ID N. 32954



UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA

DEPARTMENT OF ENGINEERING

Italian National Ph.D. in Artificial Intelligence
Health and Life Sciences
XXXVII Cycle

Data-Efficient Deep Learning for Surgical Perception in Minimally Invasive Surgery

Supervisors

Paolo Soda

Loredana Zollo

Clemente Lauretti

Candidate

Guido Manni

Abstract

Minimally invasive surgery has become the standard of care across a wide range of clinical specialities, with endoscopic and robotic procedures offering shorter recovery times, reduced complication rates, and superior ergonomics compared to open surgery. However, both paradigms rely on a monocular two-dimensional video feed as the sole visual channel into the surgical site, resulting in diminished depth perception, blind spots, and an intrinsic loss of spatial context that make tissue examination, diagnosis, and instrument manipulation substantially more demanding. Artificial intelligence, and deep learning in particular, has emerged as a transformative route to restore this missing perceptual information and to enable higher levels of surgical scene understanding. Yet its clinical deployment is hindered by two compounding barriers: the scarcity of labeled medical data and the strict real-time inference requirements of the operating room. The first contribution of this thesis addresses the recovery of three-dimensional geometry from a single endoscopic camera. We propose *BodySLAM*, a generalized monocular visual SLAM framework that integrates three components: *CycleVO*, a novel unsupervised pose estimation module that learns robust camera motion representations from unlabeled endoscopic video through a GAN-based image-to-image translation pretext task; the Zoe architecture for monocular depth estimation, leveraged without fine-tuning to exploit its strong cross-domain generalization; and a three-dimensional reconstruction module coupled with an Unscented Kalman Filter and pose-graph optimization for scale-consistent mapping. Evaluated on the three largest public endoscopic MVSLAM benchmarks across laparoscopy, gastroscopy, and colonoscopy, BodySLAM matches or outperforms four state-of-the-art methods, while CycleVO attains the lowest inference time among all competitors, simultaneously addressing the data-scarcity and real-time barriers. The second contribution targets the downstream perception task of classification under extreme label scarcity, a regime in which deep networks traditionally fail to converge. We introduce *SPARSE*, a three-player GAN-based semi-supervised learning framework that alternates supervised and unsupervised phases, leverages class-conditioned image-to-image translation to enrich feature representations from real unlabeled images rather than generating them from noise, and employs a confidence-weighted temporal ensemble for

reliable pseudo-labeling. Across eleven MedMNIST benchmarks spanning diverse imaging modalities, SPARSE achieves statistically significant improvements over six state-of-the-art semi-supervised methods with as few as five labeled samples per class, demonstrating that robust classification is attainable under data conditions that mirror the realities of clinical medical image annotation. Collectively, the contributions of this thesis address the compounding challenges of data scarcity and real-time inference that currently stand between deep learning and its clinical deployment in endoscopic and robotic surgery. By unifying unsupervised pose estimation, generalizable depth prediction, and image-translation-based semi-supervised classification into coherent, computationally efficient pipelines, the proposed approaches pave the way toward AI-enhanced surgical perception systems that are simultaneously accurate, data-efficient, and viable for the operating room.

Contents

1 Toward AI-Enhanced Surgical Perception **12**
1.1 Scope: Robot-Assisted Minimally Invasive Surgery . . . 12
1.2 Surgical Perception and Scene Understanding . . . 13
1.3 Challenges and Research Questions . . . 14
1.4 Thesis Contributions . . . 16
1.5 Chapter Organization . . . 18

2 BodySLAM: A Generalised MVSLAM Framework for Surgical Applications **19**
2.1 Introduction . . . 19
2.2 Related Works . . . 21
2.2.1 **Traditional Monocular Visual SLAM Architecture** . . . 21
2.2.2 **Challenges and Limitations of Traditional MVSLAM in Endoscopic Surgery** . . . 21
2.2.3 **Recent Advancements in Deep Learning-based Endoscopic MVSLAM** . . . 22
2.2.4 **Challenges in Deep Learning-based MVSLAM for Endoscopic Surgery** . . . 23
2.3 Methods . . . 24
2.3.1 Monocular Pose Estimation Module (MPEM): CycleVO . . . 25
2.3.2 Monocular Depth Estimation Module (MDEM): Zoe . . . 28
2.3.3 3D Reconstruction Module (3DM) . . . 29
2.4 Materials . . . 33
2.4.1 **Training Datasets for CycleVO** . . . 33
2.4.2 **Testing Datasets for Depth and Pose Estimation Modules** . . . 34
2.5 Experimental Configuration . . . 35
2.5.1 Performance Metrics and Statistical Analysis . . . 36

2.6 Results . . . 37
2.6.1 Monocular Pose Estimation Module (MPEM) . . . 38
2.6.2 Monocular Depth Estimation Module (MDEM) . . . 42
2.7 Conclusions . . . 46

3 SPARSE: Few-Shot Semi-Supervised Learning via Class-Conditioned Image Translation 49
3.1 Introduction . . . 49
3.2 Related Works . . . 51
3.3 Methods . . . 54
3.3.1 Supervised Training Phase . . . 56
3.3.2 Unsupervised Training Phase . . . 58
3.3.3 Inference configuration . . . 63
3.4 Experimental Configuration . . . 63
3.4.1 Materials . . . 63
3.4.2 Training Configuration . . . 64
3.5 Results . . . 64
3.5.1 Performance in 5-shot . . . 66
3.5.2 Performance in 50-shot . . . 68
3.5.3 Framework Component Analysis . . . 71
3.5.4 Computational Analysis . . . 74
3.6 Conclusions . . . 76

4 Conclusions and Future Directions 78
4.1 Toward Clinical Translation . . . 81

Appendices 95

A Per-Dataset Results and Dataset-Characteristic Analysis for SPARSE 95
A.1 Dataset Characteristics and Performance Patterns . . . 95
A.1.1 Discriminative Features and Image-to-Image Translation . . . 97
A.1.2 Unlabeled Dataset Size Impact . . . 98

List of Figures

1.1 The two fundamental challenges that must be simultaneously addressed for the clinical deployment of AI in endoscopic and robotic surgery. Data scarcity and real-time inference constraints act as compounding barriers that any practically viable system must overcome together. 15

2.1 Comparison of (A) a classical MVSLAM system and (B) a fully deep learning-based MVSLAM framework. The traditional approach consists of feature extraction, feature matching, motion estimation, depth map estimation, map building, and optimization components. In contrast, the deep learning-based framework replaces conventional algorithms with deep learning models for feature extraction, feature matching, motion estimation, and depth estimation, potentially simplifying the overall architecture complexity. 22

2.2 The approach takes RGB frames as input and outputs a 3D reconstruction of the surgical scene. The MPEM, utilizing CycleVO, computes relative motion between consecutive frames, outputting a motion matrix $M = [R, t_{unscaled}, 0, 1]$, where R is the rotation matrix and $t_{unscaled}$ is the unscaled translation vector. The MDEM estimates depth maps from RGB frames. The Translation Estimation Module estimates a scaled translation vector t_{scaled}, which is then combined with $t_{unscaled}$ using an Unscented Kalman Filter to correct the scale of the motion matrix. Finally, the 3DM combines the RGB frames, depth maps, and scaled pose matrices to generate and update the 3D model 24

2.3 Integration of Pose Estimation within the Cycle Consistency Framework and the Neural Network Architecture. *Left:* The diagram illustrates the integration of pose estimation within the Cycle Consistency framework. Generators GenAB and GenBA perform domain transformations, while pose networks P_{AB} and P_{BA} predict the relative pose between consecutive frames. The predicted pose M is concatenated with the latent space z to improve image-to-image translation performance. *Right:* The neural network architecture for pose estimation includes convolutional, downsampling, residual, and upsampling layers. The bottleneck is modified to accommodate pose estimation, where the generator encoder E processes concatenated frames $f_c = [f_{i-1}, f_i]$ and the pose estimation tail P produces the relative motion matrix $M = [R, t_{unscaled}, 0, 1]$, where R is the rotation matrix and $t_{unscaled}$ is the unscaled translation vector. The latent space z and predicted pose M are concatenated and fed into the generator G to produce the next frame $\hat{f}_i$. . . 26

2.4 Qualitative results of BodySLAM on a sequence of the Hamlyn dataset. Yellow frustums represent the estimated camera pose and orientation at frames 1, 500, 1000, 1500, and the final frame, demonstrating consistent pose tracking throughout the endoscopic sequence. 38

