characteristic of real colonoscopy procedures.

The contributions to the field provided by this work can be summarized as follows:

- we conduct a systematic evaluation of multiple self-supervised models across four complementary downstream tasks—polyp re-identification, retrieval, counting, and tracking—to assess their generalization and robustness in real endoscopic scenarios;
- we then study how representation learning impacts downstream performance by fine-tuning Endo-FM with different types of contrastive losses.

The Thesis is organized as follows. Chapter 2 presents the theoretical background necessary to contextualize the work, providing an overview of deep learning fundamentals and the current state of the art in self-supervised learning, both in natural and endoscopic imaging domains. Chapter 3 introduces the dataset employed for the experiments, while Chapter 4 details the models used to perform the previously described tasks. Chapter 5 reports and discusses the results obtained by evaluating these models across the four tasks, describes the fine-tuning setup adopted for Endo-FM and compares its performance with the original pre-trained model.

2

Background

This chapter provides an overview of the theoretical and methodological background underlying the work presented in this Thesis. It begins by outlining the foundations of deep learning and continues by examining the main self-supervised learning frameworks used in computer vision. Finally, the chapter focuses on recent advances in endoscopic video analysis, with particular attention to the task of polyp re-identification and counting.

2.1 FOUNDATIONS OF DEEP LEARNING

Deep Learning is a subset of machine learning that leverages deep neural architectures to automatically learn features from raw data. This is done by employing models composed of many non-linear units, known as *Artificial Neural Networks*, that progressively transform input data into semantically rich representations.

2.1.1 ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks (ANNs) are computational models inspired by the structure and functioning of the human brain, designed to recognize patterns in input data with a system of interconnected units, called *nodes* or *neurons* [1]. Each neuron receives inputs, computes a weighted sum, applies a non-linear function and passes the output to the following layers. The layers situated between the input and the output are referred to as *hidden layers*; their number defines the depth of the network and, consequently, its capacity to learn features and generalize to unseen data. A representation of an ANN is shown in Fig. 2.1 Formally, given an

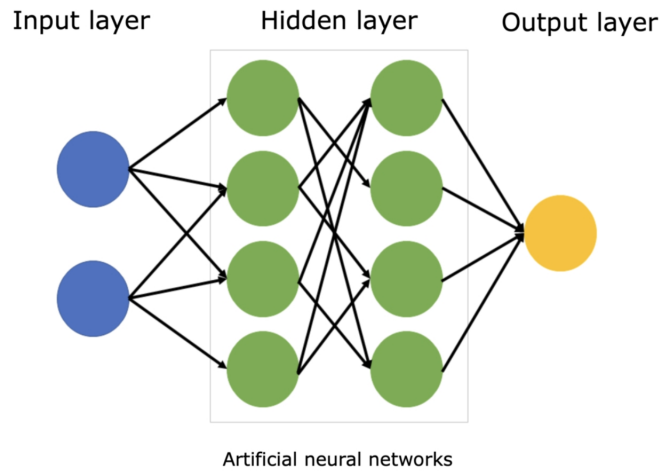


Figure 2.1: Illustration of a fully connected feed-forward neural network. Input features propagate through multiple hidden layers, where learned weights transform the representation before producing the final output.

input vector \mathbf{x} , a weight vector \mathbf{w} , a bias b and a non-linear function σ , the output of a neuron can be expressed as:

$$y(x) = \sigma(\mathbf{w} \cdot \mathbf{x} + b) \quad (2.1)$$

The training process of an ANN aims to minimize a loss function L that measures the discrepancy between the predicted output \hat{y} and the ground truth y . The model

parameters w and b are iteratively updated through *gradient-based optimization*, such as the **Stochastic Gradient Descent (SGD)** [2]. This optimization framework forms the foundation of most deep learning architectures, including more specialized variants such as Convolutional Neural Networks (CNNs), which are designed to exploit spatial correlations in image data.

2.1.2 CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks (CNNs) represent a specialized class of Artificial Neural Networks (ANNs) specifically designed to process grid-structured data, such as images. Unlike traditional ANNs, CNNs effectively address two fundamental challenges associated with visual data. First, they exhibit *translation equivariance*, enabling the model to recognize visual patterns regardless of their spatial position within the image. Second, they significantly reduce the number of learnable parameters through *parameter sharing* and *sparse connectivity* between units, thus enhancing computational efficiency.

The architecture of a typical Convolutional Neural Network (CNN) layer consists of three fundamental components. The first is the **convolutional layer**, which performs the core operation from which the architecture derives its name. Given an input image I and a **kernel** (or **filter**) function K , the kernel slides across the spatial dimensions of the image—both height and width—computing a local dot product between the kernel weights and the corresponding region of the input at each spatial location. The resulting set of local responses forms a feature map S , which encodes the spatially varying activation patterns detected by the filter. This process is formally defined by the **convolution operator** as follows:

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i - m, j - n) K(m, n) \quad (2.2)$$

Each layer usually has multiple kernels, each one of them focusing on a specific

type of feature. The **pooling layer** serves to progressively reduce the spatial dimensions of the feature maps, thus decreasing computational complexity and enhancing the model’s robustness to small translations in the input. Common pooling operations include **max pooling**, which selects the maximum activation within each local neighborhood, and **average pooling**, which computes the mean value over the same region. After several convolutional and pooling operations, the resulting feature maps are flattened and passed through one or more **fully connected layers**, which perform classification. In these layers, each neuron is connected to every neuron in the preceding layer, allowing the network to integrate the spatially distributed features extracted by the convolutional hierarchy into a compact, class-specific representation. This dense connectivity enables the model to capture complex, non-linear relationships among features and ultimately map them to the desired output classes. The whole structure of a CNN is shown in Fig. 2.2.

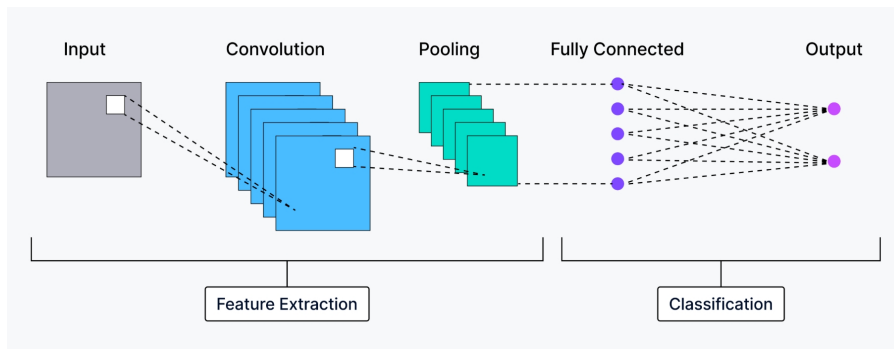


Figure 2.2: Overview of a CNN. The network extracts visual features through convolution and pooling layers, followed by fully connected layers that perform the final classification.

2.1.3 TRANSFORMERS

Transformers represent a paradigm shift in deep learning architectures, introducing a fundamental innovation known as the **self-attention** mechanism [3]. This mechanism enables the model to dynamically assess the relevance of different parts (**tokens**) of the input with respect to one another when computing contextualized

representations. Each input token is linearly projected into three distinct vectors: a **query** Q , a **key** K , and a **value** V . The attention operation then computes a similarity score between queries and keys, scales it to stabilize gradients, and normalizes the result using the softmax function. The resulting attention weights are subsequently applied to the values, yielding a weighted combination that captures the global dependencies among tokens. Formally, this process can be expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (2.3)$$

The equation above represents the output of a single attention head. In practice, however, Transformers employ **multi-head attention**, wherein multiple attention heads operate in parallel to capture relationships across different subspaces of the feature representation. Each head independently learns to focus on distinct aspects of the input dependencies, thereby enriching the overall contextual understanding. The outputs of all heads are then concatenated and projected through a linear transformation to produce the final attention output:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_0, \dots, \text{head}_n) \cdot W^o \quad (2.4)$$

The Transformer architecture is composed of two main components: an **encoder** and a **decoder** (Fig. 2.3). The encoder maps an input sequence into a latent representation by combining positional encoding with a stack of N identical layers, each consisting of a multi-head attention mechanism followed by a position-wise feed-forward network. The decoder then generates the output sequence conditioned on this latent representation through another stack of N identical layers. Each decoder layer incorporates a variant of the attention mechanism known as **masked self-attention**, which restricts each position to attend only to previous tokens in the sequence, thereby preserving the causal structure required for autoregressive generation. Both the encoder and decoder layers include fully connected

feed-forward sublayers applied independently to each position, along with residual connections and layer normalization to facilitate gradient flow and stabilize training. Finally, a linear projection layer followed by a softmax function is employed to map the decoder outputs to the target class space.

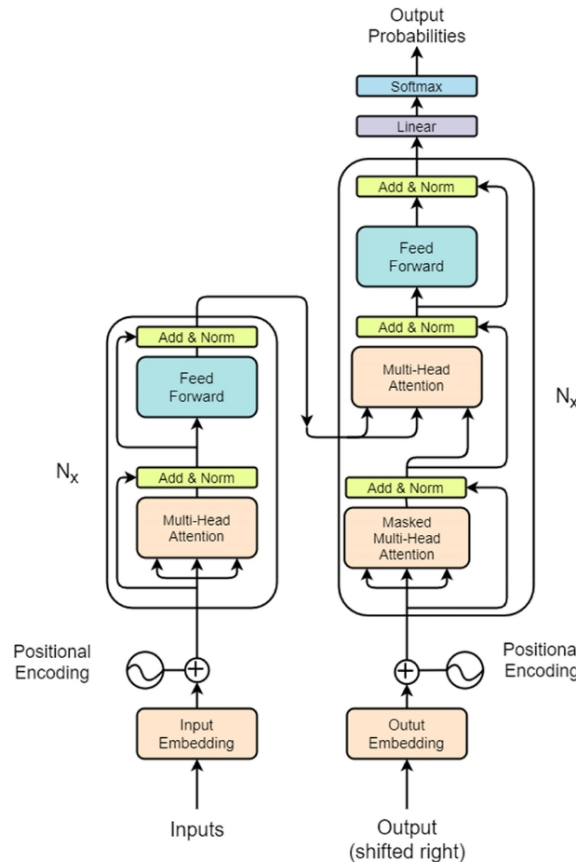


Figure 2.3: Overview of the Transformer architecture. The encoder processes the input sequence through repeated layers of self-attention and feed-forward networks, while the decoder attends both to previous outputs (via masked self-attention) and to the encoder representations to produce the final output distribution.

2.2 SELF-SUPERVISED LEARNING

In recent years, **self-supervised learning (SSL)** has emerged as one of the most promising paradigms for representation learning, particularly in domains where obtaining large amounts of labeled data is expensive or impractical—such as in

medical imaging or endoscopic video analysis. Unlike traditional supervised learning, which relies on manually annotated datasets, SSL generates output labels “intrinsicly” by the input data itself, by discovering patterns or relationships between data components [4]. In other words, the model learns useful representations without human-provided labels, by solving a pretext task designed to predict certain aspects of the input from other observed parts.

The central idea of self-supervised learning is to construct pseudo-labels or self-generated targets from the raw data, allowing the network to learn general and transferable features. For instance, in computer vision, common pretext tasks include reconstructing missing patches or distinguishing between different augmented views of the same sample. Through these tasks, the model learns to capture semantic invariances—that is, to recognize which features of the input remain consistent under transformations such as cropping, color jittering, or temporal shifts.

Formally, SSL includes different families of models, such as **contrastive** and **generative** methods. By pre-training on large collections of unlabeled data, self-supervised models can then be fine-tuned on small labeled datasets for downstream tasks—such as classification, segmentation or detection—often achieving results comparable to, or even surpassing, fully supervised counterparts. This paradigm has proven particularly effective in medical domains, where annotated data are scarce and costly to obtain.

2.2.1 CONTRASTIVE LEARNING FRAMEWORKS

The earliest SSL frameworks **SimCLR** [5] and **MoCo** [6] belong to the so-called family of **contrastive learning** methods. SimCLR introduces the idea of maximizing the agreement between two augmented views of the same image (positive pairs) against all the other samples of the batch (negatives). Augmented views can be obtained by applying transformation such as random cropping, Gaussian blur

or color jitter, as illustrated in Figure 2.4. This method encourages the model to learn invariances to visual transformation while still being able to discriminate between different instances. However, its performance is strictly dependent on the number of negatives, which limits its scalability in memory-constrained settings.



Figure 2.4: Different examples of data augmentation.

MoCo addresses this limitation by employing two encoders (query and key) and a memory queue. The two encoders process two different views of the same image, forming a positive pair (Figure 2.5). The key embedding is added to the queue, whose job is to store thousands of other keys that serve as negatives for future queries. This design decouples the number of negatives from the batch size, making contrastive learning more scalable without the need for large computational demands.

MoCov2 (Improved Baselines with Momentum Contrastive Learning, [7]) refined this framework with a few improvements: a two-layer MLP projection head (instead of using raw encoder features) that improves the invariance and feature quality, stronger data augmentations (such as a more aggressive blur) that dramatically improved generalization and longer training up to 800 epochs.

2.2.2 DINO FRAMEWORK

DINO (Distillation with No Labels) [8] extended the concept beyond contrastive learning by applying a mechanism of *self-distillation*, where a student model learns

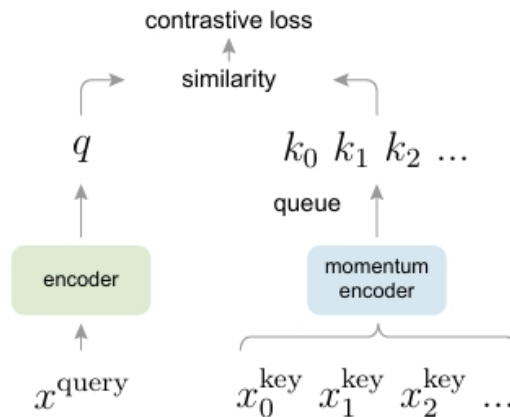


Figure 2.5: MoCo contrastive learning pipeline. The online encoder extracts a representation for the query image, whereas the momentum encoder produces representations for a large set of keys stored in a queue. The contrastive loss encourages the query to be close to its corresponding key and far from all others.

to predict the same output distribution of a teacher network, removing the need for negative samples altogether (Figure 2.6). While the teacher is updated with a standard backpropagation, the student is upgraded through an exponential moving average that also exploits the teacher’s weights.

2.2.3 MASKED AUTOENCODING METHODS

While contrastive methods focus on discriminative learning between samples, a complementary line of research aims to model contextual relationships within the data itself through masked autoencoding. Introduced in [9], in the context of Vision Transformers, the **Masked Autoencoder (MAE)** learns by reconstructing heavily masked images, forcing the encoder to capture global structure and semantic information rather than low-level pixel statistics. Its video extension, **VideoMAE** [10], applies this principle to spatiotemporal data by masking large portions of input video tokens and forcing the model to reconstruct them from visible patches. Different types of masking are shown in Fig. 2.7.

VideoMAE v2 [11] refines both the training pipeline and representation learn-

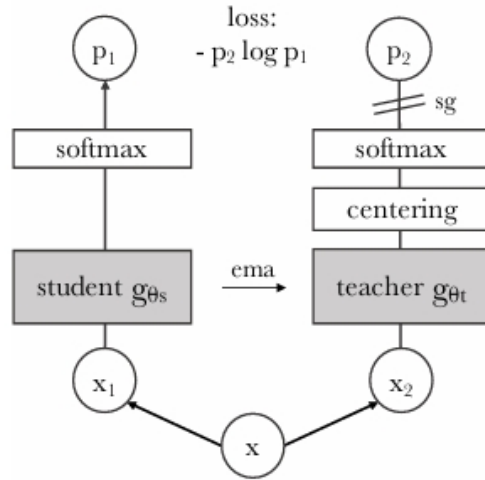


Figure 2.6: Illustration of a teacher–student self-distillation framework. The student network learns to match the teacher’s centered softmax outputs using a cross-entropy loss, while the teacher weights are updated through exponential moving average (EMA).

ing quality making masked video pretraining more scalable and transferable. The model is trained on larger and more diverse datasets; the tube masking is applied to variable temporal spans, teaching the model to handle both slow and fast motions. Moreover, VideoMAE v2 leverages multi-scale temporal modeling by sampling frames at multiple rates, which helps to better capture long-range dependencies.

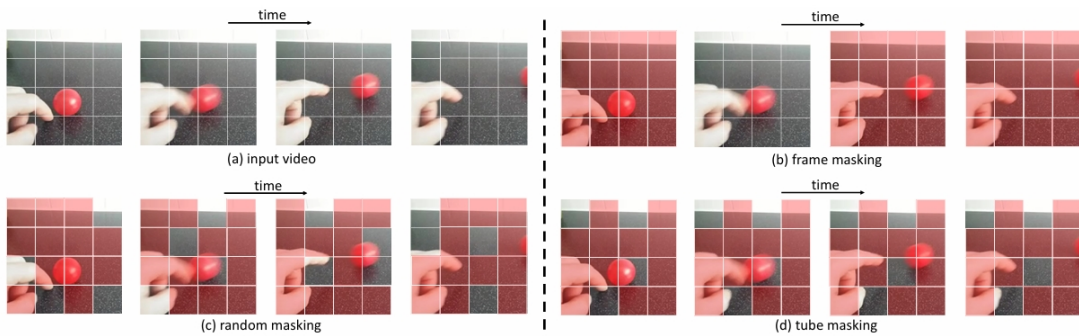


Figure 2.7: Different examples of masking in time.

Building on these advances, Hu et al. (2024) proposed **M2CRL (Multi-View**

Masked Contrastive Representation Learning) [12], the first framework to unify contrastive and masked pretraining for endoscopic video analysis. M2CRL introduces a multi-view masking mechanism that generates diverse views of the same endoscopic clip through two complementary strategies: first, an attention-guided tube masking that removes temporally coherent regions to encourage long-range context and a random tube masking, that randomly samples masked tokens in the 2D spatial domain and then extends these tokens along the temporal axis. Training involves the combination of a masked reconstruction and a contrastive loss that exploits multiple masked views. The results showed an improved performance on classification, segmentation and detection over other masked autoencoders such as VideoMAE and VideoMAE v2 and contrastive methods such as Endo-FM.

2.3 POLYP RE-IDENTIFICATION

Modern computer-aided colonoscopy systems are evolving from frame-based detectors toward entity-centric understanding, in which each polyp is treated as a unique, temporally persistent object. This shift enables clinically meaningful automation: models can identify which lesion is being observed, when, and how many times throughout a procedure. Such capability supports the automatic computation of quality indicators such as the **Adenoma Detection Rate (ADR)** and **Polyps Per Colonoscopy (PPC)** and provides a foundation for AI-assisted decision-making and reporting. Within this paradigm, several interrelated tasks emerge.

The **polyp re-identification (ReID)** task seeks to determine whether two visual observations—in short fragments, called **tracklets**—depict the same physical lesion. In colonoscopy, this is non-trivial: the same polyp may reappear after the camera retracts or repositions, often under drastically different lighting and viewing angles. Compared to object or person ReID in natural images, polyp ReID faces multiple challenges, such as the deformable nature of the tissue, illumination

changes or low inter-polyp variability, as many polyps share similar color and texture.

The work of Intrator et al. [13] introduced a self-supervised contrastive learning framework for polyp ReID based on SimCLR, eliminating the need for manual annotation. Instead of relying on image augmentations alone, they leveraged temporal coherence: temporally close frames within the same tracklet serve as positive pairs, while frames from different polyps act as negatives. Two encoder architectures were proposed: a **single-frame encoder (SFE)** that aggregates individual frame embeddings, and a **multi-view encoder (MVE)** that learns tracklet-level representations end-to-end via a transformer backbone, as represented in Fig. 2.8. This work demonstrated that self-supervised temporal pairing can produce ro-

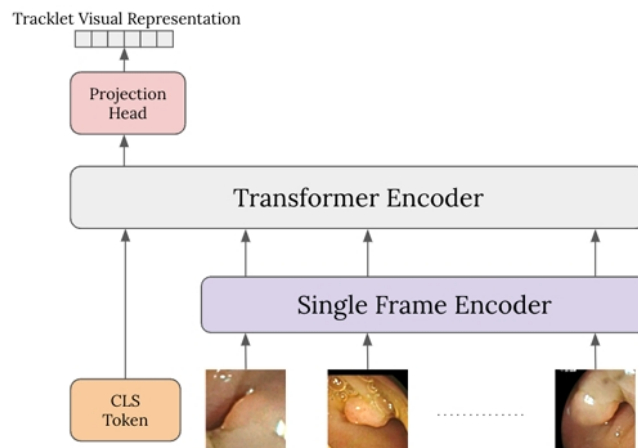


Figure 2.8: Multi-view transformer encoder.

bust appearance embeddings, significantly improving lesion-level consistency and downstream diagnostic modeling.

Building on this foundation, Parolari et al. [14] extended the problem to full-procedure colonoscopy videos. Their study reframed ReID as a two-step pipeline: learning tracklet embeddings via SimCLR-based encoders and grouping them via unsupervised clustering rather than thresholded similarity. They evaluated three

clustering algorithms—Hierarchical Clustering, HDBSCAN and Affinity Propagation—and showed that clustering-based re-association dramatically reduces the **fragmentation rate** (the average number of tracklets per polyp). Affinity Propagation achieved a $3.9\times$ improvement over previous methods, with a false positive rate of 5%. In their subsequent MICCAI 2025 paper, Parolari et al. introduced a **temporally-aware supervised contrastive framework** [15], transitioning from self-supervised to fully supervised ReID. Here, multiple tracklets of the same polyp serve as positives, while those from other lesions act as negatives. Crucially, soft temporal targets were introduced: embeddings are regularized so that temporally adjacent tracklets are mapped closer together, reflecting smooth changes in visual appearance. They also proposed a temporal penalty in the clustering stage, discouraging associations between visually similar but temporally implausible tracklets (e.g., distant in time). This temporally-aware formulation reduced fragmentation by over $2\times$ compared to prior work, setting a new state of the art for full-procedure ReID.

To evaluate the discriminative ability of re-identification models, we employ a dedicated **ReID evaluation module** that operates directly on the learned embeddings. For each test sequence, the features extracted by the encoder are concatenated into a single embedding matrix, and their corresponding polyp identities are collected. All embeddings are L2-normalized and passed through a sigmoid transformation to map values into the range $(0, 1)$. This produces a dense similarity matrix where each entry reflects the likelihood that two observations correspond to the same lesion. A binary ground-truth matrix is constructed in parallel, where entries are set to one if the two detections belong to the same identity and zero otherwise. To remove redundant self-pairs and preserve symmetry, only the lower triangular part of these matrices is retained, yielding one-dimensional vectors of predictions and binary labels. These vectors form the basis for computing standard ranking-based metrics such as the **Receiver Operating Characteristic (ROC)** curve and the

Precision–Recall (PR) curve, along with their respective areas under the curves (AUROC and AUPR). These metrics capture the model’s ability to distinguish between same-polyp and different-polyp pairs across all possible similarity thresholds, providing a threshold-independent measure of discriminative quality. Together, AUROC and AUPR summarize how well the learned embedding space supports reliable polyp re-identification, complementing clustering-based analyses and enabling fair comparison across different encoder architectures

Polyp retrieval generalizes re-identification to large-scale search: given a query polyp, the goal is to match it against a large number of candidates. In practice, retrieval performance is assessed through a dedicated retrieval evaluator module, which measures the quality of the embeddings. During testing, all feature embeddings and their corresponding identity labels are collected across the dataset and concatenated into global tensors. The embeddings are once again L2-normalized, forming the same similarity matrix described before. Finally, the ground-truth matrix is created with the same masking approach used for ReID. Evaluation is performed using ranking-based metrics. Specifically, the **Mean Average Precision (mAP)** quantifies how well relevant instances are ranked ahead of irrelevant ones, while the **Hit Rate at k (HR@k)** measures the proportion of queries whose correct match appears within the top-k retrieved candidates (with $k = 1$ and $k = 5$ in our setup). These metrics are computed for each query and then averaged across the dataset, yielding a comprehensive measure of embedding quality. Both mAP and HR@k are logged as scalar indicators of retrieval accuracy, providing a clear and interpretable evaluation of how well the learned representation space captures identity-level similarity under realistic variations in viewpoint, illumination, and tissue deformation.

Moving onto **polyp counting**, its goal is to estimate the total number of unique polyps observed during a colonoscopy by aggregating all tracklets that correspond

to the same lesion. Because the same polyp may appear multiple times, naïve frame- or detection-level counting severely overestimates the number of lesions. Intrator et al. (2023) and Parolari et al. (2025) formalized counting as an entity-level clustering problem. After tracklet embeddings are extracted, they are grouped into clusters—each representing a distinct lesion—using similarity-based algorithms. Counting then reduces to computing the number of clusters $|R|$ after re-association relative to the ground-truth number of entities $|E|$ captured by the FR:

$$FR = \frac{|R|}{|E|} \quad (2.5)$$

with $|FR| \geq 1$ where a perfect clustering yields $|FR| = 1$. To avoid over-merging, results are typically reported at a fixed false positive rate (FPR), equal to 5%. Using

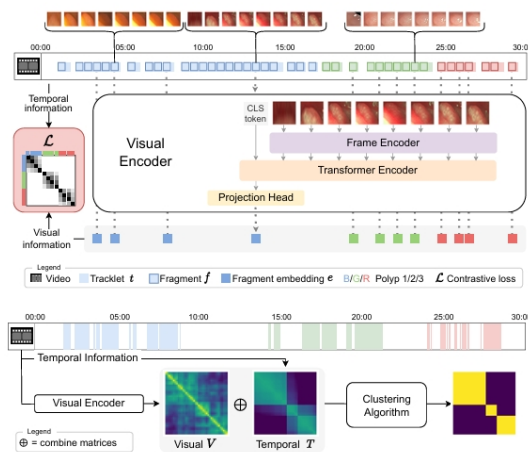


Figure 2.9: Visual encoder and clustering module.

REAL-Colon [16], Parolari et al. showed that replacing threshold-based linking with Affinity Propagation reduces fragmentation from 24.60 to 6.30, demonstrating the efficacy of unsupervised clustering for polyp counting. Their framework (Fig. 2.9) further improved results through supervised and temporally-aware contrastive learning, achieving a fragmentation rate reduction of more than twofold compared to prior self-supervised methods.

In order to assess the accuracy of polyp enumeration and determine the optimal clustering configuration, we employ a dedicated module. This component operates by comparing the predicted associations between tracklets to the ground-truth identity relationships within each video. During validation, the evaluator aggregates all feature embeddings, identity labels, video identifiers, and frame positions across the dataset. The embeddings are first L2-normalized and pairwise similarities are computed and passed through a sigmoid activation to obtain bounded similarity scores. Temporal proximity is likewise encoded as the absolute difference between frame indices, producing a second matrix that captures how close in time two detections occur. Both matrices are then stacked to form a two-channel feature tensor that integrates visual and temporal cues. For each video, the evaluator iterates over all parameter configurations of the clustering algorithm and applies them to the two-channel representation. The resulting predicted associations are compared against the binary ground-truth matrix of same-identity pairs, enabling the computation of evaluation metrics for each parameter set. These metrics are averaged across all validation videos to obtain a global performance profile. The configuration that achieves a target false positive rate of 5% is automatically selected as the optimal one, and its corresponding index is stored for subsequent use during testing. This leave-one-out validation scheme ensures that the parameters used at test time generalize well across different procedures. In the testing phase, the evaluator reuses the best configuration identified during validation to perform counting on previously unseen videos. The process mirrors that of validation: embeddings and positions are used to reconstruct pairwise similarity and distance matrices, which are then clustered according to the fixed parameter set. The predicted associations are compared against ground truth to compute standard binary classification metrics such as precision, recall, and confusion matrix statistics and more specific metrics such as the false positive rate and fragmentation rate. All results are logged and reported, providing a quantitative measure of the system’s ability

to correctly aggregate detections belonging to the same lesion within a video and to estimate the total number of distinct polyps observed during the procedure.

After performing the polyp counting task, the best clustering configuration is identified and reused for the tracking evaluation. The idea is to determine a single optimal set of clustering hyperparameters, depending on the type of algorithm, that yields the best counting accuracy and transfer this same configuration to the tracking phase. During testing, all per-frame detections are passed through the feature extractor, producing an embedding vector for each instance. Alongside these embeddings, the system keeps track of the ground-truth identity label, the corresponding video identifier and the temporal position within that video. For each video, the system constructs three pairwise matrices that describe relationships between detections. The first matrix captures appearance similarity, computed from the feature embeddings in the same way as for the previous tasks. The second matrix encodes temporal distance, obtained as the absolute difference between frame indices. This measures how far apart in time two detections occur. The third matrix represents the ground-truth relationship, a binary target where each entry is 1 if two detections correspond to the same annotated polyp and 0 otherwise. The appearance and distance matrices are then stacked along a new dimension, resulting in a two-channel tensor where the first channel encodes visual similarity and the second encodes temporal proximity. This multimodal tensor serves as input to the clustering algorithm. For each video, the algorithm produces an assignment matrix indicating which detections belong together. This output is post-processed to form clusters, each corresponding to a predicted track — that is, a temporal sequence of detections believed to depict the same polyp. Each detection is then assigned a predicted track ID corresponding to the cluster it belongs to. Once the predicted clusters have been computed, both the ground-truth and predicted tracking data are exported in the MOTChallenge format so that they can be evaluated using the standardized TrackEval framework consisting of two main metrics: IDF1

and HOTA (Higher Order Tracking Accuracy). The latter further decomposes tracking into several interpretable components, among which the most relevant are the Association Accuracy (AssA), Association Recall (AssRe) and Association Precision (AssPr). This interpretable metrics reflect the system's ability to correctly associate and maintain polyp identities across frames.

3

Dataset

3.1 REAL-COLON

The **REAL-Colon dataset** (Real-world Endoscopy Annotated Library) was conceived to bridge a persistent gap between laboratory-scale datasets and the complex, heterogeneous conditions encountered in clinical colonoscopy. Most public endoscopic datasets to date, such as CVC-ClinicDB [17] or Kvasir-SEG [18], consist of short, preselected video clips or static frames focused on individual lesions. While invaluable for the development of segmentation or detection algorithms, these datasets fail to capture the temporal and procedural variability inherent in real clinical workflows—where the camera moves dynamically through the colon, the same polyp may appear multiple times under different conditions, and long segments contain no lesions at all. The motivation behind REAL-Colon was therefore to create a benchmark that reflects the true procedural and temporal complexity of colonoscopy. Its primary goal is to support the development and evaluation of real-

world AI systems capable of operating continuously throughout an entire examination, rather than in isolated frames. Specifically, the dataset was designed to enable research in long-form video understanding, including tasks such as polyp detection, tracking, re-identification, retrieval, and counting—tasks that inherently depend on modeling appearance consistency and temporal relationships over extended time spans.

To this end, REAL-Colon provides full-procedure colonoscopy recordings collected across four medical centers, ensuring diversity in patient anatomy, bowel preparation quality, and endoscopic hardware. The final release includes 60 complete colonoscopy videos, corresponding to approximately 2.7 million high-resolution frames, each representing the entire duration of the procedure from insertion to withdrawal. Out of the 60 videos present in the dataset, 14 do not contain any polyp. This large-scale and untrimmed structure contrasts sharply with conventional datasets limited to curated polyp sequences, offering a faithful representation of the workflow dynamics encountered in clinical practice.

Every polyp that was removed during the procedure is annotated with bounding boxes for all frames in which it appears, yielding roughly 350,000 annotated frames and 132 polyps in total. In addition, each lesion is accompanied by rich metadata, including anatomical location, size, and histopathological outcome, as well as procedure-level information such as the colon segment inspected, withdrawal time, endoscope brand, and bowel preparation quality (recorded via the Boston Bowel Preparation Scale). This multi-level annotation framework allows the dataset to be used not only for computer vision benchmarks but also for clinically meaningful analyses, such as the study of inspection completeness, lesion recurrence, or polyp dwell time. A distinctive feature of REAL-Colon lies in its temporal continuity. Because the videos are untrimmed, a single lesion may appear in multiple disjoint intervals—an aspect that introduces natural fragmentation and challenges models

to maintain consistent identity over time.

Beyond its methodological contribution, REAL-Colon embodies a broader paradigm shift in medical AI dataset design. Its creators explicitly sought to move from curated datasets—where visual conditions are artificially controlled—to realistic datasets that embrace the variability, imperfections, and unpredictability of clinical imaging. This includes handling motion blur, specular reflections, occlusions by instruments, and varying illumination—all of which are unavoidable in actual procedures but seldom represented in research datasets. In doing so, REAL-Colon promotes the development of models that are not only accurate but also robust and generalizable to deployment scenarios.

3.2 DATASET SPLIT

For the re-identification and retrieval tasks, the validation and test subsets of the REAL-Colon dataset were combined into a unified evaluation split, resulting in a total of 19 full-procedure colonoscopy videos. This configuration was chosen to increase the diversity of polyp instances and imaging conditions while maintaining a clear separation from the training data. In particular, the videos used for these tasks are the last seven for each of the four cohorts.

For the counting and tracking tasks, the evaluation followed a **leave-one-out cross validation (LOOCV)** protocol. In this setting, the 19 videos from the REAL-Colon validation and test split were used, iteratively selecting one video as the test case and the remaining videos for validation in each round. In particular, the following videos were chosen for the LOOCV:

- *Cohort 001*: 001-009, 001-010, 001-014;
- *Cohort 002*: 002-009, 002-010, 002-011, 002-014;
- *Cohort 003*: 003-009, 003-010, 003-012, 003-013, 003-014, 003-015;

- *Cohort 004*: 004-009, 004-011, 004-012, 004-013, 004-014, 004-015.

All images were resized from their original resolution (1920 × 1080) to a standard size of 224 × 224.

4

Models

4.1 ENDO-FM

4.1.1 ARCHITECTURE

Endo-FM [19] is a foundation model for endoscopic video analysis that employs self-supervised pre-training on large-scale endoscopic video data. The model leverages a Video Transformer backbone trained via DINO framework to learn robust spatial-temporal representations from unlabeled endoscopic videos. At its core, Endo-FM employs a *Timesformer* architecture - a Vision Transformer (ViT) variant specifically adapted for video input - with a ViT-Base/16 backbone. The network consists of 12 encoder blocks equipped with **divided space-time attention** and a dynamic spatio-temporal positional encoding mechanism. In particular, it takes as input an endoscopic video clip $X \in \mathbf{R}^{T \times 3 \times H \times W}$ composed of T frames of size $H \times W$. Each frame is divided into $N = \frac{HW}{P^2}$ patches of size $P \times P$, and each

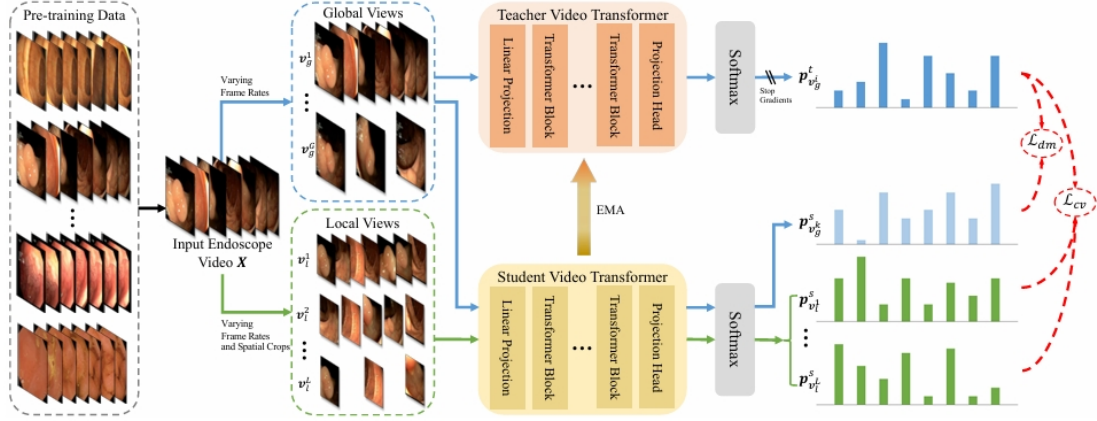


Figure 4.1: Overview of the EndoFM structure. The model is trained via distillation objectives computed between student and teacher outputs.

patch is subsequently processed and converted into a token. Specifically, given an intermediate token $z^m \in \mathbf{R}^D$ for a patch from the m -th block, the token is updated as follows:

$$z_{time}^{m+1} = MHSA_{time}(LN(z^m)) + z^m, \quad (4.1)$$

$$z_{space}^{m+1} = MHSA_{space}(LN(z_{time}^{m+1})) + z_{time}^{m+1}, \quad (4.2)$$

$$z^{m+1} = MLP(LN(z_{space}^{m+1})) + z_{space}^{m+1}, \quad (4.3)$$

where $MHSA$ denotes multi-head self-attention, LN represents layer normalization, and MLP is the multi-layer perceptron. The model uses a projection head that maps the backbone features to a high-dimensional space. The head consists of a multi-layer perceptron with GELU activations, batch normalization and a final weight-normalized linear layer. Unlike standard Vision Transformers that rely on static positional encodings, Endo-FM incorporates a dynamic spatio-temporal encoding mechanism to handle variations in spatial viewpoints, resolutions and frame rates. This design enables robust transfer across different endoscopy datasets, where both image size and temporal sampling may vary greatly. The pre-training

strategy follows a **teacher–student paradigm**, in which the student network is trained to match the output representations of the teacher. Given an input video X , two types of spatio–temporal views are generated. The global views $v_G^i \in \mathbf{R}^{T_g \times 3 \times H_g \times W_g}$ are obtained by uniformly sampling frames from X at different frame rates, while the local views $v_L^j \in \mathbf{R}^{T_l \times 3 \times H_l \times W_l}$ are generated through random spatial cropping of X , with $T_l \leq T_g$. The teacher network processes only the global views, whereas the student receives both global and local ones. The output feature vectors f are normalized using a softmax function with temperature τ to produce the probability distribution $p = \text{softmax}(\frac{f}{\tau})$.