2.5 Comparison of the performance of EndoDepth, EndoSfmLearner, AFSfMlearner, OneSLAM and CycleVO algorithms on the SCARED and ENDOSLAM datasets. Metrics evaluated include Absolute Trajectory Error (ATE), Relative Trajectory Error (RTE), and Relative Rotation Error (RRE). Each box plot shows the distribution of the errors for the respective algorithms, highlighting the median, interquartile range, and outliers. 39

2.6 Ablation study of cycle loss term modifications. This figure presents the impact of different cycle loss term settings ($\lambda_2 \neq 0$ vs $\lambda_2 = 0$) on the performance of models evaluated on the SCARED and ENDOSLAM datasets. The metrics assessed include Absolute Trajectory Error (ATE), Relative Trajectory Error (RTE), and Relative Rotation Error (RRE). Each box plot illustrates the distribution of the errors for the respective models, highlighting the median, interquartile range, and outliers. 41

2.7 Comparison of depth estimation performance for EndoDepth, EndoSfmLearner, AFSfMlearner and BodySLAM across various metrics on Hamlyn Dataset. Box plots show that BodySLAM achieves the best performance with the lowest median values for absolute relative difference, RMSE, RMSE log, and squared relative difference, and the highest median values for accuracy thresholds $\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$. EndoSfmLearner demonstrates the second best performance, while EndoDepth shows the poorest results among the three methods compared. 44

2.8 Comparison of depth estimation performance for EndoDepth, EndoSfmLearner, AFSfMlearner and BodySLAM across various metrics on SCARED Dataset. Box plots show that BodySLAM achieves the best performance with the lowest median values for absolute relative difference, RMSE log, and squared relative difference, and the highest median values for accuracy thresholds $\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$. EndoSfmLearner demonstrates the second best performance, while EndoDepth shows the poorest results among the three methods compared. Outliers beyond 1.5 times the interquartile range from the quartiles have been excluded to enhance clarity and focus on the central trends of the data. 45

3.1 Three-phase framework for semi-supervised learning with limited labeled data. Our approach integrates three specialized networks: a generator ($\mathscr{G}$) for class-conditioned image synthesis, a discriminator ($\mathscr{D}$) for authenticity assessment and classification signaling, and a dedicated classifier ($\mathscr{C}$) for the primary classification task. The generator comprises an encoder ($\mathscr{E}_{\mathscr{G}}$) and decoder ($\mathscr{D}_{\mathscr{G}}$) for image-to-image translation.(a) Initial Training Phase: Joint training of these three networks using limited labeled data. (b) Self-Supervised Pre-training: Two-part approach combining (1) ensemble-based pseudo-labeling using confidence-weighted voting and temporal ensembling, and (2) class-conditioned image translation with Wasserstein GAN training and cycle consistency. (c) Synthetic Data Enhancement Phase: Training classifier on generated samples using target class conditions as supervision signals, where (1) input images paired with target class conditions are processed by the generator to create synthetic training samples, and (2) generated samples are used to train the classifier with target class conditions as supervision signals. Fire symbols indicate trainable networks while ice symbols represent frozen weights during respective training phases. 54

3.2 Average classification accuracy as a function of the number of labeled samples per class (shots) for different unsupervised training frequencies (μ). Results are averaged across all eleven MedMNIST datasets and computed on the validation set. The parameter μ controls how frequently the unsupervised training phase is executed, with $\mu = 0$ representing supervised learning only (no unsupervised phase), $\mu = 1$ indicating unsupervised training at every epoch, and higher values ($\mu \in \{10, 25, 50, 100\}$) representing unsupervised training every μ epochs. 71

List of Tables

2.1 Overview of endoscopic datasets used for evaluating depth estimation and pose estimation modules in MVSLAM framework. Checkmarks indicate the specific modules tested on each dataset 33

2.2 Key metrics for evaluating depth prediction performance. d_i and d_i^* denote the predicted and ground-truth depth at pixel i, respectively, and N is the number of valid pixels. 36

2.3 Comparative overview of trajectory evaluation metrics used to assess the monocular pose estimation module. T_i and T_i^* denote the estimated and ground-truth poses at frame i, respectively, Δ_i is the relative transform between consecutive frames, and n is the number of frames. 37

2.4 Comparative analysis of BodySLAM against EndoDepth, EndoSfMLearner, OneSLAM, and AFSfMlearner across different metrics (ATE, RTE, RRE) for SCARED and ENDOSLAM datasets. T-Statistic and P-Value are reported for each comparison. Bold P-Values indicate statistically significant differences (p ¡ 0.05). 40

2.5 Comparison of models, datasets, and execution times. This table presents the processing times for different models across two datasets, EndoSLAM and SCARED, highlighting the computational efficiency of each approach. All computations were performed on an NVIDIA RTX 3070 GPU. 41

2.6 Statistical Performance Analysis of the Monocular Depth Estimation Module on the Hamlyn Dataset. The table details outcomes for Absolute Relative Difference, Root Mean Square Error (RMSE), RMSE Log, Squared Relative Error, and accuracy thresholds (accuracy ¡ 1.25, accuracy ¡ $1.25^{\mathbf{2}}$, accuracy ¡ $1.25^{\mathbf{3}}$). Results include pairwise comparisons of BodySLAM versus EndoDepth, EndoSfMLearner, and AFSfMLearner, assessed through t-tests with a significance level of p ¡ 0.05. 42

2.7 Statistical Performance Analysis of the Monocular Depth Estimation Module on the SCARED Dataset. The table details outcomes for Absolute Relative Difference, Root Mean Square Error (RMSE), RMSE Log, Squared Relative Error, and accuracy thresholds (accuracy ¡ 1.25, accuracy ¡ 1.25^2, accuracy ¡ 1.25^3). Results include pairwise comparisons of BodySLAM versus EndoDepth, EndoSfMLearner, and AFSfMlearner, assessed through t-tests with a significance level of p ¡ 0.05. 43

3.1 Characteristics of MedMNIST datasets used in our experiments. 64
3.2 Training hyperparameters used in our experiments 65
3.3 Model Performance Across Different Few-Shot Settings 66
3.4 5-Shot statistical comparison of model performance. The upper triangular part shows p-values from statistical comparison (significant values $p < 0.05$ highlighted in bold), while the lower triangular part shows Win-Tie-Loss (W-T-L) statistics. The r-value represents the effect size (Pearson's correlation coefficient) with interpretations: small ($r < 0.3$), medium ($0.3 \leq r < 0.5$), large ($0.5 \leq r < 0.7$), and very large ($r \geq 0.7$). Best performing model is highlighted in bold. 67
3.5 50-Shot statistical comparison of model performance. The upper triangular part shows p-values from statistical comparison (significant values $p < 0.05$ highlighted in bold), while the lower triangular part shows Win-Tie-Loss (W-T-L) statistics. The r-value represents the effect size (Pearson's correlation coefficient) with interpretations: small ($r < 0.3$), medium ($0.3 \leq r < 0.5$), large ($0.5 \leq r < 0.7$), and very large ($r \geq 0.7$). Best performing model is highlighted in bold. 69
3.6 Component ablation analysis across few-shot settings. Average accuracy and Cohen's d effect sizes when removing individual components from the full SPARSE framework. Results averaged over SPARSE and SPARSEens configurations across all eleven datasets. Positive d indicates full model outperforms ablated variant. Effect size interpretation: $|d| < 0.2$ (negligible), $0.2 \leq |d| < 0.5$ (small), $0.5 \leq |d| < 0.8$ (medium), $|d| \geq 0.8$ (large). 73
3.7 Computational complexity analysis of SPARSE framework components. Measurements are reported for input images of size 128×128×3 pixels with 8 output classes. 75

A.1 Comparison of accuracy scores across different methods on medical image datasets using 5-shot learning. Best performance for each dataset is highlighted in bold. 95
A.2 Comparison of accuracy scores across different methods on medical image datasets using 10-shot learning. Best performance for each dataset is highlighted in bold. 96
A.3 Comparison of accuracy scores across different methods on medical image datasets using 20-shot learning. Best performance for each dataset is highlighted in bold. 96
A.4 Comparison of accuracy scores across different methods on medical image datasets using 50-shot learning. Best performance for each dataset is highlighted in bold. 97

Chapter 1

Toward AI-Enhanced Surgical Perception

Minimally invasive surgery has fundamentally transformed the modern clinical practice, offering shorter recovery times, reduced complication rates, and lower overall costs compared to open procedures [58, 15]. Among its many forms, endoscopic surgery has become the preferred approach across a wide range of medical specialities [58, 11]. In a typical endoscopic procedure, a monocular camera, the endoscope, is inserted into the body through natural orifices or small incisions, providing the surgeon with a two-dimensional video feed of the internal anatomy that serves as the sole visual channel for navigating the surgical site and manipulating instruments [15]. Robotic surgery has emerged as a natural evolution of this paradigm, introducing enhanced dexterity, tremor filtering, and improved ergonomics [70, 35]. Nevertheless both paradigms share a fundamental bottleneck: a reliance on two-dimensional monocular feedback that limits depth perception, spatial understanding, and overall scene comprehension [25, 10]. Overcoming this bottleneck through artificial intelligence is the central objective of this thesis.