4.1.2 LOSS

More specifically, the pre-training process relies on two complementary schemes designed to address challenges inherent to endoscopic video data. The first, named **Cross-view Matching**, deals with the varying amount of contextual information across frames—depending on the presence or absence of lesion regions and their relative proportions within the field of view. In this approach, the student predicts the teacher’s global-view representations using its own local-view inputs. Through this mechanism, the model learns both spatial context (e.g., neighboring tissue structures within a local crop) and temporal context (e.g., anticipating lesion appearance in subsequent frames). The cross-view objective is optimized by minimizing the following loss:

$$L_{cv} = \sum_{i=1}^G \sum_{j=1}^L -p_{v_g^i}^t \cdot \log p_{v_l^j}^s \quad (4.4)$$

The second scheme, named **Dynamic Motion Matching**, addresses the variability in camera speed and range of motion that commonly occurs during endoscopic procedures. Such dynamic scenarios make it challenging to train a model that performs robustly across different motion patterns. To mitigate this issue, the

student network predicts the teacher’s global-view representations using its own global views, each sampled at different frame rates. This encourages the model to learn motion-invariant representations that remain stable under varying temporal dynamics. The corresponding loss function to be minimized is given by:

$$L_{dm} = \sum_{i=1}^G \sum_{j=1}^G 1_{i \neq j} p_{v_g^i}^t \log p_{v_g^j}^s \quad (4.5)$$

The overall pre-training objective is defined as $L_{\text{pre-train}} = L_{cv} + L_{dm}$. The student network is updated through standard backpropagation, whereas the teacher network is updated using an exponential moving average of the student’s weights, according to the rule $\varphi_t \leftarrow \alpha \varphi_{t-1} + (1 - \alpha) \theta_t$ where φ and θ represent the sets of weights of the teacher and student, respectively, t denotes the training iteration and α is the momentum hyperparameter.

4.1.3 SETUP

During pre-training, input frames are augmented using various strategies, including random horizontal flips, color jittering, and Gaussian blurring, among others. For each patch (with size $P = 16$), $G = 2$ global and $L = 8$ local views are generated, with spatial resolutions of 224×224 and 96×96 , respectively. The frame sampling rate for global views is set to $T_g \in [8, 16]$, while the one for local views is $T_l \in [2, 4, 8, 16]$.

The model is pre-trained on seven different datasets (Colonoscopic [20], SUN [21] & SUN-SEG [22], LDPolypVideo [23], Hyper-Kvasir [24], Kvasir-Capsule [25], CholecTriplet [26] and a private one) comprising colonoscopy, gastroscopy, and laparoscopy videos, amounting to a total of 32,896 videos and 5,024,101 frames. Endo-FM is then evaluated on three downstream tasks: classification, segmentation, and detection. For classification, a linear layer is appended to the backbone

and fine-tuned for 20 additional epochs on the PolypDiag dataset [27]. For segmentation, a TransUNet architecture [28] equipped with the Endo-FM backbone is trained on the CVC-12K dataset [29], while for detection, Endo-FM serves as the backbone of an STFT [30] model trained on the KUMC dataset [31].

4.2 ENDOFM-LV

4.2.1 ARCHITECTURE AND LOSS

Building upon the Endo-FM architecture, EndoFM-LV [32] extends the foundation model to learn from longer video sequences, with an average duration of approximately one minute. The overall architecture—a ViT-B/16 video transformer comprising 12 encoder blocks and a dynamic spatio-temporal encoding strategy—remains unchanged. However, to better capture long-term temporal dependencies, Endo-FM-LV introduces a key innovation: **the masked token modeling (MTM) scheme**.

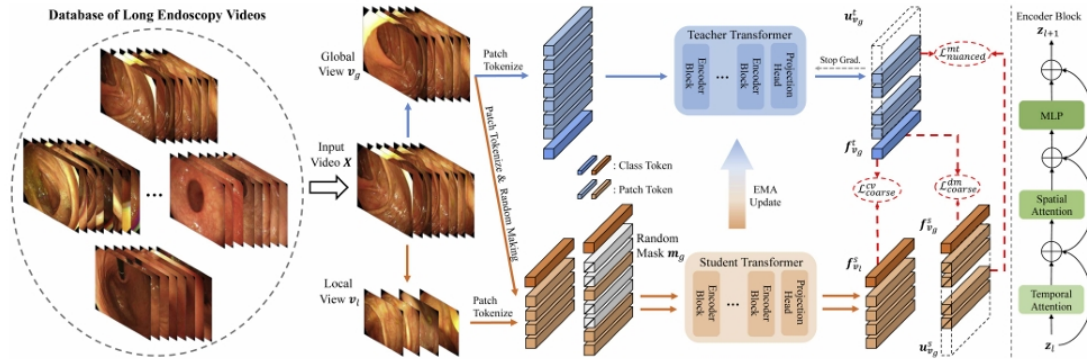


Figure 4.2: Diagram of EndoFM-LV. Global and local views are extracted from long endoscopy videos, patch-tokenized, and processed by a student transformer with random masking. A teacher transformer, updated via EMA, provides target embeddings. Coarse and nuanced contrastive losses enforce consistency across masked, global, and local representations, enabling long-range temporal modeling.

In this framework, a subset of patch tokens in the student’s input is randomly masked according to a probability γ . The model is thus required to reconstruct

the masked patches, under the assumption that this process enhances the quality and robustness of the learned feature representations. The corresponding objective can be formulated as:

$$L_{\text{nuanced}}^{mt} = \sum_{i=1}^G \sum_{k=1}^G -1_{i \neq k} m_g^{i,k} T\left(\frac{u_{v_g^i}^t}{\tau_t}\right) \log T\left(\frac{u_{v_g^k}^s}{\tau_s}\right) \quad (4.6)$$

where $m \in \{0, 1\}$ denotes the randomly sampled mask. During pre-training, a KoLeo regularizer is also introduced to encourage a uniform span of feature representations within each batch:

$$L_{\text{koleo}} = -\frac{1}{G} \sum_{i=1}^G \sum_{j=1}^G \log(\min_{j \neq i} \|f_{v_g^i}^f - f_{v_g^j}^f\|) \quad (4.7)$$

where $\min_{j \neq i} (\|f_{v_g^i}^f - f_{v_g^j}^f\|)$ represents the minimum distance between two different class token feature f from the global views output by the student model.

The overall pre-training objective combines the coarse-level cross-view and dynamic-motion losses inherited from Endo-FM with the new masked token and KoLeo regularization [33] terms:

$$L_{\text{EndoFM-LV}} = L_{\text{coarse}}^{cv} + L_{\text{coarse}}^{dm} + L_{\text{nuanced}}^{mt} + \lambda_{\text{koleo}} L_{\text{koleo}} \quad (4.8)$$

where λ_{koleo} controls the influence of the KoLeo regularizer.

4.2.2 PRE-TRAINING SETUP

The number and size of global and local views remain consistent with Endo-FM. However, the frame sampling rate for global views is set to $T_g \in [16, 32]$, while for local views it is $T_l \in [4, 8, 16, 32]$. The pre-training procedure also follows the same teacher–student paradigm: the student network is updated via backpropagation, while the teacher parameters are updated through an exponential moving average

of the student’s weights, as described in Section 4.1. The same augmentation strategies used in Endo-FM—random horizontal flips, color jittering, Gaussian blurring, and solarization—are applied here as well.

EndoFM-LV is pre-trained on three datasets comprising 6,469 colonoscopy and gastroscopy videos, each with an average duration of 68.1 seconds, amounting to a total of 13,224,974 frames. The model is tested on four downstream tasks, three of which are the same as for EndoFM with an identical setup. The fourth task, however, involves the *workflow recognition* in which EndoFM-LV serves as temporal module for extracted features and is trained on the Cholec80 dataset.

4.3 ENDOViT

EndoViT [34] is a Vision Transformer pretrained specifically on large-scale endoscopic data. The idea behind EndoViT is that domain-specific pretraining can yield more semantically meaningful features than relying only on natural image datasets such as ImageNet. However, typical benchmark datasets as, for instance, Cholec80, used for tasks such as surgical phase recognition, only contain on the order of 200,000 images which is several orders of magnitude smaller than natural images datasets used in foundation models. EndoViT was designed to overcome this issue through self-supervised pre-training on a large corpus of unlabeled endoscopic frames, named **Endo700k**. The dataset comprises over 700,000 unlabeled frames drawn from nine public datasets. The collection covers a wide spectrum of surgical procedures, including laparoscopic and robot-assisted interventions. To prevent data leakage, all images overlapping with validation and test splits of downstream datasets (Cholec80, CholecT45 and CholecSeg8) were removed. For consistency, each video was sampled at 1 frame per second (FPS) to ensure temporal diversity while avoiding redundancy.

4.3.1 ARCHITECTURE

EndoViT adopts the Vision Transformer (ViT-Base/16) architecture as its encoder. The model consists of 12 transformer encoder blocks, each equipped with multi-head self-attention and feed-forward layers, and operates on patch embeddings of dimension 768. The design provides a global receptive field, enabling context aggregation across anatomically distant regions within the endoscopic frame. The trained encoder can be easily integrated as a drop-in replacement for existing convolutional backbones, providing transformer-based representations for downstream tasks such as semantic segmentation, action recognition, or surgical phase classification. The pre-training methodology is based on the Masked Autoencoder paradigm, according to which each image is divided into non-overlapping patches and a large proportion of them is masked. The model has to reconstruct the missing patches while optimising a mean-squared error (MSE) loss. To adapt MAE to endoscopic imagery, three key innovations were introduced: the first is the **layer-wise learning rate decay** [35] in which learning rates are progressively decreased for deeper layers, with those closer to the latent space receiving greater updates, in order to stabilize training and prevent overfitting; the second is the **stochastic weight averaging (SWA)** [36]: during the final five pretraining epochs, model weights are averaged at each validation step to achieve smoother optimization landscapes and improved generalization; the last is **frequent evaluation**: performance is monitored six times per epoch, with the best SWA model selected for downstream transfer.

4.3.2 PRE-TRAINING SETUP

Pretraining is performed for 15 epochs using the AdamW optimizer with a learning rate of 1.5×10^{-3} and a batch size of 256. The schedule consists of three linear warmup epochs followed by cosine decay until epoch 10 and a constant rate from

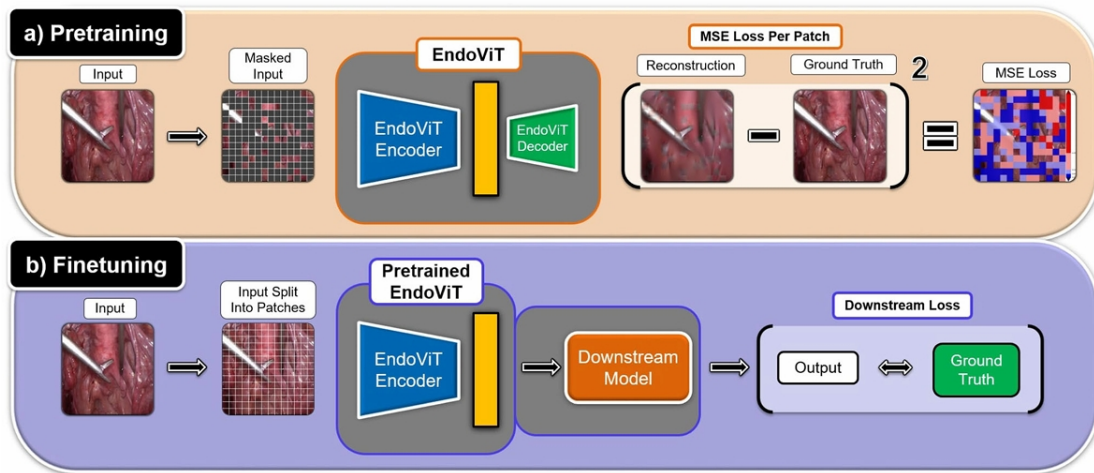


Figure 4.3: Pretraining and finetuning pipeline for EndoViT. During pretraining, masked video patches are reconstructed using an encoder–decoder architecture trained with an MSE loss. The pretrained encoder is then finetuned on downstream tasks by attaching a task-specific head.

that point on. Only simple augmentations—random resized crops and horizontal flips—are applied to maintain data realism. After pretraining, the EndoViT encoder was evaluated across three representative downstream tasks—semantic segmentation, action triplet recognition, and surgical phase recognition—using the Cholec80 dataset and its subvariants (CholecSeg8k and CholecT45). Across all downstream tasks, EndoViT consistently outperformed both convolutional and ImageNet-pretrained transformer baselines, especially under few-shot learning conditions. The benefits were most pronounced in complex, spatially detailed tasks—such as segmentation—where pixel-level reconstruction during pretraining aligns closely with the target task. The results confirm that domain-specific MAE pretraining captures semantically relevant structural cues in endoscopic imagery, enabling superior data efficiency and faster convergence. Qualitative analyses further showed that EndoViT generates more globally coherent segmentation masks and accurately reconstructs fine structures such as instrument tips and tissue boundaries—key regions that CNN-based backbones often miss. EndoViT’s success demonstrates that self-supervised pretraining on large, domain-specific corpora can close

the gap between generic and medical computer vision. By leveraging Endo7ook, it avoids the domain shift inherent in ImageNet pretraining and provides a general-purpose endoscopic encoder that can serve as the foundation for subsequent models such as EndoFM and SurgeNet. Moreover, the release of EndoViT’s pretrained weights and the Endo7ook dataset represents an important milestone for the surgical AI community, enabling reproducible research and further exploration of masked image modeling in medical domains.

4.4 SURGENET

SurgeNet [37] is a self-supervised method that follows the DINO paradigm. Like EndoFM and EndoFM-LV, the model leverages multi-view consistency between augmented crops of the same frame to enforce alignment of visual embeddings across scales and perspectives. It is trained on **SurgeNetXL**, which includes complete surgical videos spanning 23 distinct procedures, from robotic and laparoscopic operations to gynecologic and gastrointestinal surgeries. The dataset amounts to more than 4.7 million frames, collected from both institutional clinical recordings and publicly available surgical videos on YouTube. Its variants can either be *procedure-specific* or vary in size (e.g. the **SurgeNet** or **SurgeNet Small** datasets).

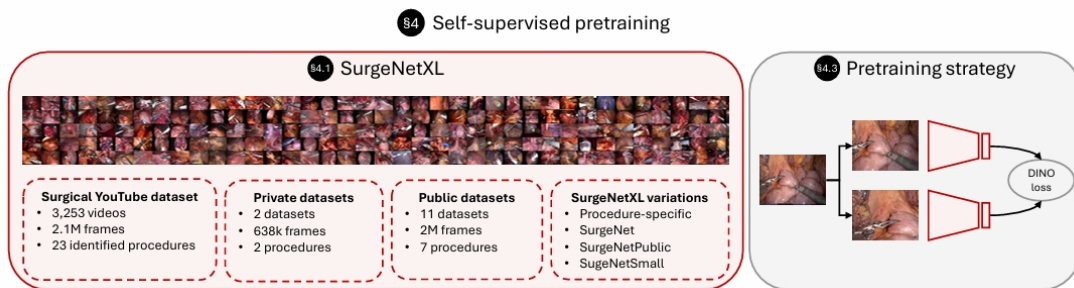


Figure 4.4: Overview of the SurgeNetXL self-supervised pretraining pipeline. The model is trained on a large collection of surgical videos sourced from YouTube, private and public datasets, using a DINO-style self-distillation objective.

During pretraining, two parallel networks (teacher and student) process different

augmentations of the same input frame. The student network is optimized to match the teacher’s output probability distribution, while the teacher parameters are updated via an exponential moving average of the student weights. This strategy allows the model to progressively stabilize and sharpen its representation space without explicit supervision. The resulting embeddings capture both high-level semantic consistency and low-level structural detail relevant to surgical tools and tissue regions.

To explore the architectural trade-offs of self-supervised pretraining in the surgical domain, three backbone families were evaluated: **ConvNeXtV2-Tiny** [38] (CNN-based); **PVTv2-B2** [39] (Transformer-based) and **CAFormer-S18** [40], a hybrid architecture combining convolutional feature extraction with transformer-based self-attention.

Models are initialized with ImageNet weights and are trained with batch sizes of 544. Optimization is performed with AdamW [41] and a cosine learning rate schedule, using the DINO objective for 50 epochs, even though ablation studies indicate that performance continues to improve even beyond 50 epochs.

To assess the generality of the learned representations, the pretrained SurgeNet models are fine-tuned on six downstream datasets, covering three main task categories: semantic segmentation of organs, instruments, and surgical landmarks; phase recognition for workflow analysis and temporal context understanding; critical View of safety (CVS) classification, a clinically relevant task that evaluates whether key biliary structures are sufficiently exposed during cholecystectomy. Ablation experiments further reveal that procedure-specific pretraining (e.g., on only laparoscopic cholecystectomy videos) benefits related datasets but limits transferability. In contrast, the multi-procedure pretraining strategy adopted by SurgeNet yields stronger generalization to unseen surgical contexts. Moreover, t-SNE visualizations of the learned embeddings reveal distinct and well-separated clusters corre-

sponding to surgical procedures and tool types, confirming that the model captures meaningful semantic structure even in the absence of labels.

Conceptually, SurgeNet can be viewed as the natural evolution of the EndoFM family of models. While Endo-FM and EndoFM-LV pioneered temporal modeling and masked-token learning for gastrointestinal endoscopy, SurgeNet expands this paradigm to multi-procedure surgical contexts, frame-level self-distillation, and cross-domain representation learning. The underlying principle remains the same—leveraging massive amounts of unlabeled surgical video through teacher–student SSL—but SurgeNet significantly scales both the breadth (dataset diversity) and depth (training duration, backbone exploration) of this approach.

5

Results

This chapter presents an in-depth analysis of the performance of the proposed models across the four evaluation tasks. It begins by defining the metrics adopted for each task, followed by a detailed description of the experimental setup and the corresponding results for each case. Beyond the performance of the pretrained representations, we further examine how fine-tuning EndoFM with different objectives—supervised, temporally-aware, and intra–inter contrastive—affects downstream behavior. The fine-tuned models are evaluated on the same four tasks to quantify the benefits and limitations of each loss formulation.

5.1 PERFORMANCE METRICS

5.1.1 REID

The performance of the models on ReID is assessed using metrics such as the **Area Under the Receiver Operating Characteristic Curve (AUROC)** and the **Area Under the Precision–Recall Curve (AUPR)**.

The AUROC measures the model’s ability to discriminate between positive and negative pairs across all possible decision thresholds. It is obtained by plotting the *True Positive Rate (TPR)* against the *False Positive Rate (FPR)*, both defined as:

$$TPR = \frac{TP}{TP + FN} \quad FPR = \frac{FP}{FP + TN} \quad (5.1)$$

and computing the area under this curve. An AUROC value of 1 indicates perfect discrimination, whereas a value of 0.5 corresponds to random guessing.

The AUPR, on the other hand, focuses on the trade-off between *precision* (the proportion of correctly identified positives among all *predicted* positives) and *recall* (the proportion of correctly identified positives among all *actual* positives). The area under the PR curve summarizes how well the model maintains high precision as recall increases or, in other words, it measures how well the model is at catching true positives without having too many false positives. Together, AUROC and AUPR provide complementary perspectives on a model’s discriminative capability and robustness to threshold selection.

5.1.2 RETRIEVAL

The evaluation of the retrieval task relies on ranking-based metrics that quantify how effectively the learned embeddings support identity-level search. Specifically, we report the **Hit Rate at k (HR@ k)** and the **Mean Average Precision (mAP)**,

two complementary measures commonly used in large-scale retrieval. $HR@k$ computes the proportion of queries for which at least one correct match appears among the top- k retrieved items, thereby capturing the likelihood that a query’s true counterpart is retrieved within the first few ranked results. In our experiments, we report both $HR@1$ and $HR@5$, which respectively indicate perfect top-1 matching and retrieval performance within the top five candidates. The Mean Average Precision provides a more global assessment of the ranking quality by averaging the precision obtained at each correct retrieval position across all queries. Unlike $HR@k$, which focuses on the presence of a correct match within a cutoff, mAP accounts for the entire ranking order and the number of relevant items retrieved.

5.1.3 COUNTING

In the context of polyp counting, performance is assessed in terms of how accurately the method groups tracklets belonging to the same polyp entity while avoiding incorrect associations. Two metrics are used in this case: the **Fragmentation Rate (FR)** and the **False Positive Rate (FPR)**.

The Fragmentation Rate measures how effectively the algorithm re-associates tracklets corresponding to the same physical lesion. It is defined as the average number of tracklets into which each polyp entity is split after clustering. Formally, if R denotes the set of re-associated tracklets and E the set of true polyp entities, the fragmentation rate is computed as:

$$FR = \frac{|R|}{|E|} \quad (5.2)$$

A perfectly consistent clustering, where each polyp corresponds to a single cluster, achieves $FR = 1$, whereas higher values indicate over-fragmentation due to under-association of tracklets belonging to the same polyp. This metric captures the completeness of re-identification and is averaged across videos to ensure a balanced eval-

uation, independent of the number of polyps per video. However, FR alone does not penalize incorrect associations, i.e., cases where tracklets from different polyps are erroneously merged into the same cluster. For this reason, we also employ the False Positive Rate, which quantifies the proportion of incorrect tracklet-to-polyp associations. Specifically, the FPR is defined as the fraction of re-associated tracklets that are assigned to the wrong polyp entity among all non-matching pairs. In practice, clustering hyperparameters are tuned on the validation set to achieve a fixed target FPR (i.e., 5%) and the corresponding FR at that threshold is reported as the final measure of counting performance. We also employ precision and recall as standard metrics.

5.1.4 TRACKING

Finally, to quantitatively assess the performance of the polyp tracking task, we adopt the **Higher Order Tracking Accuracy (HOTA)** and the **IDF1** metrics, both of which provide a principled evaluation of identity consistency over time. Unlike frame-level detection scores, these metrics focus on how accurately the tracker maintains the association between detections and their corresponding polyp identities across consecutive frames.

The **HOTA** metric decomposes tracking performance into three interpretable components: **Association Accuracy (AssA)**, **Association Precision (AssPr)**, and **Association Recall (AssRe)**. Association Accuracy measures the overall quality of temporal associations between detections belonging to the same ground-truth identity, combining the effects of precision and recall into a single, balanced term. Association Precision quantifies how many of the predicted identity links correspond to correct associations, while Association Recall measures how many of the true identity links were successfully recovered by the tracker. These association metrics operate at the level of object trajectories rather than individual frames, making

them particularly suitable for medical video analysis where each lesion corresponds to a temporally coherent entity.

Complementary to HOTA, the **IDF_I** score is defined as

$$IDF_I = \frac{2IDTP}{2IDTP + IDFP + IDFN} \quad (5.3)$$

where *IDTP*, *IDFP* and *IDFN* represent the Identity True Positives, False Positives and False Negatives, respectively. It evaluates how consistently the tracker assigns detections to the correct identity labels throughout the sequence. In contrast to frame-based precision and recall, IDF_I considers the temporal continuity of identity assignments, penalizing both identity switches and missed associations. In the context of polyp tracking, high IDF_I and HOTA scores jointly indicate that the model not only detects lesions accurately but also preserves their identity through time, ensuring that each polyp is tracked as a distinct, persistent entity across the colonoscopy video.

5.2 PERFORMANCE ANALYSIS – BASELINE EVALUATION

We evaluate all models under two complementary temporal settings: variable-length tracklets and fixed-length temporal fragments. In the first setting, each tracklet is processed in its entirety, allowing the embedding to capture the natural variability in duration that arises from differences in polyp visibility, camera motion, and procedural dynamics. Because tracklets vary in length, the dimensionality of intermediate representations may also differ across samples. In addition to this realistic, procedure-driven scenario, we also conduct controlled experiments using uniform temporal windows. All models are evaluated on fragments consisting of 8 consecutive frames, while EndoFM-LV—whose pretraining relies on longer temporal

contexts—is also tested with 32-frame fragments. These fixed-length evaluations allow us to isolate the effect of temporal window size and to assess how each architecture performs under standardized temporal conditions, independent of tracklet duration.

For the re-identification and retrieval tasks, all frames are resized to 224×224 pixels. Additional experiments are conducted on spatially cropped regions, where each bounding box is uniformly scaled by a factor $\varphi \in \{1, 2, 3, 5, 10\}$ to quantify the impact of spatial context on embedding quality. Tables 5.1 - 5.2 report the ReID results obtained using variable-length tracklets across the 19 videos of the validation and test splits of the REAL-Colon dataset, while Tables 5.3 - 5.5 present the corresponding results under fixed fragment-length settings.

Table 5.1: ReID results across crop factors on REAL-Colon for EndoFM and EndoFM-LV with variable length temporal fragments.

EndoFM			EndoFM-LV		
Crop factor	AUPR	AUROC	Crop factor	AUPR	AUROC
Full Size	0.342	0.803	Full Size	0.213	0.669
$\varphi = 1$	0.240	0.730	$\varphi = 1$	0.170	0.644
$\varphi = 2$	0.295	0.763	$\varphi = 2$	0.176	0.664
$\varphi = 3$	0.321	0.783	$\varphi = 3$	0.182	0.675
$\varphi = 5$	0.334	0.795	$\varphi = 5$	0.193	0.676
$\varphi = 10$	0.343	0.806	$\varphi = 10$	0.205	0.668

A comparison between variable-length tracklets and fixed-length fragments reveals notable differences across models. EndoFM exhibits greater performance variability when representations are computed over the entire tracklet. In this setting, its ReID accuracy decreases for tightly cropped images but improves when more spatial context is available. Conversely, when restricted to fixed 8-frame fragments, EndoFM produces markedly more stable results, showing little sensitivity to crop scale. EndoFM-LV benefits more consistently from longer fragments, achieving

Table 5.2: ReID results across crop factors on REAL-Colon for EndoViT and SurgeNet with variable length temporal fragments.

EndoViT			SurgeNet		
Crop factor	AUPR	AUROC	Crop factor	AUPR	AUROC
Full Size	0.262	0.685	Full Size	0.332	0.792
$\varphi = 1$	0.133	0.650	$\varphi = 1$	0.137	0.746
$\varphi = 2$	0.194	0.706	$\varphi = 2$	0.269	0.800
$\varphi = 3$	0.220	0.709	$\varphi = 3$	0.326	0.809
$\varphi = 5$	0.229	0.701	$\varphi = 5$	0.339	0.810
$\varphi = 10$	0.241	0.684	$\varphi = 10$	0.331	0.801

stronger performance when the entire temporal sequence is used rather than short, fixed-length clips. For EndoViT and SurgeNet, the trend is similar to that observed for EndoFM: both models perform better with increasing spatial context and degrade under aggressive cropping, but they also display greater variability under the 8-frame setting, suggesting a stronger dependence on temporal redundancy or longer-range dynamics.

Focusing on the variable length fragments, we can immediately highlight the surprising behavior of EndoFM-LV, which—despite being conceived as an improved version of EndoFM through the addition of masked token modeling and longer temporal pretraining—consistently underperforms not only its predecessor but also the other models considered. Interestingly, a recurring pattern emerges: all architectures that incorporate masked token modeling—EndoFM-LV and EndoViT—, achieve lower re-identification performance compared to those relying purely on self-distillation. This suggests that, while masked modeling objectives improve generalization and robustness in global scene understanding, they may suppress the fine-grained discriminative cues required to distinguish visually similar polyps.

As for retrieval, a different pattern emerges in the retrieval task. Across all architectures, performance is consistently higher when embeddings are computed from

Table 5.3: ReID results across crop factors on REAL-Colon for EndoFM when fragment length is fixed to 8.

EndoFM		
Crop factor	AUPR	AUROC
Full Size	0.287	0.709
$\phi = 1$	0.287	0.788
$\phi = 2$	0.290	0.731
$\phi = 3$	0.285	0.722
$\phi = 5$	0.284	0.719
$\phi = 10$	0.285	0.709

Table 5.4: ReID results across crop factors on REAL-Colon for EndoFM-LV with fragment length fixed to 8 and 32.

EndoFM-LV, fragm. length = 8			EndoFM-LV, fragm. length = 32		
Crop factor	AUPR	AUROC	Crop factor	AUPR	AUROC
Full Size	0.164	0.582	Full Size	0.198	0.578
$\phi = 1$	0.145	0.593	$\phi = 1$	0.178	0.587
$\phi = 2$	0.161	0.605	$\phi = 2$	0.193	0.609
$\phi = 3$	0.164	0.612	$\phi = 3$	0.195	0.618
$\phi = 5$	0.163	0.607	$\phi = 5$	0.199	0.614
$\phi = 10$	0.163	0.590	$\phi = 10$	0.200	0.588

fixed 8-frame fragments rather than from full variable-length tracklets. When using the entire tracklet, all models exhibit a decrease in performance and more variability across spatial crop scales (Tables 5.6 - 5.7). On the other hand, short, uniform clips appear to provide more stable and discriminative representations for ranking all views of the same polyp, likely because they reduce temporal heterogeneity across queries and gallery samples (Tables 5.8 - 5.10). Furthermore, we observe once again the unexpected underperformance of EndoFM-LV, whose retrieval metrics are consistently lower than those of both its predecessor EndoFM and the other models. The model nonetheless mirrors the behavior of the others, with its strongest results obtained when only short temporal fragments are provided. The same trend characterizes all models trained with a masked token mod-

Table 5.5: ReID results across crop factors on REAL-Colon for EndoViT and SurgeNet with fragment length fixed to 8.

EndoViT			SurgeNet		
Crop factor	AUPR	AUROC	Crop factor	AUPR	AUROC
Full Size	0.218	0.589	Full Size	0.198	0.578
$\varphi = 1$	0.165	0.640	$\varphi = 1$	0.178	0.587
$\varphi = 2$	0.215	0.668	$\varphi = 2$	0.193	0.609
$\varphi = 3$	0.238	0.671	$\varphi = 3$	0.195	0.618
$\varphi = 5$	0.241	0.653	$\varphi = 5$	0.199	0.614
$\varphi = 10$	0.226	0.609	$\varphi = 10$	0.200	0.588

Table 5.6: Retrieval results across crop factors on REAL-Colon with variable length temporal fragments.

EndoFM				EndoFM-LV			
Crop factor	HR@1	HR@5	mAP	Crop factor	HR@1	HR@5	mAP
Full Size	0.891	0.969	0.510	Full Size	0.700	0.867	0.350
$\varphi = 1$	0.784	0.925	0.383	$\varphi = 1$	0.525	0.762	0.241
$\varphi = 2$	0.834	0.931	0.414	$\varphi = 2$	0.515	0.746	0.243
$\varphi = 3$	0.853	0.928	0.436	$\varphi = 3$	0.527	0.76	0.256
$\varphi = 5$	0.876	0.947	0.458	$\varphi = 5$	0.632	0.813	0.282
$\varphi = 10$	0.894	0.965	0.492	$\varphi = 10$	0.655	0.829	0.318

eling objective which generally lag behind the EndoFM and SurgeNet. The overall performance with both settings shows an improvement in performance when more context is given, as in ReID.

For the polyp counting task, each model was evaluated on full size images and whole tracklets, without additional cropping, using five distinct clustering algorithms: **Temporal Affinity Propagation (T-AP)**, **Threshold-based Clustering**, **Affinity Propagation (AP)**, **DBSCAN**, and **HDBSCAN**. These methods were selected to assess the robustness of the learned embeddings under different grouping paradigms. Affinity-based approaches such as AP and T-AP operate by iteratively exchanging messages between data points to identify representative *exem-*

Table 5.7: Retrieval results across crop factors on REAL-Colon with variable length temporal fragments.

EndoViT				SurgeNet			
Crop factor	HR@1	HR@5	mAP	Crop factor	HR@1	HR@5	mAP
Full Size	0.878	0.955	0.482	Full Size	0.955	0.996	0.499
$\varphi = 1$	0.669	0.853	0.290	$\varphi = 1$	0.762	0.893	0.349
$\varphi = 2$	0.730	0.871	0.321	$\varphi = 2$	0.821	0.910	0.381
$\varphi = 3$	0.754	0.885	0.351	$\varphi = 3$	0.848	0.938	0.428
$\varphi = 5$	0.787	0.920	0.382	$\varphi = 5$	0.887	0.959	0.48
$\varphi = 10$	0.832	0.935	0.424	$\varphi = 10$	0.902	0.974	0.528

Table 5.8: Retrieval results across crop factors on REAL-Colon with fragment length fixed to 8.

EndoFM			
Crop factor	HR@1	HR@5	mAP
Full Size	0.955	0.997	0.467
$\varphi = 1$	0.943	0.977	0.384
$\varphi = 2$	0.964	0.985	0.404
$\varphi = 3$	0.977	0.991	0.423
$\varphi = 5$	0.982	0.994	0.437
$\varphi = 10$	0.973	0.996	0.458

plars, with the temporal variant introducing an additional constraint that penalizes associations between tracklets that are far apart in time, thereby improving temporal consistency in video sequences. The Threshold-based method instead groups tracklets according to distances not exceeding a fixed threshold. Density-based methods, namely DBSCAN and HDBSCAN, define clusters as regions of high sample density separated by sparse regions, automatically rejecting outliers. While DBSCAN relies on manually set parameters such as distance radius and minimum cluster size, HDBSCAN extends it with a hierarchical formulation that adapts to variable density, providing improved robustness to heterogeneous data distributions. Collectively, these algorithms enable a comprehensive evaluation of the discriminative quality and temporal coherence of the embeddings produced by

Table 5.9: Retrieval results across crop factors on REAL-Colon with fragment length fixed to 8 and 32.

EndoFM-LV, f. l. = 8				EndoFM-LV, f. l. = 32			
Crop factor	HR@1	HR@5	mAP	Crop factor	HR@1	HR@5	mAP
Full Size	0.869	0.962	0.330	Full Size	0.788	0.931	0.391
$\varphi = 1$	0.808	0.917	0.269	$\varphi = 1$	0.680	0.845	0.322
$\varphi = 2$	0.818	0.921	0.278	$\varphi = 2$	0.704	0.841	0.333
$\varphi = 3$	0.830	0.929	0.288	$\varphi = 3$	0.722	0.854	0.337
$\varphi = 5$	0.858	0.940	0.299	$\varphi = 5$	0.761	0.894	0.357
$\varphi = 10$	0.875	0.951	0.314	$\varphi = 10$	0.763	0.905	0.379

Table 5.10: Retrieval results across crop factors on REAL-Colon with fragment length fixed to 8.