1.1 Scope: Robot-Assisted Minimally Invasive Surgery

The term *surgical robotics* is broad and encompasses a heterogeneous set of paradigms, including teleoperated master–slave platforms for laparoscopic and endoluminal procedures, cooperatively controlled hand-held tools, image-guided stereotactic robots for neurosurgery and orthopaedics, steerable catheters for endovascular interventions, and emerging micro- and nano-scale robots for targeted therapy [52, 15]. Across this landscape, the common trait that defines the clinical context addressed in this thesis is a specific combination of constraints: the surgical field is accessed through small incisions or natural orifices; the primary visual channel is a monocular endoscopic camera; and the procedure must be carried

out under real-time visual feedback. We refer to this setting as *robot-assisted minimally invasive surgery* (RA-MIS), and it is the scenario in which the methods proposed in this thesis are intended to operate.

Concrete systems that exemplify RA-MIS include teleoperated multi-arm platforms such as the da Vinci Surgical System (Intuitive Surgical) [24, 35], the Hugo™ RAS system (Medtronic) [9], the Versius system (CMR Surgical) [78], and the Senhance Surgical System (Asensus Surgical) [63], as well as flexible robotic endoscopy platforms such as Monarch (Auris / Johnson & Johnson) [55] and the Ion endoluminal system (Intuitive Surgical) [27]. Across all these platforms, the surgeon operates from a console or external controller while relying, as the primary or sole visual cue to the operative field, on a monocular or stereo-from-monocular digital endoscopic video stream. Consequently, any AI-based perception module that restores depth, rebuilds three-dimensional scene geometry, or classifies tissue and lesions from the video feed can, in principle, be deployed on top of the standard video output of such systems without requiring new sensing hardware. This is precisely the deployment surface targeted by the methods proposed in the following chapters: rather than committing to a single manufacturer or mechanical architecture, the contributions of this thesis are designed to operate on the endoscopic video stream that is common to all of the platforms above.

The BodySLAM framework introduced in Chapter 2 directly targets the depth-estimation and 3D-reconstruction stages of such pipelines, while the SPARSE framework of Chapter 3 targets downstream classification tasks (e.g., tissue or lesion recognition) that operate on the same video stream under clinical annotation constraints.

1.2 Surgical Perception and Scene Understanding

To understand why AI is needed, it is worth examining these perceptual limitations in greater detail. The loss of depth perception and the complex topological and photometric properties of tissue produce blind spots and substantial appearance variations that make the examination and diagnosis of internal anatomical structures difficult [10, 15]. Consequently, important lesions can be missed or misdiagnosed, and entire tissue areas can be overlooked. Furthermore, endoscope and instrument manipulation within deformable and narrow anatomical cavities demands a high level of expertise and dexterity [73, 15]. These difficulties are further aggravated by the intrinsic nature of the surgical scene. Soft tissues deform continuously under instrument interaction, respiration, and pulsation [69], while instruments, anatomy, and the boundaries of the field of view frequently occlude one another. Illumination is equally problematic, since the light source moves together with the endoscope and produces abrupt brightness variations, specular highlights, and strongly non-uniform

lighting [45, 92, 71]. The operative field is additionally degraded by blood, irrigation fluids, and surgical smoke, which introduce transient artifacts and further occlude the underlying anatomy [38, 39]. Artificial intelligence, and deep learning in particular, has emerged as a transformative solution to these limitations [15, 96]. On one hand, AI can directly restore the perceptual information lost in monocular endoscopy: deep learning-based methods for monocular depth estimation and ***simultaneous localization and mapping*** (SLAM) can reconstruct the three-dimensional geometry of the surgical scene from a single camera view [66, 42]. On the other hand, AI can enhance surgical procedures beyond what unaided human perception can achieve. Computer-assisted detection systems have been shown to substantially reduce the miss rate of adenomas during colonoscopy, with convolutional neural networks reaching accuracy exceeding 95% for colorectal polyp detection [83, 64, 80], and modelling analyses projecting significant reductions in colorectal cancer incidence when deployed at scale [2]. Beyond detection, surgical instrument segmentation, workflow recognition, gesture recognition, and skill assessment are essential building blocks for comprehensive computer-assisted intervention platforms [15, 96, 44], and AI-driven perception is laying the groundwork for increasing levels of surgical autonomy [15].

1.3 Challenges and Research Questions

Despite the remarkable promise of AI in this field, its effective deployment in endoscopic and robotic surgery is hindered by two fundamental and interrelated challenges: **data scarcity** and **real-time computational requirements**. These two challenges are summarized in Figure 1.1 and are discussed in detail below, each paired with the research question it motivates.

Data Scarcity

In the broader machine learning literature, data scarcity has been addressed through unsupervised learning, semi-supervised learning, transfer learning, data augmentation, and synthetic data generation [81, 49, 60]. In the medical and endoscopic domain, however, this problem takes on a particularly acute form. Patient privacy regulations severely restrict the collection and sharing of medical imaging data, and only clinicians with specialized training can accurately annotate surgical images and videos, making large-scale annotation efforts prohibitively expensive [47]. In the specific context of endoscopic surgery, the situation presents a subtle but important nuance: the critical bottleneck is not the absence of raw data, since endoscopic procedures inherently generate hours of continuous video rich in visual

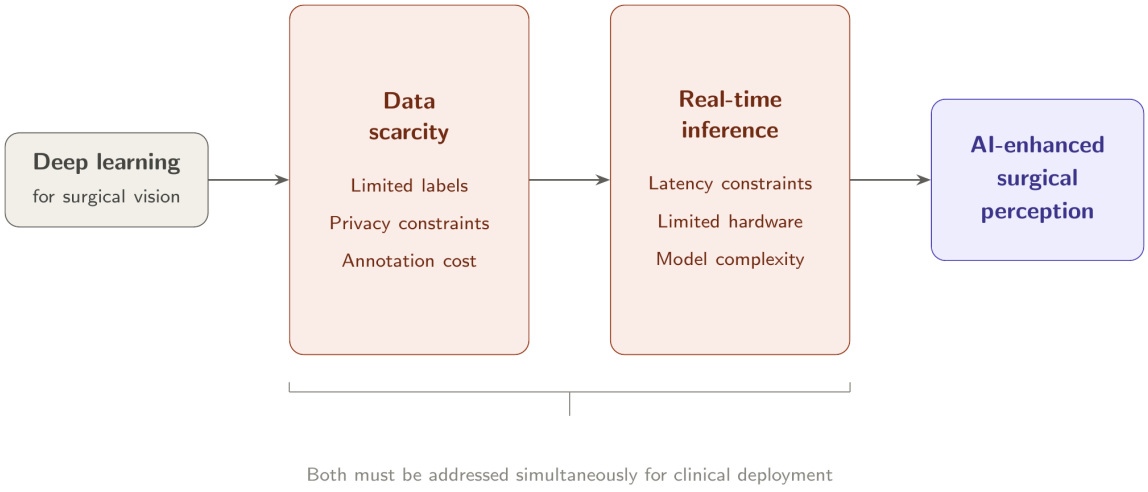

Figure 1.1: The two fundamental challenges that must be simultaneously addressed for the clinical deployment of AI in endoscopic and robotic surgery. Data scarcity and real-time inference constraints act as compounding barriers that any practically viable system must overcome together.

information. The true bottleneck is rather the scarcity of *labeled* data, as providing frame-level annotations for lengthy surgical videos is extraordinarily labor-intensive and expensive. This scarcity of labels is compounded by the heterogeneity of surgical data. Endoscopic video is acquired with different optical systems, resolutions, fields of view, and color responses across manufacturers and clinical centers, so that models trained at a single site or on a single device often exhibit degraded performance when transferred elsewhere [44, 67]. As a consequence, inter-center and inter-device variability turns generalization, rather than in-domain accuracy alone, into a central requirement for surgical perception, and it makes learning from a single small labeled dataset even less sufficient. The problem is exacerbated by the fact that, in contrast with mainstream computer vision, the field still lacks large, standardized, and openly available benchmarks spanning multiple procedures, centers, and imaging systems, a scarcity that limits reproducibility and complicates the fair comparison of competing methods [67, 44]. This observation motivates the development of methods that can learn effectively from large pools of unlabeled data with only minimal supervision. Whether such methods can achieve accurate and generalizable representations for surgical perception and scene understanding remains an open question that drives the first research direction of this thesis:

Research Question 1

Can deep learning methods learn accurate and generalizable representations for surgical perception and scene understanding while operating under severe labeled data constraints?