EndoViT				SurgeNet			
Crop factor	HR@1	HR@5	mAP	Crop factor	HR@1	HR@5	mAP
Full Size	0.945	0.991	0.438	Full Size	0.955	0.996	0.499
$\varphi = 1$	0.891	0.955	0.319	$\varphi = 1$	0.931	0.975	0.359
$\varphi = 2$	0.916	0.965	0.342	$\varphi = 2$	0.953	0.981	0.387
$\varphi = 3$	0.936	0.976	0.360	$\varphi = 3$	0.964	0.987	0.423
$\varphi = 5$	0.951	0.984	0.381	$\varphi = 5$	0.978	0.993	0.456
$\varphi = 10$	0.964	0.987	0.410	$\varphi = 10$	0.976	0.995	0.490

each model in the context of lesion-level counting.

When clustering tracklets with Temporal Affinity Propagation, the overall results are consistent across all models, with EndoViT achieving the highest false positive rate (Fig. 5.1) but also the lowest fragmentation rate (Fig. 5.2), and EndoFM showing the inverse trend, indicating that its embeddings are more discriminative but less permissive in associating tracklets belonging to the same lesion. EndoFM-LV performs comparably to SurgeNet in terms of fragmentation, although it presents a slightly higher FPR and a notably lower recall. In general, models that rely purely on self-distillation, such as EndoFM and SurgeNet, tend to produce fewer false associations, whereas those using masked modeling, such as EndoFM-LV and En-

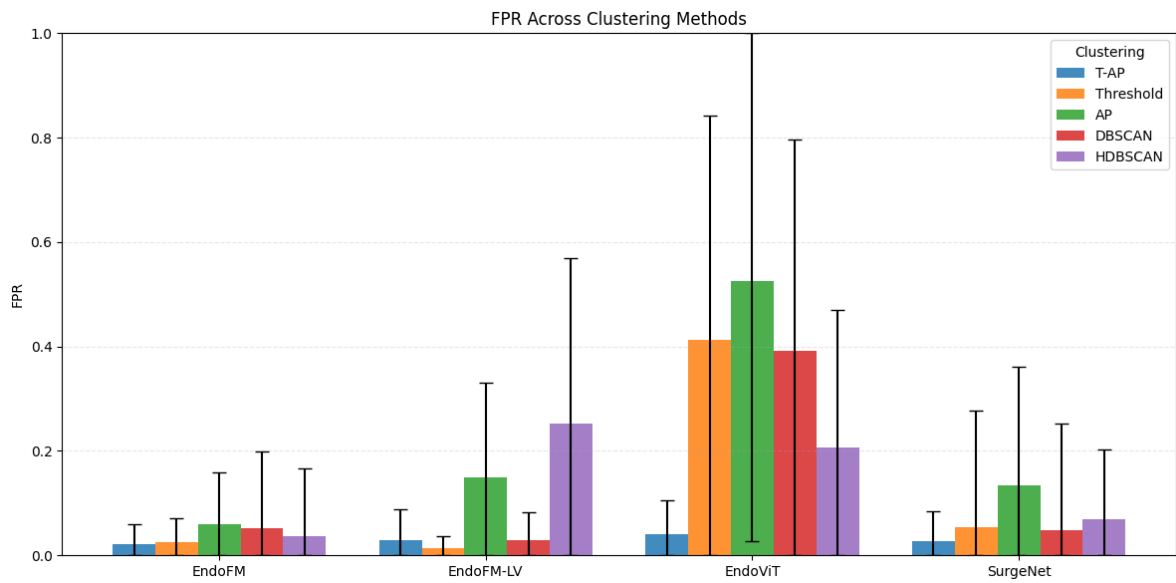


Figure 5.1: False Positive Rate across different models.

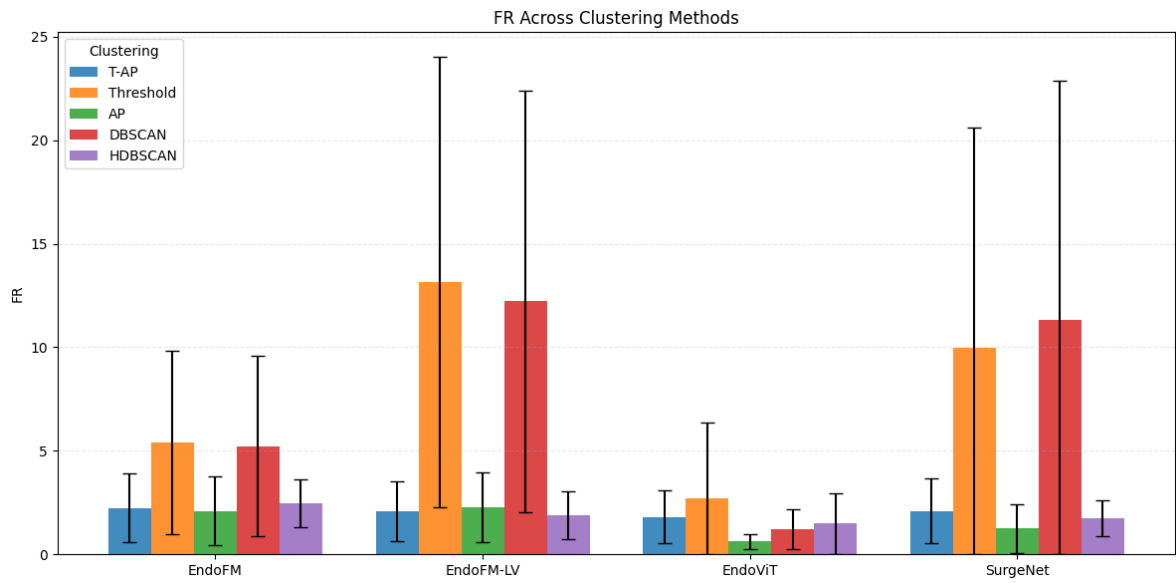


Figure 5.2: Fragmentation Rate across different models.

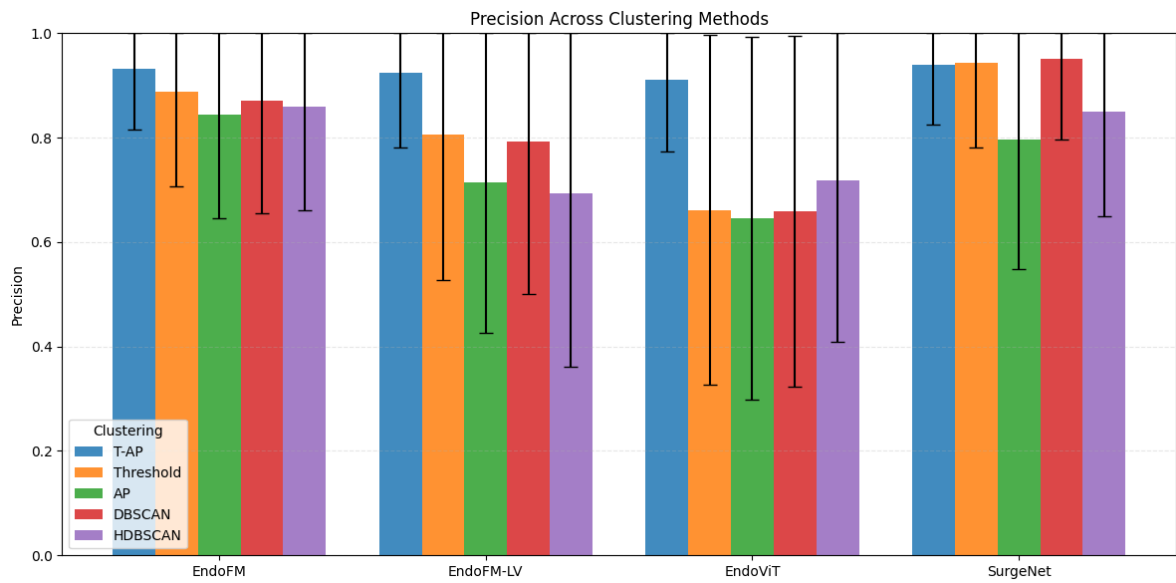


Figure 5.3: Precision across different models.

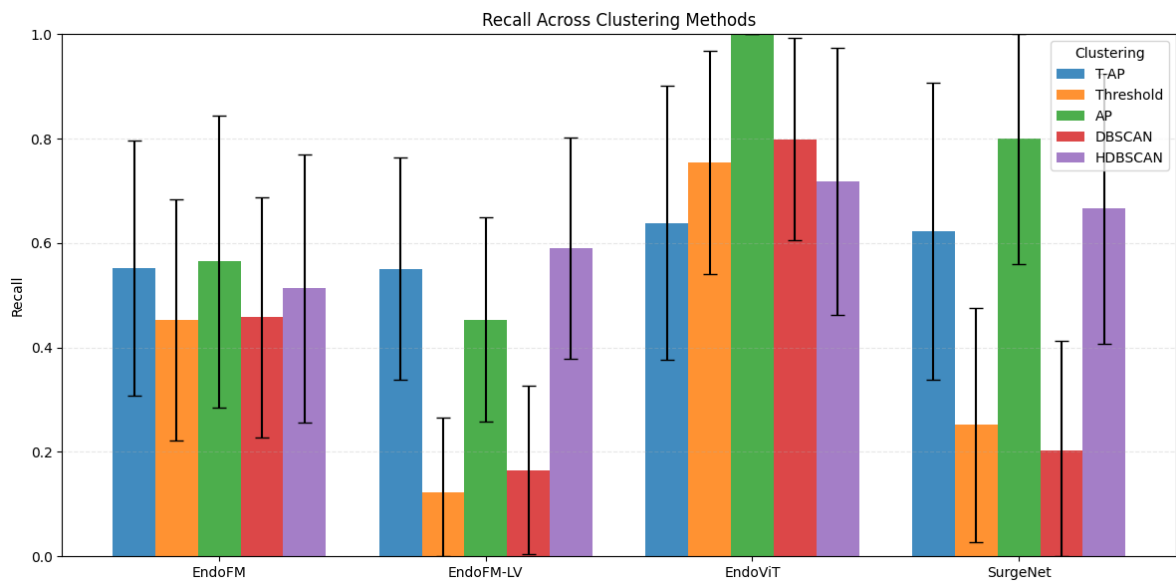


Figure 5.4: Recall across different models.

doViT, generate broader, context-rich embeddings that enhance continuity at the cost of over-merging.

This trend remains evident across the remaining clustering algorithms. With the Threshold based method, EndoFM, EndoFM-LV, and SurgeNet yield high FR but low FPR, while EndoViT again exhibits the opposite behavior, confirming its tendency to merge visually similar tracklets. Affinity Propagation produces results largely consistent with the temporal variant, though EndoViT’s performance diverges, showing the lowest FR and the highest FPR, emphasizing the importance of temporal regularization in mitigating excessive associations. DBSCAN behaves similarly to the threshold-based approach: EndoViT reports a fragmentation rate below one again—indicating that the number of predicted clusters is smaller than the actual number of lesions—at the expense of a very high FPR, exceeding 50%, as its density-based formulation often merges distinct tracklets located in dense regions of the embedding space. Finally, HDBSCAN yields stable results across all models, with comparable FR values and moderate sensitivity to embedding variations; nonetheless, EndoFM-LV and EndoViT still display higher FPRs than EndoFM and SurgeNet, confirming their inclination toward over-clustering.

Figures 5.5 - 5.8 report the quantitative results of the polyp tracking evaluation using the best hyperparameters for each one of the five clustering algorithms. Across most settings, the temporal and affinity-based methods tend to outperform the density-based ones, highlighting the importance of temporal and relational cues for maintaining consistent associations between tracklets. Among the evaluated models, EndoFM manages to keep a moderately high association precision across most algorithms at the expense of a reduced recall. As expected from the performance of EndoFM-LV on the counting task, the model is able to keep efficiently track of the polyps when using Temporal Affinity Propagation but experiences a drop in performance when the Threshold-based method or DBSCAN are involved.

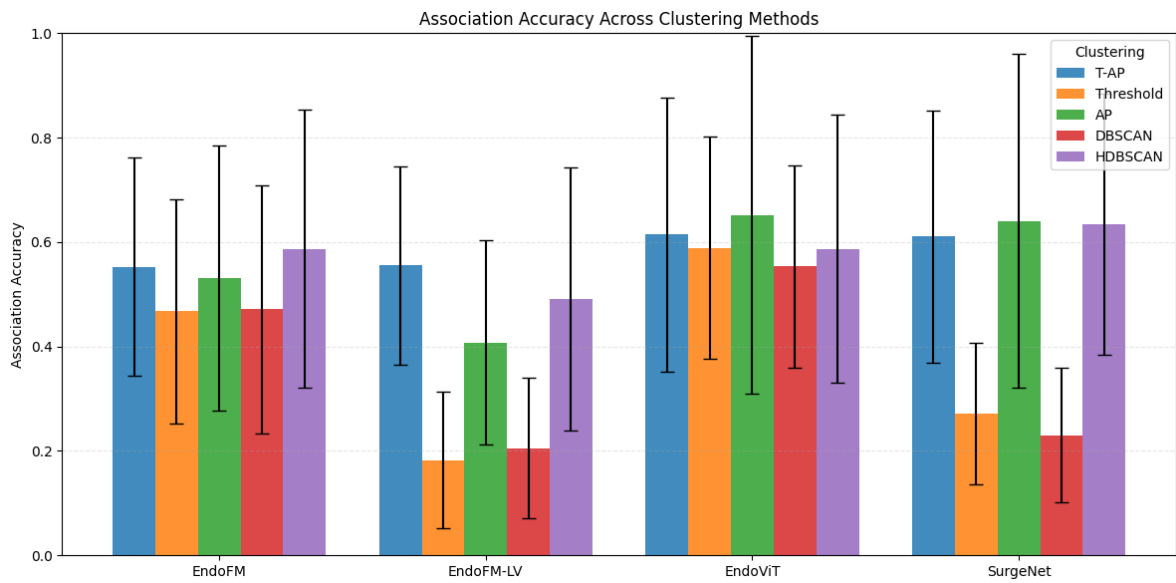


Figure 5.5: Association Accuracy across different models.

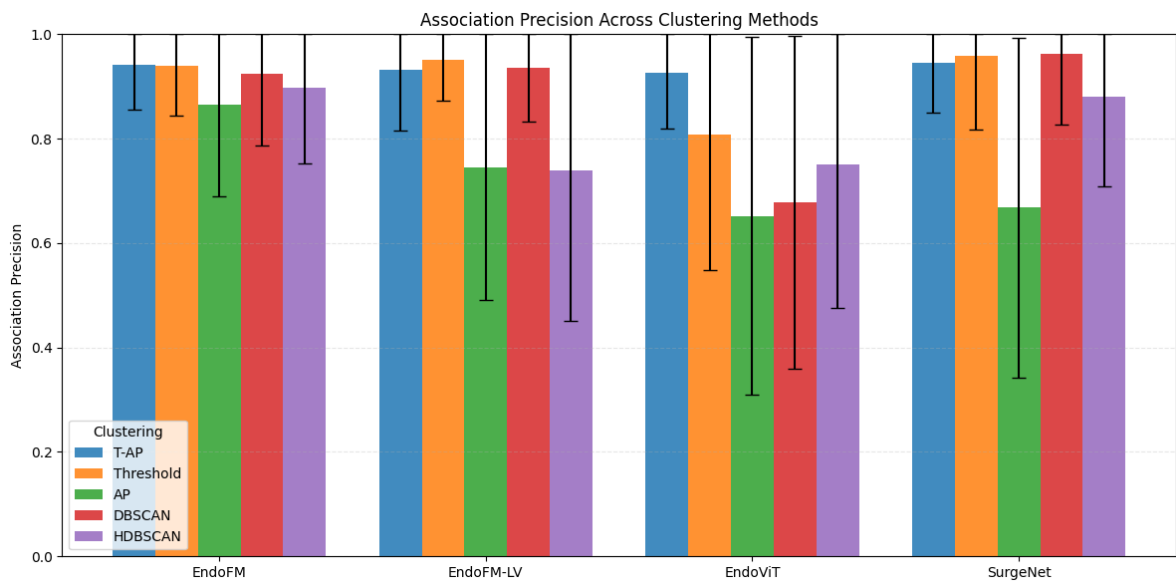


Figure 5.6: Association Precision across different models.

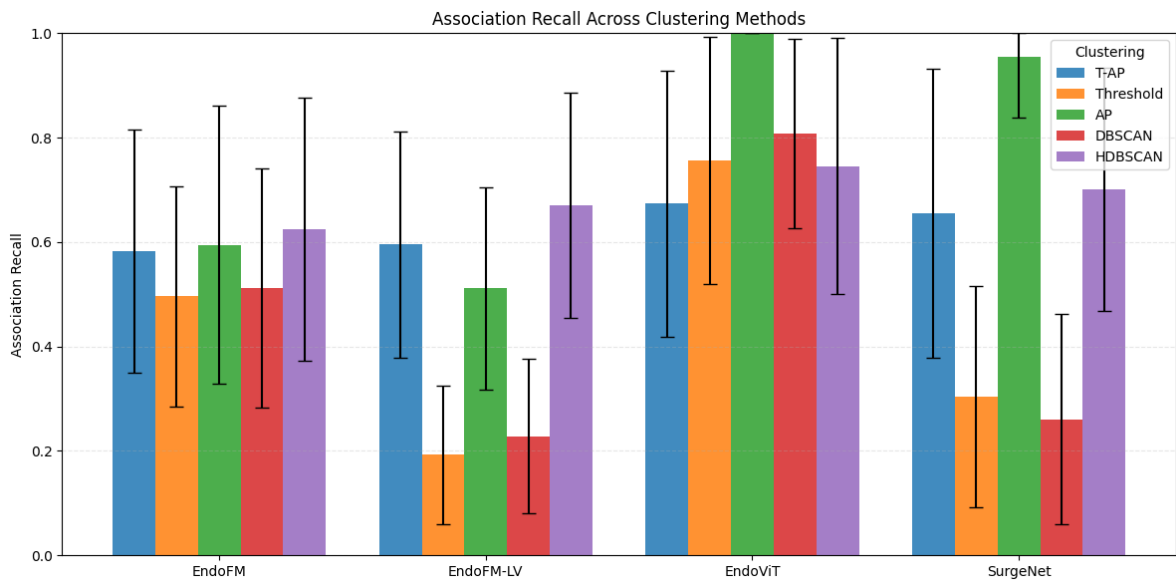


Figure 5.7: Association Recall across different models.