Real-Time Computational Requirements

For an AI system to be clinically useful during a surgical procedure, it must produce its predictions within the tight latency constraints imposed by the dynamics of surgery [15]. State-of-the-art deep learning models are often computationally demanding, and their inference time does not always satisfy real-time requirements. This is compounded by the practical reality of surgical environments, where computational resources may be limited and where integrating multiple AI functionalities within a single platform imposes a trade-off between model tractability, robustness, and accuracy [15]. Moreover, the heterogeneity and visual complexity of the surgical environment discussed above, spanning inter-center variability, tissue deformation, and severe appearance changes, push toward larger and more expressive models in order to attain robust and generalizable performance, which in turn increases their computational cost and inference latency. Crucially, this challenge does not exist in isolation: the more complex models generally needed to learn robust representations from limited data are precisely the ones most likely to violate real-time constraints, making the two challenges mutually reinforcing. Whether it is possible to design AI systems that are simultaneously data-efficient and computationally viable for clinical deployment constitutes the second open question addressed in this thesis:

Research Question 2

Can AI systems for surgical perception and scene understanding satisfy the strict real-time inference requirements of clinical deployment without sacrificing accuracy or generalizability?

1.4 Thesis Contributions

To pursue the objective of advancing surgical perception and scene understanding while simultaneously tackling data scarcity and real-time inference, this thesis presents two research contributions, each addressing a complementary facet of the overarching challenge. The **first contribution**, BodySLAM [46], directly addresses the perceptual limitations of endoscopic surgery by proposing a generalized deep learning-based Monocular Visual SLAM framework

for surgical applications. BodySLAM integrates three key components: *CycleVO*, a novel unsupervised monocular pose estimation module that learns robust camera motion representations from unlabeled endoscopic video through a GAN-based image-to-image translation pretext task; the Zoe architecture for monocular depth estimation, which provides superior generalization across diverse endoscopic scenarios without fine-tuning; and a three-dimensional reconstruction module that fuses pose and depth information into a coherent surgical map. Crucially, CycleVO is entirely unsupervised, requiring no labeled pose data for training, and achieves the lowest inference time among all compared pose estimation methods, addressing both RQ1 and RQ2 simultaneously.

The **second contribution**, SPARSE [47], tackles data scarcity in medical image classification through a novel GAN-based semi-supervised learning framework specifically designed for extremely low labeled-data regimes. SPARSE introduces a three-phase training framework that alternates between supervised learning on minimal labeled data and unsupervised learning that leverages abundant unlabeled images through class-conditioned image-to-image translation, combined with an ensemble-based pseudo-labeling mechanism for reliable label estimation. Evaluated across eleven benchmark datasets with as few as five labeled samples per class, SPARSE demonstrates statistically significant improvements over six state-of-the-art semi-supervised methods, directly addressing RQ1.

Together, these contributions form a coherent research trajectory that progresses from restoring the lost perceptual information in endoscopic surgery to enabling robust learning under extreme data constraints, all while maintaining the computational efficiency required for clinical deployment.

It is worth situating these two contributions within the broader trajectory of contemporary machine learning, which is increasingly shaped by large-scale self-supervised pretraining and by foundation models, including vision transformers and multimodal architectures, that learn transferable representations from large unlabeled corpora and adapt to downstream tasks with limited task-specific supervision. The strategies pursued in this thesis are complementary to this trend rather than alternative to it. On one hand, BodySLAM already builds on a strongly pretrained and well-generalizing depth model, namely the Zoe architecture mentioned above, and is benchmarked against a foundation-model approach for endoscopy, the tracking-any-point method OneSLAM [77]. On the other hand, both CycleVO and SPARSE can be regarded as data-efficient mechanisms for adapting or augmenting such large models under the specific constraints of the operating room, where annotations are scarce and inference budgets are tight. This perspective reinforces the unifying principle of the thesis, that surgical perception should be decoupled from the availability of dense annotations.

1.5 Chapter Organization

The structure of this thesis reflects this methodological progression and is composed of a collection of research articles, each addressing a specific dimension of the challenges outlined above. Chapter 2 presents BodySLAM, detailing the design, implementation, and evaluation of the generalized MVSLAM framework for endoscopic surgery, with a focus on the unsupervised CycleVO module and its implications for data-efficient and real-time surgical perception (RQ1, RQ2). Chapter 3 introduces SPARSE, describing the semi-supervised learning framework for medical image classification under extreme label scarcity and demonstrating its effectiveness across a comprehensive suite of medical imaging benchmarks (RQ1). Chapter 4 synthesizes the key insights gathered across both contributions, discusses their implications for the development of clinically useful AI systems, outlines open challenges and future directions, and emphasizes the continued need for AI solutions that are simultaneously accurate, data-efficient, and computationally viable for real-world surgical deployment.

Chapter 2

BodySLAM: A Generalised Monocular Visual SLAM Framework for Surgical Applications

The contents of this chapter are partly or entirely adapted from our work [46], to which the reader is referred for further details. The chapter has been reformulated to fit the overall narrative of this thesis, with minor editorial changes and additional context connecting the material to the surgical robotics scenario outlined in Chapter 1 and to the subsequent contributions.

2.1 Introduction

Laparoscopy has become the preferred surgical method for various procedures across medical specialties due to its minimally invasive nature [67]. However, a significant limitation of this technique is its reliance on two-dimensional views [26], and inferring depth from 2D images can be challenging for surgeons and requires years of experience [74].

To address this challenge, there has been growing interest in integrating 3D imaging technology into surgical procedures. Studies have shown that 3D imaging reduces cognitive strain on surgeons, enhances procedural performance, and increases the accuracy of instrument manipulation [72, 82]. These benefits lead to better procedural outcomes and shortened learning curves for surgical trainees [76, 23, 68].

In recent years, camera-based tracking and mapping methods have gained popularity in surgical 3D imaging research [67]. Approaches such as mosaicking and Shape from Template have been explored, but each faces specific limitations in the surgical context. Mosaick-

ing, which stitches together overlapping images to create a comprehensive view of the scene [7, 4, 37], struggles with illumination variations, camera motion, moving objects, and noise, which can lead to artifacts and reduced visual quality in surgical environments. Shape from Template estimates the 3D shape of a deformable object by aligning a known template to observed images [16, 91], but, it is affected by template size and complexity, image quality and noise, rotation and scaling issues, partial occlusion, and the presence of similar shapes or clutter in surgical scenes.

In contrast, Simultaneous Localization and Mapping (SLAM) has emerged as a promising solution to address these limitations. SLAM algorithms simultaneously estimate the camera pose and construct a map of the environment in real-time, effectively handling the dynamic nature of surgical scenes. However, implementing SLAM in endoscopic procedures presents significant challenges due to hardware limitations and environmental factors. The use of monocular cameras, necessitated by size constraints, and the absence of odometry sensors force SLAM to rely solely on visual information for localization and mapping. This approach is known as Monocular Visual SLAM (MVSLAM).

The surgical environment compounds these difficulties through low-texture surfaces, lighting variations [45, 92], organ deformation [69], and the presence of blood, fluids, and smoke [38, 39]. These factors make robust feature extraction and matching particularly difficult for traditional feature-based MVSLAM methods, such as ORB-SLAM [13] and Structure from Motion [40].

Recent works [54, 51, 71, 77] have explored deep learning-based approaches to address these challenges. However, these methods still face difficulties in generalizing to unseen surgical environments due to data scarcity and the wide variability of surgical scenes. To overcome these limitations, in this chapter we present a robust deep learning MVSLAM approach that effectively works in diverse surgical settings, referred to as *BodySLAM* in the following. It introduces a newly developed model and state-of-the-art that, together, overcome the issues of low-texture surfaces, lighting variations, and generalization to unseen datasets. Our key contributions are:

- CycleVO: a novel GAN-based pose estimation module that enhances accuracy and robustness in surgical settings while addressing data scarcity. This unsupervised method applies concepts from CycleGAN and InfoGAN to learn robust pose representations from unlabeled endoscopic video sequences.

- Integration of the Zoe model [8] for monocular depth estimation: we leverage its superior generalization capabilities across diverse datasets without fine-tuning, marking the first use of such a well-generalized model for depth prediction in endoscopic MVSLAM.