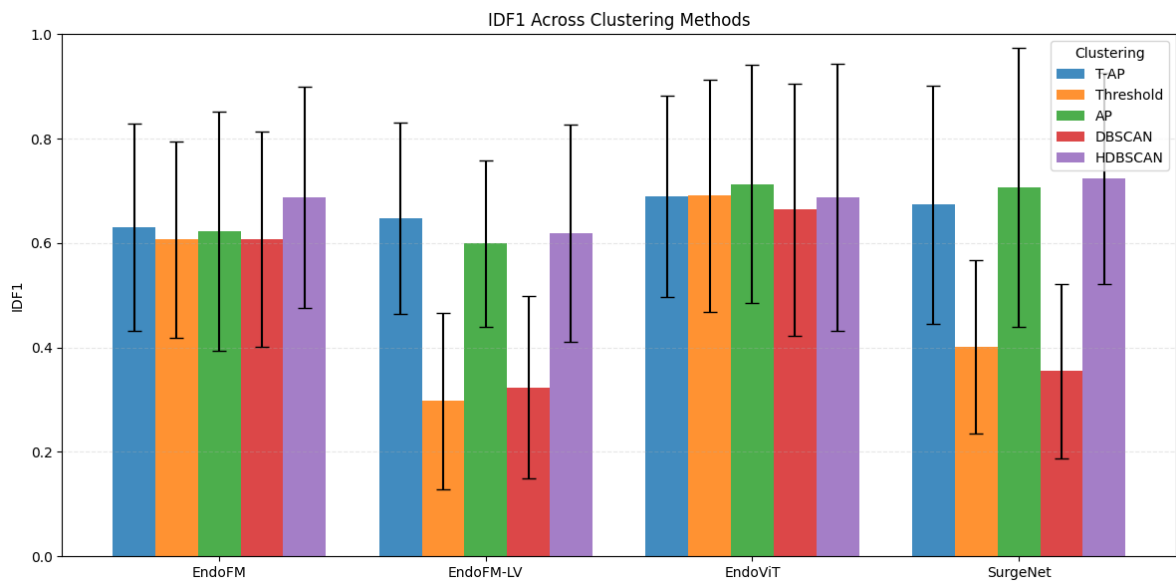


Figure 5.8: IDF1 across different models.

Both Affinity Propagation and HDBSCAN yield intermediate results. EndoViT notably achieves a recall equal to 1 when using Affinity Propagation and more balanced metrics with the other algorithms. SurgeNet displays a consistently good behavior with the affinity-based methods and with HDBSCAN as well, while its performance decreases with the remaining two methods.

When comparing the clustering algorithms, Temporal Affinity Propagation provides the most stable and balanced performance across all models, reflecting the performance seen in the counting task, with consistently high precision and moderate recall, while standard Affinity Propagation yields nearly equivalent results, in terms of accuracy, to the temporal variant but tends to yield higher recalls according to the type of model employed. The Threshold method leads to lower accuracy and recall, indicating that a fixed similarity cutoff struggles to generalize across varying embedding distributions. While DBSCAN yields results that are comparable to those obtained with the Threshold-based method, HDBSCAN manages to improve them across all the four metrics, achieving similar results to the ones obtained with Temporal Affinity Propagation. The overall tracking performance reflects the same inter-model patterns identified in the counting task, suggesting that the models' ability to maintain temporal consistency aligns with their capacity to correctly group tracklets during counting.

5.3 FINE-TUNING AND RE-EVALUATION OF ENDOFM

In this section, we present the fine-tuning of the EndoFM model on the training split of the REAL-Colon dataset, followed by its evaluation across the same four downstream tasks: re-identification, retrieval, counting, and tracking. The objective of this stage is to assess whether adapting EndoFM to the REAL-Colon domain can improve its overall performance across these tasks, compared to the pretrained self-supervised version. To this end, three fine-tuning objectives were

explored in order to understand how different loss formulations can affect the model’s ability to produce consistent and identity-preserving representations suitable for entity-level analysis in colonoscopy videos.

5.3.1 LOSSES

In all cases, tracklet embeddings are L2-normalized, and pairwise similarities are computed via temperature-scaled dot-products:

$$z_{ij} = \frac{f_i^T f_j}{\tau}, \quad q_{ij} = \frac{\exp(z_{ij})}{\sum_{k \neq i} \exp(z_{ik})} \quad (5.4)$$

where $f_i \in R^d$ denotes the embedding of sample i , and τ is a temperature parameter. The diagonal terms are masked to prevent self-similarities from contributing. Let l_i denote the identity label, τ_i the temporal position within the video, and v_i the video index of sample i . All losses minimize a row-wise cross-entropy:

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^B \sum_{\substack{j=1 \\ j \neq i}}^B p_{ij} \log q_{ij}, \quad (5.5)$$

where B denotes the batch size. For the supervised multi-positive contrastive loss, all samples in the batch sharing the same identity label l_i are treated as positives for anchor i . Let

$$P(i) = \{j \neq i \mid l_j = l_i\} \quad (5.6)$$

denote the set of positive indices for sample i . The target distribution assigns equal mass to all positives:

$$p_{ij} = \begin{cases} \frac{1}{|P(i)|}, & \text{if } j \in P(i), \\ 0, & \text{otherwise.} \end{cases} \quad (5.7)$$

Substituting p_{ij} into the general loss definition yields the supervised multi-positive contrastive objective.

To incorporate temporal smoothness and, thus obtain the temporally aware loss formulation, we weight positive pairs according to their temporal distance. Let t_i denote the temporal position of sample i , and $\lambda \geq 0$ a decay parameter. We define an unnormalized weight

$$\tilde{p}_{ij} = \exp(-\lambda |t_i - t_j|) \mathbb{1}[\ell_j = \ell_i, j \neq i], \quad (5.8)$$

and obtain a normalized target distribution via

$$p_{ij} = \frac{\tilde{p}_{ij}}{\sum_{\substack{k=1 \\ k \neq i}}^B \tilde{p}_{ik}}. \quad (5.9)$$

This encourages temporally closer positives to have higher weight in the contrastive objective.

Finally, the intra–inter multi-positive loss combines an intra-class, temporally-weighted term with an inter-class term based on video identity. Let v_i denote the video index of sample i , and $\gamma \geq 0$ a weighting factor for the inter-video term. We define

$$\tilde{p}_{ij} = \underbrace{\exp(-\lambda |t_i - t_j|) \mathbb{1}[\ell_j = \ell_i, j \neq i]}_{\text{intra-class, temporally weighted}} + \underbrace{\gamma \mathbb{1}[v_j = v_i, j \neq i]}_{\text{inter-class, same video}}, \quad (5.10)$$

and normalize row-wise to obtain

$$p_{ij} = \frac{\tilde{p}_{ij}}{\sum_{\substack{k=1 \\ k \neq i}}^B \tilde{p}_{ik}}. \quad (5.11)$$

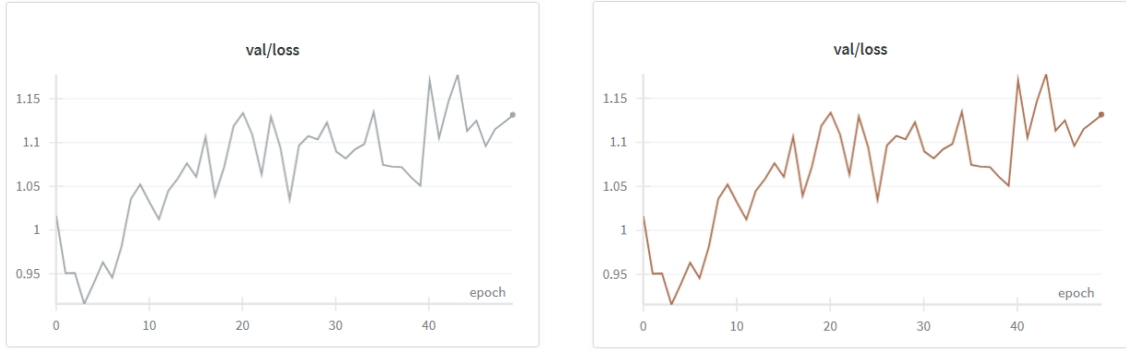
Plugging this distribution into the general loss yields the intra–inter multi-positive contrastive objective.

5.3.2 SETUP

All experiments adopt a consistent optimization setup: the network is fine-tuned for **50 epochs** using a **learning rate** of 10^{-5} and a **weight decay** of 10^{-4} . The **batch size** is set to 8, due to GPU memory limitations. This also constraints the **number of views** to either 2 or 4. **Temporal fragments** are sampled from tracklets with a fixed length of 8 frames and a **temporal stride** of 4 between consecutive fragments. Frames are spatially resized to a **resolution** of 224×224 pixels before feeding them into the network. Since the number of frames in a tracklet is not guaranteed to be a multiple of the fragment length, incomplete trailing fragments are discarded. During validation, the same temporal and spatial configuration is maintained, using an **evaluation batch size** of 16 and a **fixed number of views** equal to 2. This ensures consistency between training and validation data distributions. All contrastive losses employ a **temperature parameter** $\tau = 0.25$. For the temporally-aware objective, the **temporal decay** factor is set to $\lambda = 1$, whereas for the intra-inter formulation we additionally set the **inter-video weight** to $\gamma = 0.2$. These hyperparameters were selected to balance the contribution of temporal smoothness and inter-video regularization without overwhelming the supervised identity signal. The training of the whole network was conducted on a NVIDIA RTX A4000.

5.3.3 RESULTS

When training with only two views per polyp identity, the temporally-aware contrastive objective exhibits a convergence behavior that is indistinguishable from the supervised loss (Fig. 5.9). This effect is expected: with 2 views, each identity provides only a single positive pair per batch. As a consequence, the temporal weighting term introduced in the temporally-aware formulation becomes constant within each pair.



(a) Supervised loss.

(b) Temporally aware loss.

Figure 5.9: Validation loss comparison for the same model trained with different objectives. When using 2 views per polyp, the two training strategies exhibit identical validation loss trends.

In contrast, the intra–inter contrastive loss behaves differently under the same setting, as it additionally leverages inter-video relationships that remain active even when temporal diversity is limited. Nevertheless, for methodological consistency and to ensure that all three objectives are evaluated under conditions where temporal structure is properly represented, we restrict downstream experiments to models trained with four views per identity, where multiple positive temporal relationships can influence the optimization.

When increasing the number of views per polyp to $k = 4$, the temporal structure becomes sufficiently expressive to drive distinct learning dynamics across all three loss objectives. In particular, the temporally-aware loss now benefits from multiple positive temporal relationships, resulting in a steeper and more stable convergence compared to the supervised objective. The intra–inter contrastive formulation further amplifies this effect by leveraging both within-identity temporal consistency and cross-identity dissimilarity, encouraging better separation in the embedding space. Figure 5.10 highlights the different optimization behaviors observed during training. The final classification performance for the three training objectives is reported in Table 5.11. Top-k accuracy evaluates the correctness of the top-ranked retrieval results. Top-1 accuracy measures the proportion of queries for which the

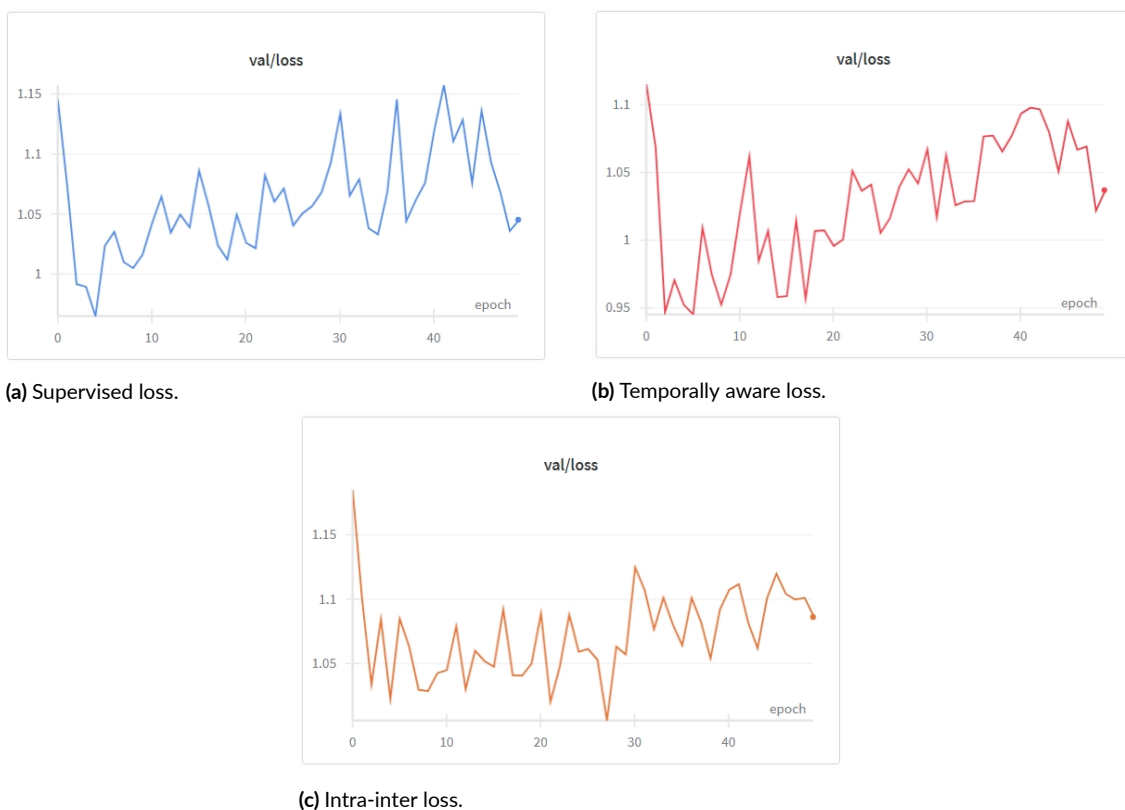


Figure 5.10: Validation loss trends for the three loss objectives with $k = 4$ views per identity.

highest-scoring retrieved tracklet corresponds to the same polyp identity. Top-5 accuracy extends this criterion by considering the first five retrieved candidates, counting a query as correct if at least one of them matches the ground-truth identity. These metrics quantify how reliably the model prioritizes the correct polyp among the most similar retrieved instances, providing an intuitive measure of ranking quality. Metrics are expressed as Top-1 and Top-5 accuracy on both the validation and test sets.

Overall, the three losses achieve comparable results, with the supervised and temporally-aware objectives showing slightly better performance, particularly in Top-1 accuracy.

Re-identification performance. After finetuning EndoFM under the three train-

Table 5.11: Performance comparison of EndoFM finetuning with $k = 4$ views per identity. Metrics are reported as Top-1 and Top-5 classification accuracy on validation and test sets.

Loss Objective	Val Top-1	Val Top-5	Test Top-1	Test Top-5
Supervised	71.81%	99.07%	72.97%	98.18%
Temporally-aware	73.67%	98.87%	72.38%	98.56%
Intra-Inter	70.15%	98.54%	70.99%	98.13%

ing objectives, we evaluate their impact on the four tasks starting with re-identification. Table 5.12 summarizes the resulting AUPR and AUROC values for all loss functions and cropping configurations. For clarity and direct comparison, we also report the performance of the pretrained EndoFM encoder, previously introduced in Section 5.2. This allows quantifying the improvements brought by each finetuning objective under consistent evaluation settings.

Table 5.12: Polyp re-identification performance of EndoFM under different loss objectives. We report AUPR and AUROC for full-frame images and for polyp-centred crops obtained with five different scaling factors. The pretrained EndoFM encoder (no finetuning) is included as a reference baseline.

Training	Metric	Full	Crop scale factor				
			1	2	3	5	10
Pretrained	AUPR	0.342	0.240	0.295	0.321	0.334	0.343
	AUROC	0.803	0.725	0.763	0.783	0.795	0.806
Supervised CE	AUPR	0.385	0.150	0.273	0.344	0.362	0.421
	AUROC	0.939	0.759	0.841	0.880	0.904	0.930
Temporally-aware	AUPR	0.440	0.222	0.307	0.376	0.404	0.462
	AUROC	0.938	0.797	0.835	0.871	0.903	0.930
Intra-Inter	AUPR	0.401	0.249	0.326	0.377	0.382	0.420
	AUROC	0.927	0.809	0.835	0.863	0.889	0.914

Fine-tuning EndoFM consistently improves performance on the ReID task across all image settings, both when using full-frame images and when evaluating on different cropping factors. In most conditions, all three loss functions outperform the pretrained baseline. The only exceptions occur with the supervised objective

under the tightest crops ($\varphi = 1$ and $\varphi = 2$), where the AUPR score slightly degrades, although AUROC still improves, indicating that ranking quality remains beneficial even when precision–recall trade-offs are less favorable.

Among the three objectives, the temporally-aware loss achieves the strongest improvements when more contextual information is present (larger crops), suggesting that temporal continuity better constrains the embedding structure when anatomical surroundings are available. Conversely, the intra–inter contrastive loss performs best on tightly zoomed regions, where suppressing inter-polyp similarity becomes crucial to disambiguate visually similar polyps with limited context.

Retrieval performance. In addition to pairwise re-identification, we evaluate the three training objectives in a retrieval setting, where each query polyp is matched against a gallery of candidate instances. As in the ReID evaluation, we compare full-frame images with polyp-centred crops obtained by scaling the detection bounding boxes with five different factors. The resulting retrieval performance for all loss functions and cropping configurations is summarized in Table 5.13.

In this task, the pretrained EndoFM encoder offers slightly higher performance than the fine-tuned models when images are tightly cropped, indicating that its original representation is already effective at distinguishing polyps based primarily on local texture. However, as more contextual information is retained in the image, the benefits of fine-tuning become increasingly evident. In particular, the intra–inter contrastive objective achieves superior results when zoomed-in regions still contain subtle contextual cues, while the temporally-aware loss yields the most consistent gains under full-frame conditions, where temporal structure and anatomical surroundings jointly contribute to more reliable identity retrieval.

Counting performance. Tables 5.14 - 5.18 compare the performance of the evaluated models in the polyp counting task. The results on the polyp counting task

Table 5.13: Polyp retrieval performance of EndoFM under different loss objectives. We report mAP and Hit Rate at rank 1 and 5 (HR@1, HR@5) for full-frame images and for polyp-centred crops obtained with five different scaling factors. The pretrained EndoFM encoder (no finetuning) is included as a reference baseline.

Training	Metric	Full	Crop scale factor				
			1	2	3	5	10
Pretrained	HR@1	0.891	0.784	0.834	0.853	0.876	0.894
	HR@5	0.969	0.925	0.931	0.928	0.947	0.965
	mAP	0.51	0.383	0.414	0.436	0.458	0.492
Supervised CE	HR@1	0.879	0.717	0.754	0.802	0.847	0.880
	HR@5	0.963	0.876	0.894	0.908	0.932	0.948
	mAP	0.622	0.335	0.391	0.452	0.499	0.577
Temporally-aware	HR@1	0.917	0.770	0.802	0.841	0.874	0.908
	HR@5	0.982	0.902	0.921	0.933	0.952	0.966
	mAP	0.665	0.366	0.411	0.473	0.523	0.607
Intra-inter	HR@1	0.910	0.765	0.834	0.857	0.875	0.900
	HR@5	0.985	0.905	0.933	0.946	0.957	0.977
	mAP	0.663	0.375	0.430	0.485	0.524	0.598

indicate that clustering performance strongly depends on the chosen clustering method. When using Temporal Affinity Propagation, all losses achieve comparable results, with the supervised objective yielding the lowest fragmentation rate but also the highest false positive rate. In contrast, the temporally-aware loss provides the best balance, achieving low fragmentation while keeping false positives under control.

When switching to threshold-based clustering or DBSCAN, the pretrained EndoFM encoder remains the strongest baseline, suggesting that its embedding space is already well structured for simple clustering approaches. Affinity Propagation, however, generally achieves the lowest fragmentation rates across models — even outperforming its temporal variant — although this improvement comes at the cost of a higher false positive rate. Finally, HDBSCAN produces results similar to Temporal Affinity Propagation in terms of fragmentation, but with increased false

positives.

Overall, the temporally-aware fine-tuning strategy proves to be the most effective when a balance between fragmentation and false detection is required, confirming the importance of leveraging temporal consistency for robust polyp identity clustering.

Temporal Affinity Propagation				
Model	FPR	FR	Precision	Recall
Pretrained	0.021 ± 0.038	2.238 ± 1.651	0.932 ± 0.116	0.552 ± 0.244
Supervised	0.031 ± 0.056	1.920 ± 1.296	0.932 ± 0.126	0.617 ± 0.254
Temporally aware	0.027 ± 0.046	2.047 ± 1.474	0.937 ± 0.110	0.588 ± 0.253
Intra inter	0.030 ± 0.056	2.065 ± 1.641	0.936 ± 0.108	0.596 ± 0.258

Table 5.14: Polyp counting metrics using T-AP as clustering algorithm across the analyzed models.

Threshold				
Model	FPR	FR	Precision	Recall
Pretrained	0.025 ± 0.047	5.401 ± 4.444	0.888 ± 0.182	0.453 ± 0.231
Supervised	0.053 ± 0.197	13.863 ± 14.900	0.856 ± 0.281	0.166 ± 0.220
Temporally aware	0.030 ± 0.068	10.818 ± 12.537	0.907 ± 0.179	0.364 ± 0.276
Intra inter	0.040 ± 0.083	10.393 ± 9.451	0.871 ± 0.224	0.269 ± 0.270

Table 5.15: Polyp counting metrics using Threshold-based method as clustering algorithm across the analyzed models.

Affinity Propagation				
Model	FPR	FR	Precision	Recall
Pretrained	0.060 ± 0.100	2.094 ± 1.668	0.845 ± 0.200	0.565 ± 0.280
Supervised	0.114 ± 0.164	1.458 ± 1.474	0.807 ± 0.280	0.686 ± 0.260
Temporally aware	0.080 ± 0.116	1.711 ± 1.718	0.836 ± 0.205	0.701 ± 0.295
Intra inter	0.087 ± 0.114	1.570 ± 1.460	0.820 ± 0.210	0.724 ± 0.296

Table 5.16: Polyp counting metrics using AP as clustering algorithm across the analyzed models.

Tracking performance. In the polyp tracking task, the temporal variant of Affinity Propagation yields the most consistent improvements, with both the temporally-

DBSCAN				
Model	FPR	FR	Precision	Recall
Pretrained	0.052 ± 0.147	5.234 ± 4.373	0.870 ± 0.216	0.458 ± 0.230
Supervised	0.030 ± 0.090	12.604 ± 13.931	0.867 ± 0.259	0.210 ± 0.232
Temporally aware	0.063 ± 0.206	11.637 ± 14.040	0.890 ± 0.227	0.331 ± 0.273
Intra inter	0.063 ± 0.206	11.637 ± 14.040	0.890 ± 0.227	0.331 ± 0.273

Table 5.17: Polyp counting metrics using DBSCAN as clustering algorithm across the analyzed models.

HDBSCAN				
Model	FPR	FR	Precision	Recall
Pretrained	0.038 ± 0.058	2.471 ± 1.164	0.860 ± 0.199	0.513 ± 0.257
Supervised	0.131 ± 0.168	2.119 ± 1.282	0.780 ± 0.262	0.608 ± 0.228
Temporally aware	0.081 ± 0.119	2.233 ± 1.469	0.812 ± 0.238	0.591 ± 0.254
Intra inter	0.158 ± 0.238	1.735 ± 1.144	0.768 ± 0.278	0.652 ± 0.267

Table 5.18: Polyp counting metrics using HDBSCAN as clustering algorithm across the analyzed models.

aware and intra–inter losses outperforming the pretrained baseline as well as the supervised objective. This confirms that temporal structure and inter-identity relationships play a crucial role in stabilizing identity assignments across frames.

In contrast, threshold clustering and DBSCAN favor the pretrained EndoFM encoder, which achieves the best overall tracking accuracy, while the supervised fine-tuning strategy performs the worst across these settings. A similar trend is observed with HDBSCAN, where the pretrained model again provides the strongest performance and the supervised version the weakest. However, for HDBSCAN, the temporally-aware and intra–inter losses still achieve better tracking metrics than those obtained with threshold clustering and DBSCAN, indicating that these objectives produce embeddings that are more compatible with density-based clustering.

Finally, standard Affinity Propagation also favors the temporally-aware and intra–inter embeddings, mirroring the behavior of its temporal variant and reinforcing

the advantage of loss functions that explicitly exploit temporal consistency or contrastive inter-identity separation.

Temporal Affinity Propagation				
Model	ASSA	ASSPR	ASSRE	IDF_I
Pretrained	0.553 ± 0.209	0.942 ± 0.087	0.583 ± 0.233	0.631 ± 0.198
Supervised	0.551 ± 0.231	0.938 ± 0.106	0.648 ± 0.249	0.676 ± 0.209
Temporally aware	0.583 ± 0.206	0.940 ± 0.097	0.625 ± 0.250	0.652 ± 0.202
Intra inter	0.589 ± 0.191	0.944 ± 0.091	0.638 ± 0.250	0.659 ± 0.178

Table 5.19: Polyp tracking metrics using T-AP as clustering algorithm across the analyzed models.

Threshold				
Model	ASSA	ASSPR	ASSRE	IDF_I
Pretrained	0.468 ± 0.215	0.940 ± 0.096	0.496 ± 0.212	0.607 ± 0.188
Supervised	0.198 ± 0.145	0.936 ± 0.171	0.239 ± 0.213	0.294 ± 0.179
Temporally aware	0.368 ± 0.240	0.941 ± 0.117	0.405 ± 0.267	0.497 ± 0.246
Intra inter	0.284 ± 0.221	0.924 ± 0.144	0.336 ± 0.262	0.397 ± 0.237

Table 5.20: Polyp tracking metrics using Threshold-based method as clustering algorithm across the analyzed models.

Affinity Propagation				
Model	ASSA	ASSPR	ASSRE	IDF_I
Pretrained	0.531 ± 0.254	0.864 ± 0.175	0.595 ± 0.266	0.623 ± 0.229
Supervised	0.338 ± 0.251	0.799 ± 0.246	0.704 ± 0.241	0.699 ± 0.236
Temporally aware	0.635 ± 0.280	0.854 ± 0.184	0.729 ± 0.278	0.716 ± 0.235
Intra inter	0.643 ± 0.278	0.844 ± 0.182	0.747 ± 0.276	0.731 ± 0.222

Table 5.21: Polyp tracking metrics using AP as clustering algorithm across the analyzed models.

DBSCAN				
Model	ASSA	ASSPR	ASSRE	IDF_I
Pretrained	0.472 ± 0.238	0.924 ± 0.137	0.511 ± 0.229	0.608 ± 0.206
Supervised	0.258 ± 0.205	0.945 ± 0.132	0.285 ± 0.220	0.370 ± 0.231
Temporally aware	0.320 ± 0.225	0.925 ± 0.180	0.377 ± 0.265	0.447 ± 0.242
Intra inter	0.292 ± 0.195	0.940 ± 0.109	0.330 ± 0.225	0.423 ± 0.200

Table 5.22: Polyp tracking metrics using DBSCAN as clustering algorithm across the analyzed models.

HDBSCAN				
Model	ASSA	ASSPR	ASSRE	IDF_I
Pretrained	0.587 ± 0.266	0.897 ± 0.144	0.625 ± 0.252	0.688 ± 0.211
Supervised	0.465 ± 0.205	0.813 ± 0.215	0.634 ± 0.210	0.370 ± 0.231
Temporally aware	0.543 ± 0.237	0.848 ± 0.196	0.629 ± 0.234	0.649 ± 0.207
Intra inter	0.526 ± 0.239	0.792 ± 0.257	0.680 ± 0.251	0.633 ± 0.204

Table 5.23: Polyp tracking metrics using HDBSCAN as clustering algorithm across the analyzed models.

6

Conclusion

The work presented in this thesis revolves around the study of tracklet-based representation learning for polyp-level identity analysis in colonoscopy videos. The central goal was to understand whether current self-supervised foundation models—originally designed for broad endoscopic scene understanding—can also capture the fine-grained, instance-level cues required to recognise the same polyp across time, retrieve all its appearances, count the number of distinct lesions in a procedure, and maintain identity consistency during tracking. To explore this, we structured the project around four existing foundation models (EndoFM, EndoFM-LV, EndoViT and SurgeNet), each representing a different pre-training philosophy. We evaluated how their learned representations transfer to four downstream tasks—re-identification, retrieval, counting, and tracking—using tracklets extracted from REAL-Colon, a realistic dataset containing full colonoscopy procedures with polyp annotations.

The first part of the work focused on evaluating the models with different temporal structures. We explored two complementary settings: one based on variable-length tracklets, which reflects the natural heterogeneity of colonoscopy videos, and another based on fixed-length temporal fragments, which offers a controlled and standardised temporal context. Our results show that performance varies significantly depending on the model and on the evaluation regime. For instance, EndoFM and SurgeNet behave more stably when working with full tracklets and larger spatial context, while showing almost no variation with 8-frame windows. EndoFM-LV, on the other hand, clearly benefits from longer temporal spans, achieving its strongest performance when given full tracklets or 32-frame clips. EndoViT generally follows the same trends but with lower overall performance.

Alongside temporal structure, we also examined the effect of spatial context, ranging from tight polyp crops to full-frame images. The results reveal a consistent and intuitive pattern: aggressive cropping often removes the contextual cues necessary for stable identity discrimination, whereas increasing spatial context leads to more reliable embedding geometry. EndoFM and SurgeNet, in particular, appear to strongly leverage global scene information, whereas EndoFM-LV and EndoViT—both incorporating masked token modelling—tend to struggle with the fine-grained distinctions required for polyp identity, despite being strong general-purpose learners.

A particularly interesting and somewhat unexpected finding emerges from comparing the pure self-distillation-based models (EndoFM and SurgeNet) with those using masked modelling (EndoFM-LV and EndoViT). Despite being designed as an improved version of EndoFM, EndoFM-LV consistently underperforms its predecessor across nearly all identity-related tasks. Likewise, EndoViT—though very strong for segmentation and global tissue understanding—shows limited ability, in some settings, to discriminate similar polyps. These observations suggest that

masked token modelling, while excellent for reconstruction and general visual reasoning, may suppress some of the fine-grained discriminative cues required to distinguish visually similar lesions. This distinction becomes clear in ReID and retrieval, where self-distilled representations remain robust across crops and temporal settings, while masked models often collapse when similarity ranking is required across the entire gallery.

Moreover, we fine-tuned EndoFM using REAL-Colon and three different contrastive objectives: supervised contrastive loss, temporally-aware contrastive loss, and intra–inter loss. Despite their conceptual differences, the supervised and temporally-aware objectives behave identically when only two views per identity are available, as temporal weighting degenerates to a constant factor. Increasing to four views exposes clearer differences: the temporally-aware and intra–inter losses produce more structured and stable embedding spaces. Across tasks, we observe that fine-tuning generally improves identity discrimination, especially when more spatial context is available. The temporally-aware loss tends to excel in settings where longer temporal consistency is required (e.g., tracking and ReID with full frames), while the intra–inter loss performs strongly under aggressive cropping, where local appearance is dominant. Interestingly, the pretrained EndoFM remains very competitive—sometimes even superior—when only tight crops are provided, suggesting that its original self-distilled features are already optimised for local polyp texture.

In counting and tracking experiments, the interaction between the embedding space and the clustering algorithm becomes especially important. Temporal Affinity Propagation proves to be the most stable method across all models, often providing the best balance between fragmentation and false positives. Threshold-based clustering and DBSCAN, instead, show much higher sensitivity to the geometry of the embedding space, favouring the pretrained EndoFM in many cases. Clean pat-

terns also emerge in the tracking metrics: fine-tuned temporally-aware and intra-inter models outperform the pretrained version under affinity-based clustering, but the pretrained model remains the strongest under simple thresholding methods.

Overall, we believe that this work demonstrates both the potential and the current limitations of self-supervised foundation models for entity-level analysis in endoscopy. Through a unified and comprehensive benchmark, we have shown that existing models can indeed support identity-based tasks when provided with appropriate temporal and spatial context, but also that their performance strongly depends on the pre-training objective, the temporal regime, and the structure of the downstream task. More importantly, we highlight that not all self-supervised objectives are equally suited to identity preservation: masked token modelling appears less aligned with the needs of fine-grained polyp-level discrimination than self-distillation, at least in its current form. These insights, together with the tracklet-centric evaluation framework introduced in this thesis, pave the way for future work on designing more targeted representations for lesion-level reasoning.

Although the results obtained throughout this thesis are encouraging, the proposed framework still presents several limitations that deserve consideration. The most relevant constraint concerns the fine-tuning stage: due to GPU memory limitations, training was performed with relatively small batch sizes, which restricted the number of distinct views per identity that could be included in each batch. This limitation directly impacts contrastive objectives, especially those that rely on multiple positive samples to fully shape the embedding geometry. While the models still benefit from fine-tuning, it is reasonable to assume that larger batch sizes—or gradient accumulation strategies—could further stabilise the learned representations and potentially lead to stronger performance across all tasks.

A second limitation relates to the absence of real-time evaluation. While the pro-

posed framework is well-suited for offline analysis, deploying it in a clinical workflow would require assessing its behaviour under streaming conditions, including its ability to handle frame-by-frame updates, abrupt camera motion, occlusions and changing viewpoints. Without such real-time testing, the readiness of these models for integration into decision-support tools remains an open question.

An immediate next step would be to extend the evaluation beyond the REAL-Colon dataset and investigate how well foundation models generalise to gastroscopy or capsule endoscopy datasets. Such experiments would not only test robustness, but also reveal whether the observed behaviours—such as the sensitivity of masked-model pretraining to fine-grained identity cues—persist across modalities.

A

Tables

Temporal Affinity Propagation				
Model	FPR	FR	Precision	Recall
EndoFM	0.021 ± 0.038	2.238 ± 1.651	0.932 ± 0.116	0.552 ± 0.244
EndoFM-LV	0.030 ± 0.060	2.087 ± 1.447	0.924 ± 0.143	0.551 ± 0.212
EndoViT	0.040 ± 0.064	1.815 ± 1.257	0.912 ± 0.139	0.639 ± 0.262
SurgeNet	0.027 ± 0.058	2.087 ± 1.563	0.940 ± 0.116	0.622 ± 0.285

Table A.1: Polyp counting metrics using T-AP as clustering algorithm across the analyzed models.

Threshold				
Model	FPR	FR	Precision	Recall
EndoFM	0.025 ± 0.047	5.401 ± 4.444	0.888 ± 0.182	0.453 ± 0.231
EndoFM-LV	0.013 ± 0.024	13.140 ± 10.880	0.806 ± 0.28	0.124 ± 0.142
EndoViT	0.413 ± 0.43	2.685 ± 3.672	0.662 ± 0.335	0.755 ± 0.213
SurgeNet	0.055 ± 0.223	9.978 ± 10.613	0.943 ± 0.162	0.252 ± 0.224

Table A.2: Polyp counting metrics using Threshold-based method as clustering algorithm across the analyzed models.

Affinity Propagation				
Model	FPR	FR	Precision	Recall
EndoFM	0.060 ± 0.100	2.094 ± 1.668	0.845 ± 0.2	0.565 ± 0.280
EndoFM-LV	0.150 ± 0.181	2.261 ± 1.678	0.714 ± 0.288	0.454 ± 0.195
EndoViT	0.526 ± 0.499	0.626 ± 0.361	0.645 ± 0.347	1 ± 0
SurgeNet	0.134 ± 0.227	1.248 ± 1.191	0.797 ± 0.248	0.8 ± 0.241

Table A.3: Polyp counting metrics using Affinity Propagation as clustering algorithm across the analyzed models.

DBSCAN				
Model	FPR	FR	Precision	Recall
EndoFM	0.052 ± 0.147	5.234 ± 4.373	0.870 ± 0.216	0.458 ± 0.230
EndoFM-LV	0.029 ± 0.053	12.217 ± 10.183	0.792 ± 0.291	0.165 ± 0.162
EndoViT	0.526 ± 0.499	0.626 ± 0.361	0.645 ± 0.347	1 ± 0
SurgeNet	0.049 ± 0.204	11.310 ± 11.564	0.950 ± 0.154	0.202 ± 0.211

Table A.4: Polyp counting metrics using DBSCAN as clustering algorithm across the analyzed models.