We present an in-depth validation procedure demonstrating its robust performance across various endoscopic scenarios.

To rigorously test our approach and demonstrate its generalization capabilities, we conducted comprehensive evaluations using the largest publicly available endoscopic MVSLAM dataset to date. Our experiments span three surgical environments: laparoscopy, gastroscopy, and colonoscopy. We utilized the Hamlyn [54], EndoSLAM [51], and SCARED [1] datasets, each offering unique features for validating our proposed method. We also benchmarked our approach against four state-of-the-art methods [54, 51, 71, 77], which are overviewed in section 2.2.3 and then presented in section 2.5. The remainder of this chapter is organized as follows: section 3.2 provides an overview of the related work. Section 2.3 describes our proposed approach, detailing the three primary modules of our framework. Section 2.4 presents the materials used in this study, including the training, validation, and testing datasets. Section 2.5 outlines the experimental configuration, including the comparative analysis against the four state-of-the-art approaches, performance metrics, and statistical analysis. Section 2.6 presents the experimental results. Finally, section 2.7 provides concluding remarks.

2.2 Related Works

2.2.1 Traditional Monocular Visual SLAM Architecture

A classical Monocular Visual SLAM (MVSLAM) system typically consists of several integrated components, as illustrated in Figure 2.1A. The framework begins with a feature extraction module, which identifies salient features (like ORB, SURF, and SIFT [57, 41, 6]) within the images. A feature-matching algorithm establishes correspondences between features across consecutive frames, which are then processed by motion estimation and depth map estimation algorithms to determine relative motion and generate sparse depth maps. A map-building algorithm uses these outputs to reconstruct the environmental model, while an optimization module refines estimates through bundle adjustment [79] and pose graph optimization to refine the map and pose estimates

2.2.2 Challenges and Limitations of Traditional MVSLAM in Endoscopic Surgery

Traditional feature-based MVSLAM methodologies face notable challenges in endoscopic surgery due to the distinctive nature of surgical images. Endoscopic images often lack dis-

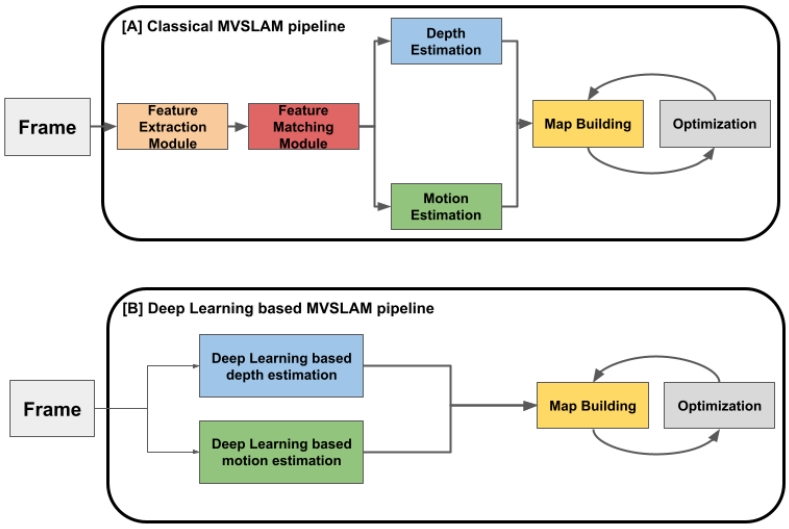

Figure 2.1: Comparison of (A) a classical MVSLAM system and (B) a fully deep learning-based MVSLAM framework. The traditional approach consists of feature extraction, feature matching, motion estimation, depth map estimation, map building, and optimization components. In contrast, the deep learning-based framework replaces conventional algorithms with deep learning models for feature extraction, feature matching, motion estimation, and depth estimation, potentially simplifying the overall architecture complexity.

cernible features and are subject to lighting variations, posing substantial difficulties for feature extraction and matching algorithms. For example, a modified ORB-SLAM [43], using limited number of keyframes for dense reconstruction, struggled with soft, textureless tissue common in minimally invasive surgery. Early SLAM techniques, such as those relying on Structure from Motion [93], aimed to concurrently track camera motion and reconstruct the 3D environment. However, these methods were severely limited by the low-texture environments typical of minimal invasive surgery. Structure from Motion relies on robust wide-baseline feature matching, which becomes untenable in texture-deficient surgical scenes. Moreover, Structure from Motion's computational expense is another drawback. It requires processing all images collectively to optimize 3D structures and camera poses, making real-time surgical applications challenging.

2.2.3 Recent Advancements in Deep Learning-based Endoscopic MVSLAM

To address the limitations of traditional MVSLAM methods in endoscopic surgery, researchers have turned to deep learning-based approaches. As shown in Figure 2.1B, a fully deep learning-based MVSLAM framework replaces the conventional feature matching, fea-

ture extraction, motion estimation, and depth estimation algorithms with deep learning models. These models can learn to extract relevant features and estimate depth and motion directly from the input images, even in the presence of low-texture environments and lighting variations.

Deep learning models offer several advantages over traditional MVSLAM methods for endoscopic surgery. They adapt to surgical environments by learning to extract features and estimate depth and motion directly from data, eliminating handcrafted algorithms that struggle with low-texture scenes.

Significant advancements have been made in deep learning-based endoscopic MVSLAM. EndoSLAM [51] represents a fully end-to-end approach, leveraging convolutional neural networks and attention mechanisms for monocular depth and pose estimation in gastroscopy. This work demonstrated the feasibility of deep learning solutions for domain-specific endoscopic SLAM. Hybrid approaches have also emerged, combining deep learning with traditional techniques. Recasens et al. [54] proposed the Endo-Depth-and-Motion framework, which integrates MonoDepthV2 [29] for depth estimation with a PTAM-inspired [34] photometric method for pose estimation. While this approach showed improvements over purely traditional methods, it still encountered limitations in extended procedures. Addressing a critical challenge in endoscopic environments, Shao et al. [71] introduced the concept of appearance flow to handle brightness inconsistency. Their self-supervised framework employs a generalized dynamic image constraint, relaxing the conventional brightness constancy assumption and enhancing robustness to varying illumination conditions. Most recently, Teufel et al. [77] presented OneSLAM, a generalized monocular SLAM approach for endoscopy utilizing tracking any point (TAP) foundation models. This method achieves sparse correspondences across multiple frames for joint optimization of camera poses and anatomical geometry. Notably, OneSLAM demonstrates zero-shot generalization across various endoscopic domains without retraining, while maintaining or surpassing the performance of domain-specific approaches. These advancements demonstrate deep learning's growing capabilities in endoscopic SLAM, from domain-specific to generalized approaches with broader surgical applicability.

2.2.4 Challenges in Deep Learning-based MVSLAM for Endoscopic Surgery

Despite recent advancements, deep learning-based MVSLAM for endoscopic surgery faces several challenges. The complex surgical environment presents several obstacles [69, 38, 39]. A major hurdle is the requirement for large, diverse datasets to train generalizable models,

which are costly and time-consuming to collect and annotate. Additionally, the computational demands of deep learning models can impede real-time performance in resource-limited surgical settings. Striking a balance between model accuracy and efficiency remains crucial for practical implementation.

2.3 Methods

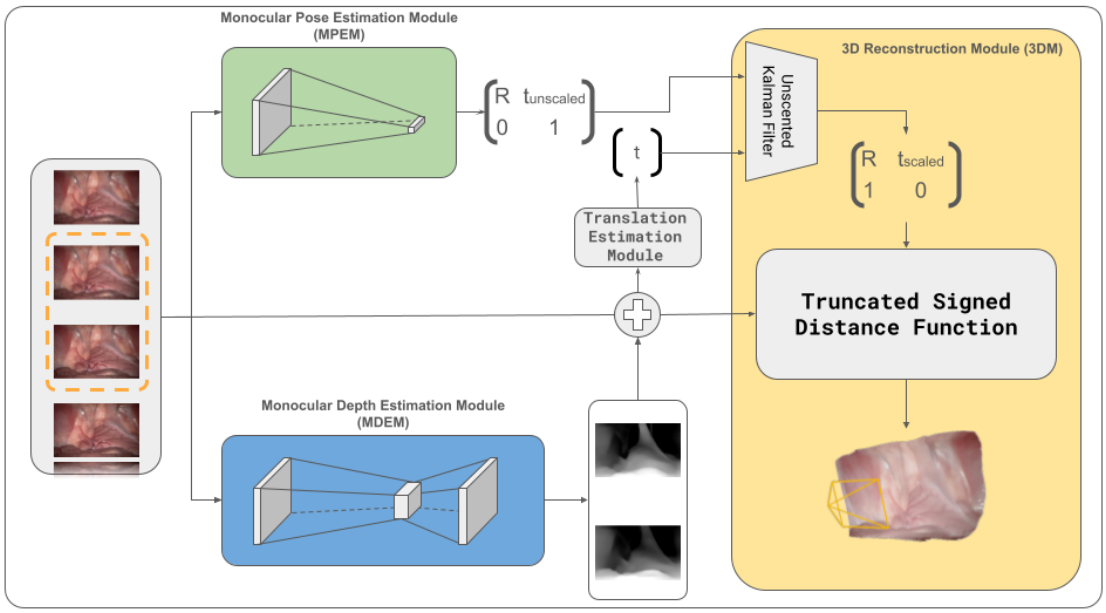

Figure 2.2: The approach takes RGB frames as input and outputs a 3D reconstruction of the surgical scene. The MPEM, utilizing CycleVO, computes relative motion between consecutive frames, outputting a motion matrix $M = [R, t_{unscaled}, 0, 1]$, where R is the rotation matrix and $t_{unscaled}$ is the unscaled translation vector. The MDEM estimates depth maps from RGB frames. The Translation Estimation Module estimates a scaled translation vector t_{scaled}, which is then combined with $t_{unscaled}$ using an Unscented Kalman Filter to correct the scale of the motion matrix. Finally, the 3DM combines the RGB frames, depth maps, and scaled pose matrices to generate and update the 3D model

The proposed MVSLAM approach is composed of three key modules (Figure 2.2): the Monocular Pose Estimation Module (MPEM), the Monocular Depth Estimation Module (MDEM), and the 3D Reconstruction Module (3DM). The framework ensures that while MPEM and MDEM operate independently, their outputs work together to enable effective 3DM functioning. The MPEM (section 2.3.1) uses our novel CycleVO network to estimate relative camera motion between consecutive frames. It outputs a motion matrix $M = [R, t_{unscaled}, 0, 1]$, where R is the rotation matrix and $t_{unscaled}$ is the unscaled translation vector due to monocular scale ambiguity. Simultaneously, the MDEM (section 2.3.2)

estimates the depth map of the scene from the RGB input. Given that the translation component ($t_{unscaled}$) of the relative motion matrix is unscaled, a dedicated module, comprising a translation estimation module and an Unscented Kalman Filter, is tasked with scale correction. The translation estimation module estimates a scaled translation vector t_{scaled}, which is then combined with $t_{unscaled}$ using the UKF to correct the scale of the motion matrix. The 3DM (section 2.3.3) then employs the outputs from both the MDEM and MPEM. It utilizes the depth map from the MDEM and the pose estimations from the MPEM to generate an initial point cloud of the current field of view of the endoscopic camera. This point cloud is carefully merged with the broader scene point cloud, leveraging the pose information to ensure precise alignment and integration.

2.3.1 Monocular Pose Estimation Module (MPEM): CycleVO

The MPEM introduces *CycleVO*, a novel unsupervised method designed to address the challenge of data scarcity in endoscopic environments. While CycleVO incorporates architectural ideas inspired by CycleGAN [97] for cyclic image translation and concepts from InfoGAN [18] for latent space manipulation, it uniquely applies and extends these concepts to pose estimation. The key innovation lies in learning the estimation of relative camera pose between consecutive frames through the pre-task of image-to-image translation. We define this pre-task as a function $G : f_{i-1} \rightarrow f_i$, which generates the current frame f_i from the previous frame f_{i-1}. Similarly, we define the function $F : f_i \rightarrow f_{i-1}$, which generates the previous frame f_{i-1} from the current frame f_i. To improve the performance of this image-to-image translation pre-task, we draw inspiration from InfoGAN's manipulation of the latent space through the maximization of mutual information. We propose to manipulate the latent space z produced by the generator encoder E by concatenating a prediction of the relative pose M between the two consecutive frames to z. The underlying concept is that if the predicted pose M is correct, the image-to-image translation task $G(f_{i-1}, M) \approx f_i$ will also be successful. Conversely, if the estimated pose is incorrect, the pre-task will fail. This novel approach creates a synergy between pose estimation and image generation, enabling the model to learn robust pose representations from unlabeled endoscopic video sequences. The CycleVO generator performs two key tasks (right panel of Figure 2.3):

- Predicting the pose M between two concatenated consecutive frames (f_{i-1} and f_i). This is done by passing the concatenated frame $f_c = [f_{i-1}, f_i]$ through the generator encoder E, which then feeds into the pose estimation tail P. The pose network P outputs the relative motion matrix M in quaternion form to avoid gimbal lock and improve numerical stability. The quaternion is immediately converted to a matrix

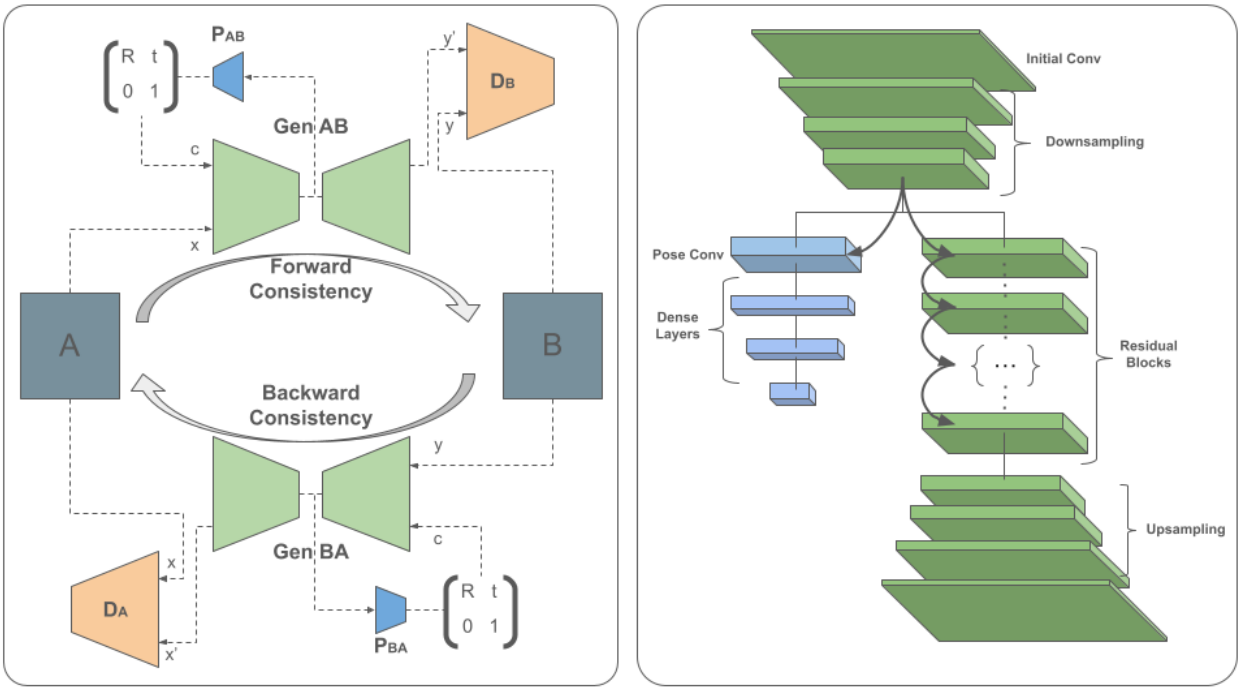

Figure 2.3: Integration of Pose Estimation within the Cycle Consistency Framework and the Neural Network Architecture. *Left:* The diagram illustrates the integration of pose estimation within the Cycle Consistency framework. Generators GenAB and GenBA perform domain transformations, while pose networks P_{AB} and P_{BA} predict the relative pose between consecutive frames. The predicted pose M is concatenated with the latent space z to improve image-to-image translation performance. *Right:* The neural network architecture for pose estimation includes convolutional, downsampling, residual, and upsampling layers. The bottleneck is modified to accommodate pose estimation, where the generator encoder E processes concatenated frames $f_c = [f_{i-1}, f_i]$ and the pose estimation tail P produces the relative motion matrix $M = [R, t_{unscaled}, 0, 1]$, where R is the rotation matrix and $t_{unscaled}$ is the unscaled translation vector. The latent space z and predicted pose M are concatenated and fed into the generator G to produce the next frame $\hat{f}_i$.

representation $M = [R, t_{unscaled}, 0, 1]$, where R is the rotation matrix and $t_{unscaled}$ is the unscaled translation vector, for better integration within the SLAM pipeline. The final linear layer of the pose network is responsible for outputting M.