HDBSCAN				
Model	FPR	FR	Precision	Recall
EndoFM	0.038 ± 0.058	2.471 ± 1.164	0.860 ± 0.199	0.513 ± 0.257
EndoFM-LV	0.253 ± 0.316	1.882 ± 1.147	0.692 ± 0.332	0.590 ± 0.212
EndoViT	0.208 ± 0.263	1.484 ± 1.488	0.718 ± 0.309	0.719 ± 0.256
SurgeNet	0.069 ± 0.134	1.725 ± 0.866	0.851 ± 0.201	0.666 ± 0.259

Table A.5: Polyp counting metrics using HDBSCAN as clustering algorithm across the analyzed models.

Temporal Affinity Propagation				
Model	ASSA	ASSPR	ASSRE	IDF_t
EndoFM	0.553 ± 0.209	0.942 ± 0.087	0.583 ± 0.233	0.631 ± 0.198
EndoFM-LV	0.556 ± 0.190	0.931 ± 0.116	0.595 ± 0.216	0.647 ± 0.183
EndoViT	0.615 ± 0.209	0.926 ± 0.107	0.674 ± 0.255	0.689 ± 0.193
SurgeNet	0.611 ± 0.241	0.946 ± 0.096	0.655 ± 0.277	0.674 ± 0.228

Table A.6: Polyp tracking metrics using T-AP as clustering algorithm across the analyzed models.

Threshold				
Model	ASSA	ASSPR	ASSRE	IDF_I
EndoFM	0.468 ± 0.215	0.940 ± 0.096	0.496 ± 0.212	0.607 ± 0.188
EndoFM-LV	0.183 ± 0.130	0.952 ± 0.079	0.193 ± 0.132	0.298 ± 0.169
EndoViT	0.589 ± 0.260	0.808 ± 0.259	0.756 ± 0.237	0.691 ± 0.222
SurgeNet	0.271 ± 0.136	0.958 ± 0.141	0.304 ± 0.212	0.401 ± 0.166

Table A.7: Polyp tracking metrics using Threshold-based methods as clustering algorithm across the analyzed models.

Affinity Propagation				
Model	ASSA	ASSPR	ASSRE	IDF_I
EndoFM	0.531 ± 0.254	0.864 ± 0.176	0.595 ± 0.266	0.623 ± 0.229
EndoFM-LV	0.408 ± 0.195	0.745 ± 0.255	0.511 ± 0.193	0.599 ± 0.160
EndoViT	0.652 ± 0.342	0.652 ± 0.342	1 ± 0	0.713 ± 0.229
SurgeNet	0.641 ± 0.320	0.668 ± 0.325	0.955 ± 0.117	0.707 ± 0.267

Table A.8: Polyp tracking metrics using AP as clustering algorithm across the analyzed models.

DBSCAN				
Model	ASSA	ASSPR	ASSRE	IDF_I
EndoFM	0.472 ± 0.238	0.924 ± 0.137	0.511 ± 0.229	0.608 ± 0.206
EndoFM-LV	0.206 ± 0.135	0.936 ± 0.102	0.228 ± 0.148	0.324 ± 0.174
EndoViT	0.553 ± 0.279	0.678 ± 0.319	0.808 ± 0.181	0.664 ± 0.242
SurgeNet	0.230 ± 0.129	0.962 ± 0.134	0.260 ± 0.201	0.355 ± 0.168

Table A.9: Polyp tracking metrics using DBSCAN as clustering algorithm across the analyzed models.

HDBSCAN				
Model	ASSA	ASSPR	ASSRE	IDF_I
EndoFM	0.587 ± 0.266	0.897 ± 0.144	0.897 ± 0.144	0.625 ± 0.252
EndoFM-LV	0.491 ± 0.252	0.740 ± 0.288	0.670 ± 0.216	0.618 ± 0.208
EndoViT	0.587 ± 0.318	0.750 ± 0.275	0.746 ± 0.245	0.687 ± 0.256
SurgeNet	0.635 ± 0.250	0.881 ± 0.172	0.701 ± 0.234	0.723 ± 0.203

Table A.10: Polyp tracking metrics using HDBSCAN as clustering algorithm across the analyzed models.

Table A.11: RelD rankings.

Crop 1, fragm. 8			Crop 1, tracklet		
Model	AUPR	AUROC	Model	AUPR	AUROC
EndoFM	0.287	0.788	EndoFM	0.240	0.725
SurgeNet	0.185	0.765	EndoFM-LV	0.170	0.644
EndoViT	0.165	0.640	SurgeNet	0.138	0.746
EndoFM-LV	0.145	0.593	EndoViT	0.133	0.650

Table A.12: RelD rankings.

Crop 5, fragm. 8			Crop 5, tracklet		
Model	AUPR	AUROC	Model	AUPR	AUROC
SurgeNet	0.321	0.764	SurgeNet	0.339	0.810
EndoFM	0.284	0.719	EndoFM	0.334	0.795
EndoViT	0.241	0.653	EndoViT	0.223	0.701
EndoFM-LV	0.163	0.607	EndoFM-LV	0.193	0.676

Table A.13: RelD rankings based on AUPR.

Full Size, fragm. 8			Full Size, tracklet		
Model	AUPR	AUROC	Model	AUPR	AUROC
EndoFM	0.287	0.709	EndoFM	0.342	0.803
SurgeNet	0.280	0.582	SurgeNet	0.332	0.792
EndoViT	0.218	0.589	EndoViT	0.262	0.686
EndoFM-LV	0.164	0.582	EndoFM-LV	0.213	0.669

Table A.14: Retrieval ranking.

Crop 1, fragm. 8				Crop 1, tracklet			
Model	HR@1	HR@5	mAP	Model	HR@1	HR@5	mAP
EndoFM	0.943	0.977	0.384	EndoFM	0.784	0.925	0.383
SurgeNet	0.931	0.975	0.359	SurgeNet	0.762	0.893	0.349
EndoViT	0.891	0.955	0.319	EndoViT	0.669	0.853	0.290
EndoFM-LV	0.808	0.917	0.269	EndoFM-LV	0.525	0.762	0.241

Table A.15: Retrieval ranking.

Crop 5, fragm. 8				Crop 5, tracklet			
Model	HR@1	HR@5	mAP	Model	HR@1	HR@5	mAP
EndoFM	0.982	0.994	0.437	SurgeNet	0.887	0.959	0.480
SurgeNet	0.978	0.993	0.456	EndoFM	0.876	0.947	0.458
EndoViT	0.951	0.984	0.381	EndoViT	0.787	0.920	0.382
EndoFM-LV	0.858	0.940	0.299	EndoFM-LV	0.632	0.813	0.282

Table A.16: Retrieval ranking.

Full Size, fragm. 8				Full Size, tracklet			
Model	HR@1	HR@5	mAP	Model	HR@1	HR@5	mAP
SurgeNet	0.955	0.996	0.499	SurgeNet	0.955	0.996	0.499
EndoFM	0.955	0.997	0.467	EndoFM	0.891	0.969	0.510
EndoViT	0.955	0.991	0.438	EndoViT	0.878	0.955	0.482
EndoFM-LV	0.869	0.962	0.330	EndoFM-LV	0.700	0.867	0.350

Temporal Affinity Propagation				
Model	FPR	FR	Precision	Recall
EndoViT	0.040 ± 0.064	1.815 ± 1.257	0.912 ± 0.139	0.639 ± 0.262
SurgeNet	0.027 ± 0.058	2.087 ± 1.563	0.940 ± 0.116	0.622 ± 0.285
EndoFM-LV	0.030 ± 0.060	2.087 ± 1.447	0.924 ± 0.143	0.551 ± 0.212
EndoFM	0.021 ± 0.038	2.238 ± 1.651	0.932 ± 0.116	0.552 ± 0.244

Table A.17: Polyp counting rankings using T-AP based on the FR.

Temporal Affinity Propagation				
Model	ASSA	ASSPR	ASSRE	IDF1
SurgeNet	0.611 ± 0.241	0.946 ± 0.096	0.655 ± 0.277	0.674 ± 0.228
EndoFM	0.553 ± 0.209	0.942 ± 0.087	0.583 ± 0.233	0.631 ± 0.198
EndoViT	0.615 ± 0.209	0.926 ± 0.107	0.674 ± 0.255	0.689 ± 0.193
EndoFM-LV	0.556 ± 0.190	0.931 ± 0.116	0.595 ± 0.216	0.647 ± 0.183

Table A.18: Polyp tracking rankings using T-AP based on the IDF1.

Table A.19: RelD rankings based on AUPR.

Full Size, fragm. 8		
Model	AUPR	AUROC
Temporally aware	0.440	0.938
Intra inter	0.401	0.927
Supervised	0.385	0.939
Pretrained	0.287	0.709

Table A.20: Retrieval ranking.

Full Size, tracklet			
Model	HR@1	HR@5	mAP
Temporally aware	0.917	0.982	0.665
Intra inter	0.910	0.985	0.663
Pretrained	0.891	0.969	0.510
Supervised	0.879	0.963	0.622

References

- [1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [2] Z. Zhang, “Improved adam optimizer for deep neural networks,” 2018 *IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, 2018.
- [3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [4] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, and D. Tao, “A survey on self-supervised learning: Algorithms, applications, and future trends,” 2024. [Online]. Available: <https://arxiv.org/abs/2301.05712>
- [5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” 2020. [Online]. Available: <https://arxiv.org/abs/2002.05709>
- [6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” 2020. [Online]. Available: <https://arxiv.org/abs/1911.05722>
- [7] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” 2020. [Online]. Available: <https://arxiv.org/abs/2003.04297>

- [8] 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>
- [9] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” 2021. [Online]. Available: <https://arxiv.org/abs/2111.06377>
- [10] Z. Tong, Y. Song, J. Wang, and L. Wang, “Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training,” 2022. [Online]. Available: <https://arxiv.org/abs/2203.12602>
- [11] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao, “Videomae v2: Scaling video masked autoencoders with dual masking,” 2023. [Online]. Available: <https://arxiv.org/abs/2303.16727>
- [12] K. Hu, Y. Xiao, Y. Zhang, and X. Gao, “Multi-view masked contrastive representation learning for endoscopic video analysis,” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 47 987–48 014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/55cb562b1f5af71f6707f3ff3c7941e6-Paper-Conference.pdf
- [13] Y. Intrator, N. Aizenberg, A. Livne, E. Rivlin, and R. Goldenberg, *Self-supervised Polyp Re-identification in Colonoscopy*. Springer Nature Switzerland, 2023, p. 590–600. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-43904-9_57
- [14] L. Parolari, A. Cherubini, L. Ballan, and C. Biffi, “Towards polyp counting in full-procedure colonoscopy videos,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.10054>

- [15] —, “Temporally-aware supervised contrastive learning for polyp counting in colonoscopy,” 2025. [Online]. Available: <https://arxiv.org/abs/2507.02493>
- [16] C. Biffi, G. Antonelli, S. Bernhofer, C. Hassan, D. Hirata, M. Iwatate, A. Maieron, P. Salvagnini, and A. Cherubini, “Real-colon: A dataset for developing real-world ai applications in colonoscopy,” 2024. [Online]. Available: <https://arxiv.org/abs/2403.02163>
- [17] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, “Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians,” *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0895611115000567>
- [18] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, “Kvasir-seg: A segmented polyp dataset,” 2019. [Online]. Available: <https://arxiv.org/abs/1911.07069>
- [19] Z. Wang, C. Liu, S. Zhang, and Q. Dou, “Foundation model for endoscopy video analysis via large-scale self-supervised pre-train,” 2024. [Online]. Available: <https://arxiv.org/abs/2306.16741>
- [20] P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux, and A. Bartoli, “Computer-aided classification of gastrointestinal lesions in regular colonoscopy,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 9, pp. 2051–2063, 2016.
- [21] M. Misawa, S. ei Kudo, Y. Mori, K. Hotta, K. Ohtsuka, T. Matsuda, S. Saito, T. Kudo, T. Baba, F. Ishida, H. Itoh, M. Oda, and K. Mori, “Development

- of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video),” *Gastrointestinal Endoscopy*, vol. 93, no. 4, pp. 960–967.e3, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0016510720346551>
- [22] G.-P. Ji, G. Xiao, Y.-C. Chou, D.-P. Fan, K. Zhao, G. Chen, and L. Van Gool, “Video polyp segmentation: A deep learning perspective,” *Machine Intelligence Research*, vol. 19, no. 6, p. 531–549, Nov. 2022. [Online]. Available: <http://dx.doi.org/10.1007/s11633-022-1371-y>
- [23] —, “Video polyp segmentation: A deep learning perspective,” *Machine Intelligence Research*, vol. 19, no. 6, p. 531–549, Nov. 2022. [Online]. Available: <http://dx.doi.org/10.1007/s11633-022-1371-y>
- [24] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, D. Johansen, C. Griwodz, H. K. Stensland, E. Garcia-Ceja, P. T. Schmidt, H. L. Hammer, M. A. Riegler, P. Halvorsen, and T. de Lange, “HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy,” *Scientific Data*, vol. 7, no. 1, p. 283, 2020. [Online]. Available: <https://doi.org/10.1038/s41597-020-00622-y>
- [25] S. D. C. Team, “Metadata record for: Kvasir-Capsule, a video capsule endoscopy dataset,” 5 2021. [Online]. Available: https://springernature.figshare.com/articles/dataset/Metadata_record_for_Kvasir-Capsule_a_video_capsule_endoscopy_dataset/14178905
- [26] C. I. Nwoye, D. Alapatt, T. Yu, A. Vardazaryan, F. Xia, Z. Zhao, T. Xia, F. Jia, Y. Yang, H. Wang, D. Yu, G. Zheng, X. Duan, N. Getty, R. Sanchez-Matilla, M. Robu, L. Zhang, H. Chen, J. Wang, L. Wang, B. Zhang, B. Gerats, S. Raviteja, R. Sathish, R. Tao, S. Kondo, W. Pang, H. Ren,

- J. R. Abbing, M. H. Sarhan, S. Bodenstedt, N. Bhasker, B. Oliveira, H. R. Torres, L. Ling, F. Gaida, T. Czempiel, J. L. Vilaça, P. Morais, J. Fonseca, R. M. Egging, I. N. Wijma, C. Qian, G. Bian, Z. Li, V. Balasubramanian, D. Sheet, I. Luengo, Y. Zhu, S. Ding, J.-A. Aschenbrenner, N. E. van der Kar, M. Xu, M. Islam, L. Seenivasan, A. Jenke, D. Stoyanov, D. Mutter, P. Mascagni, B. Seeliger, C. Gonzalez, and N. Padoy, “Cholectriplet2021: A benchmark challenge for surgical action triplet recognition,” *Medical Image Analysis*, vol. 86, p. 102803, May 2023. [Online]. Available: <http://dx.doi.org/10.1016/j.media.2023.102803>
- [27] Y. Tian, G. Pang, F. Liu, Y. Liu, C. Wang, Y. Chen, J. W. Verjans, and G. Carneiro, “Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection,” 2022. [Online]. Available: <https://arxiv.org/abs/2203.12121>
- [28] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” 2021. [Online]. Available: <https://arxiv.org/abs/2102.04306>
- [29] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, “Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians,” *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0895611115000567>
- [30] L. Wu, Z. Hu, Y. Ji, P. Luo, and S. Zhang, “Multi-frame collaboration for effective endoscopic video polyp detection via spatial-temporal feature transformation,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy,

- S. Speidel, Y. Zheng, and C. Essert, Eds. Cham: Springer International Publishing, 2021, pp. 302–312.
- [31] G. Wang, “Replication Data for: Colonoscopy Polyp Detection and Classification: Dataset Creation and Comparative Evaluations,” 2021. [Online]. Available: <https://doi.org/10.7910/DVN/FCBUOR>
- [32] Z. Wang, C. Liu, L. Zhu, T. Wang, S. Zhang, and Q. Dou, “Improving foundation model for endoscopy video analysis via representation learning on long sequences,” *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 5, pp. 3526–3536, 2025.
- [33] A. Sablayrolles, M. Douze, C. Schmid, and H. Jégou, “Spreading vectors for similarity search,” 2019. [Online]. Available: <https://arxiv.org/abs/1806.03198>
- [34] B. D, H. F, Özsoy E, C. T, and N. N., “Endovit: pretraining vision transformers on a large collection of endoscopic images,” *Int J Comput Assist Radiol Surg.*, 2024.
- [35] K. You, M. Long, J. Wang, and M. I. Jordan, “How does learning rate decay help modern neural networks?” 2019. [Online]. Available: <https://arxiv.org/abs/1908.01878>
- [36] P. Izmailov, D. Podoprikin, T. Garipov, D. Vetrov, and A. G. Wilson, “Averaging weights leads to wider optima and better generalization,” 2019. [Online]. Available: <https://arxiv.org/abs/1803.05407>
- [37] T. J. M. Jaspers, R. L. P. D. de Jong, Y. Li, C. H. J. Kusters, F. H. A. Bakker, R. C. van Jaarsveld, G. M. Kuiper, R. van Hillegersberg, J. P. Ruurda, W. M. Brinkman, J. P. W. Pluim, P. H. N. de With, M. Breeuwer, Y. A. Khalil, and F. van der Sommen, “Scaling up self-supervised learning

- for improved surgical foundation models,” 2025. [Online]. Available: <https://arxiv.org/abs/2501.09436>
- [38] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A convnet for the 2020s,” 2022. [Online]. Available: <https://arxiv.org/abs/2201.03545>
- [39] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 548–558.
- [40] W. Yu, C. Si, P. Zhou, M. Luo, Y. Zhou, J. Feng, S. Yan, and X. Wang, “Metaformer baselines for vision,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 2, pp. 896–912, 2024.
- [41] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2019. [Online]. Available: <https://arxiv.org/abs/1711.05101>

Acknowledgments

My heartfelt thanks go to Prof. Lamberto Ballan and Dr. Luca Parolari, whose support and availability, over these months, have played a crucial role in bringing this thesis to completion.

I am also deeply grateful to my family and friends for always being by my side; I truly could not have made it without them.