- Generating the next frame (image-to-image translation pre-task). First, the previous frame f_{i-1} is passed through the encoder E. The resulting latent space $z = E(f_{i-1})$ is then concatenated with the predicted pose M from the first task. This concatenated latent space $[z, M]$ is then fed into the generator G, which produces the next frame $\hat{F}i = G(z, M)$.

We now discuss the objective loss function, which has been optimized for the task of pose estimation. The objective function consists of three terms:

- The adversarial loss for matching the distribution of generated images to the data distribution in the target domain.
- The cycle consistency loss to prevent the learned mappings on frames G and F from contradicting each other.
- The cycle consistency loss on pose to prevent the learned mappings from contradicting each other regarding the estimated pose.

The full objective loss is defined as follows:

$$\begin{aligned} \mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{\text{adv}}(G, D_Y, X, Y) \\ +\mathcal{L}_{\text{adv}}(F, D_X, Y, X) + \lambda_1 \mathcal{L}_{\text{cycImg}}(G, F) \\ +\lambda_2 \mathcal{L}_{\text{cycPose}}(G, F) \end{aligned} \tag{2.1}$$

where the first two terms, $\mathcal{L}_{\text{adv}}(G, D_Y, X, Y)$ and
$\mathcal{L}_{\text{adv}}(F, D_X, Y, X)$, are the adversarial losses. The third term $\lambda_1 \mathcal{L}_{\text{cycImg}}(G, F)$ is the classical cycle consistency loss on the generated images. The final term $\lambda_2 \mathcal{L}_{\text{cycPose}}(G, F)$ is a custom term added to enforce pose consistency between the estimated poses from the generated images. To enforce pose consistency, the term $\mathcal{L}_{\text{cycPose}}(G, F)$ is defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{cycPose}}(G, F) = \mathcal{L}_{\text{chordal}}(R_{i-1,i}, \hat{R}_{i-1,i}) \\ +\mathcal{L}_1(t_{i-1,i}, \hat{t}_{i-1,i}) + \mathcal{L}_{\text{chordal}}(R_{i,i-1}, \hat{R}_{i,i-1}) \\ +\mathcal{L}_1(t_{i,i-1}, \hat{t}_{i,i-1}) \end{aligned} \tag{2.2}$$

where $R_{i-1,i}$ and $\hat{R}_{i-1,i}$ are the rotation components of the estimated pose between the real frames ($f_{(i-1)real}$, $f_{(i)real}$) and the generated frames ($f_{(i-1)recov}$, $F_{(i)real}$), respectively.

Similarly, $t_{i-1,i}$ and $\hat{t}_{i-1,i}$ are the translation vectors of the estimated pose between the real frames ($f_{(i-1)real}$, $f_{(i)real}$) and the generated frames ($f_{(i-1)recov}$, $f_{(i)real}$), respectively. The terms $R_{i,i-1}$, $\hat{R}_{i,i-1}$, $t_{i,i-1}$, and $\hat{t}_{i,i-1}$ are defined analogously for the mapping f. The chordal loss $\mathcal{L}_{\text{chordal}}$ is defined as:

$$\mathcal{L}_{\text{chordal}}(R, \hat{R}) = |R - \hat{R}|_F \tag{2.3}$$

where $|\cdot|_F$ denotes the Frobenius norm. The $\mathcal{L}1$ loss is defined as:

$$\mathcal{L}_1(t, \hat{t}) = \sum i = 1^n |t_i - \hat{t}i| \tag{2.4}$$

where n is the dimension of the translation vector. The chordal loss constrains the rotation component of the estimated relative pose, while the $\mathcal{L}1$ loss constrains the translation component.
The overall workflow (left part of Figure 2.3) is as follows:

- The concatenated frame $f_c = [Fi - 1, f_i]$ is passed through the network to predict the relative pose $M = P(E(F_c))$.
- The previous frame $fi - 1$ is then passed through the encoder E for the image-to-image translation task, producing the latent space $z = E(Fi - 1)$.
- At the encoder level, the predicted pose M is concatenated with the latent space z, resulting in $[z, M]$.
- The concatenated latent space $[z, M]$ is then fed into the generator G, which produces the next frame $\hat{F}_i = G(z, M)$.

2.3.2 Monocular Depth Estimation Module (MDEM): Zoe

The Monocular Depth Estimation Module estimates depth from endoscopic camera frames, which is essential for subsequent 3D reconstruction.
To address data scarcity in endoscopic depth estimation, we employ the Zoe architecture [8]. This marks the first use of such a well-generalized model in endoscopic SLAM.
Zoe model utilizes a distinctive two-stage framework:

1. Pre-training on a wide array of datasets for relative depth estimation, which fosters excellent generalization ability. The datasets used for pre-training include.
2. Adding heads for metrics depth estimation and fine-tuning on metrics depth datasets.

This approach allows Zoe to maintain metric scale while benefiting from pre-training generalization capabilities.
Zoe architecture features a novel metrics bins module, which takes multi-scale features from the MiDaS [53] decoder and predicts bin centres crucial for metrics depth prediction. The module predicts all bin centres at the initial stage and adjusts them at subsequent decoder layers using attractor layers, enabling more accurate and adaptable depth estimation.
Another innovative aspect of Zoe is its use of the log-binomial method for final metrics depth prediction, instead of the conventional softmax approach. This method linearly combines the bin centres, weighted by their probability scores, enhancing the model's ability to accurately predict depth in a structured and ordered manner.
Integrating Zoe into our MVSLAM approach represents a significant advancement in-depth estimation for endoscopic applications. Its architecture, combining the strengths of relative and metrics depth estimation and utilizing novel components, makes it exceptionally suited for the challenging and variable conditions of endoscopic surgery.
In our approach, we chose to assess Zoe's out-of-the-box performance to evaluate its generalization capabilities across domains, without retraining or fine-tuning the model specifically for the surgical context (section 2.6.2).

2.3.3 3D Reconstruction Module (3DM)

The 3D reconstruction module generates 3D models from endoscopic images through a multi-phase process. The key steps are:

1. **Pose Estimate Refinement**: refining the pose estimate provided by the MPEM through multiple optimization steps. This refinement is needed because initial pose estimates contain accumulated drift and noise that would propagate through the reconstruction process and degrade the quality of the final 3D model.

2. **Point Cloud Construction and Integration**: constructing point clouds from the endoscopic camera perspective using pseudo-RGBD frames produced by the MDEM, then aligning and integrating them with the pre-existing reconstructed scene. This process is necessary to convert 2D endoscopic images into a coherent global 3D representation while maintaining spatial consistency across multiple viewpoints.

Pose Estimate Refinement

The pose refinement process addresses both scale ambiguity inherent to monocular pose estimation and accumulated drift errors. To address this scale ambiguity, we implement

a dedicated translation estimation module that establishes a consistent metric scale across frames by leveraging the depth predictions from the MDEM. The key insight is that the depth maps generated by the MDEM provide an implicit scale that, when used to compute 3D correspondences between frames, allows us to recover a scaled translation vector t_scaled that maintains consistency throughout the sequence. The module computes a scaled translation vector through the following process:

1. **Feature Extraction and Matching**: the module extracts visual features (SIFT or ORB) from consecutive RGB frames and establishes correspondences between them. These features serve as anchor points for computing the scale-consistent transformation.

2. **Depth-based 3D Correspondence**: for each matched feature pair, the module retrieves the corresponding depth values from the MDEM-generated depth maps. These depth values transform the 2D correspondences into 3D correspondences through back-projection:

$$P_{3D} = \begin{bmatrix} (u - c_x) \times d/f_x \\ (v - c_y) \times d/f_y \\ d \end{bmatrix} \tag{2.5}$$

   where (u, v) are the pixel coordinates, d is the predicted depth from the MDEM, (f_x, f_y) are the focal lengths, and (c_x, c_y) is the principal point. Crucially, the predicted depth d establishes the metric scale that ensures consistency across the entire reconstruction.

3. **Scale-Consistent Translation Estimation**: using the 3D correspondences from both frames, the module computes the translation that aligns these points while maintaining consistency with the depth scale:

$$t_{\text{scaled}} = \frac{1}{N} \sum_{i=1}^{N} \left(P^{(i)} - P^{(i-1)} \right) \tag{2.6}$$

   where N is the number of valid 3D correspondences. This translation vector t_{scaled} inherits the scale from the depth predictions, ensuring that the pose estimates are compatible with the 3D point clouds generated from the same depth maps.

4. Unscented Kalman Filter (UKF) Fusion: an UKF fuses the unscaled translation vector $t_unscaled$ obtained from the MPEM with the scaled translation vector t_scaled combining the accurate directional information from the deep learning model with the metric scale from the depth-based correspondences. The state vector of the UKF is defined

as:

$$x = [t_x, t_y, t_z]^T \tag{2.7}$$

where t_x, t_y, and t_z are the components of the translation vector. The measurement model of the UKF is given by:

$$z = [t_{scaled_x}, t_{scaled_y}, t_{scaled_z}]^T \tag{2.8}$$

where t_{scaled_x}, t_{scaled_y}, and t_{scaled_z} are the components of the scaled translation vector obtained from the RGB-D odometry or the scaling factor computation.

The UKF operates in two main steps:

1. prediction step: the UKF predicts the next state $\hat{x}$ using the unscaled translation vector $t_{unscaled}$ obtained from the MPEM:

$$\hat{x} = f(x, t_{unscaled}) \tag{2.9}$$

where f is the state transition function that uses the MPEM to predict the next state.

2. Update step: the UKF updates the predicted state $\hat{x}$ using the scaled translation vector t_{scaled} obtained from the RGB-D odometry or the scaling factor computation:

$$x = \hat{x} + K(z - h(\hat{x})) \tag{2.10}$$

where K is the Kalman gain, z is the measurement vector containing the scaled translation, and h is the measurement function that maps the predicted state to the measurement space.

The corrected translation vector $t_{corrected}$ is then obtained from the updated state of the UKF:

$$t_{corrected} = [x_1, x_2, x_3]^T \tag{2.11}$$

where x_1, x_2, and x_3 are the scaled translation components of the updated state vector.

To mitigate accumulated drift errors, the pose refinement process employs pose graph optimization to ensure global consistency across the entire trajectory. The pose graph maintains a representation of all camera poses and their relative transformations as a graph structure where nodes represent camera poses and edges encode the spatial constraints between them. The pose graph optimization problem is formulated as a nonlinear least squares

minimization:

$$\arg\min_{\mathbf{T}_i} \sum_{(i,j)\in\mathcal{E}} \left\| \log\left(\hat{\mathbf{T}}_{ij}^{-1} \cdot \mathbf{T}_i^{-1} \cdot \mathbf{T}_j\right) \right\|_{\mathbf{\Omega}_{ij}}^2 \tag{2.12}$$

where $\mathbf{T}_i$ and $\mathbf{T}_j$ are the global poses of frames i and j, $\hat{\mathbf{T}}_{ij}$ is the measured relative transformation between them, $\mathcal{E}$ is the set of edges in the pose graph, and $\mathbf{\Omega}_{ij}$ is the information matrix encoding measurement uncertainty for edge (i, j). Our optimization strategy operates at two temporal scales:

- **Local Optimization**: every 500 frames, we trigger pose graph optimization using the Levenberg-Marquardt algorithm to redistribute accumulated errors across recent poses. This periodic optimization is necessary because individual pose estimates accumulate small errors that compound over time, degrading the overall trajectory accuracy. The information matrix for each edge is computed based on the confidence of feature matches and depth measurements, with edge pruning threshold of 0.05 to remove unreliable constraints that could introduce additional errors into the optimization process.

- **Global Map Update**: Following pose graph optimization, the entire 3D reconstruction is updated by re-integrating all point clouds with the optimized poses into the TSDF volume. This complete reconstruction update is required to ensure consistency between the refined trajectory and the reconstructed geometry, as the pose corrections would otherwise create misalignments between the camera positions and the previously integrated 3D structure.

Point Cloud Construction and Integration

This module converts the 2D endoscopic images into 3D point clouds using the pseudo-RGBD frames generated by combining RGB images with depth maps from the MDEM. The construction process involves back-projecting each pixel to 3D space using the camera intrinsic parameters and the predicted depth values from the depth estimation module. Upon constructing the individual point clouds, the module aligns and integrates the newly acquired point cloud with the pre-existing reconstructed scene, guided by the refined pose estimates. The module utilizes the Truncated Signed Distance Function (TSDF) [21] technique to convert the point cloud data into a comprehensive volumetric representation. The TSDF is a volumetric representation that stores the signed distance to the nearest surface at each voxel. It allows for efficient fusion of depth information from multiple viewpoints and enables the generation of a smooth and continuous surface reconstruction.

2.4 Materials

In this study, we employed a comprehensive approach to train, validate, and test our proposed method, utilizing the diverse datasets listed in table 2.1 to ensure robustness and generalizability in endoscopic applications.

Dataset	Depth	Pose	Procedure	Frames
Hamlyn [54]	✓		Laparoscopy	78,160
EndoSLAM [51]		✓	Gastroscopy, Colonoscopy	76,837
SCARED [1]	✓	✓	Laparoscopy	17,206

Table 2.1: Overview of endoscopic datasets used for evaluating depth estimation and pose estimation modules in MVSLAM framework. Checkmarks indicate the specific modules tested on each dataset

2.4.1 Training Datasets for CycleVO

The training and validation processes were performed on the CycleVO module because it was the only model we trained in this study. Two distinct datasets were utilized: an extensive internal dataset for training and selected subsets from the EndoSLAM dataset [51] for validation. For training, we leveraged a large internal unlabeled dataset comprising more than 300 hours of gastroscopy and prostatectomy videos, collected from 100 patients undergoing these procedures at Fondazione Policlinico Universitario Campus Bio-Medico. In total, this training dataset contained 2,250,900 frames, allowing us to effectively train our pose estimation network on a substantial amount of endoscopic data. We prepared training data by extracting individual frames and center-cropping them to 128x128 pixels. This approach enhances robustness by training with a reduced receptive field while avoiding resizing distortions. This preserves relevant visual information while maintaining spatial integrity. To validate CycleVO, we used the following subsets from the EndoSLAM dataset, which are completely separate from the training data: HighCam Colon IV Tumor-Free Trajectory 1, LowCam Colon IV Tumor-Free Trajectory 1, HighCam Colon IV Tumor-Free Trajectory 5, and LowCam Colon IV Tumor-Free Trajectory 5. These subsets account for four video sequences with a total of 3024 frames, providing a diverse range of scenarios to assess CycleVO's inference performance. Importantly, there is no overlap between training and validation datasets, as EndoSLAM subsets were used exclusively for validation.

2.4.2 Testing Datasets for Depth and Pose Estimation Modules

We evaluated our depth and pose estimation modules using three comprehensive datasets representing the largest publicly available endoscopic MVSLAM datasets. These datasets, summarized in Table 2.1, provide diverse scenarios and data type to assess the robustness and generalizability of our method:

1. Hamlyn Dataset [54]: We employed an enhanced version of the Hamlyn dataset, which contains stereo images from endoscopic interventions on porcines. This dataset, augmented with depth ground truth derived from stereo images using Libelas, consists of 78,160 frames across 20 videos, making it suitable for testing the depth estimation module of our method. However, it lacks pose ground truth, preventing pose estimation evaluation.

2. EndoSLAM Dataset [51]: In addition to its use in validating CycleVO, the EndoSLAM dataset was also employed for testing the pose estimation module of our method. The subsets used for validating CycleVO were excluded from the testing process to maintain a clear separation between validation and testing data. While the EndoSLAM dataset includes depth data, we did not use it for testing depth estimation due to its synthetic nature. The dataset comprises both ex-vivo and synthetic porcine data, with 35 sub-datasets totaling 76,837 frames spanning various gastrointestinal areas, providing pose data for a comprehensive evaluation of our method's performance in pose estimation.

3. SCARED Dataset [1]: The SCARED dataset, recorded using a da Vinci Xi surgical robot, consists of 7 training and 2 test sets, each corresponding to a single porcine subject. Each set contains 4 to 5 keyframes representing distinct scene views, with structured light patterns projected for dense stereo reconstruction. The movement of the endoscope enabled the capture of camera poses relative to the robot base, and the projection of 10-bit Gray code patterns onto the scene ensured unique pixel encoding for efficient stereo matching and depth recovery. The SCARED dataset provides both depth maps and ground truth poses, making it suitable for testing both the depth and pose estimation modules of our method.

By leveraging these diverse datasets, we were able to thoroughly test our depth estimation and pose estimation modules across various scenarios and data types, ensuring their robustness and generalizability in endoscopic applications.

2.5 Experimental Configuration

We validated our MVSLAM approach through comprehensive comparison against four state-of-the-art endoscopic MVSLAM methods.

EndoSLAM[1] [51], also known as Endo-SfMLearner, is an unsupervised monocular visual odometry and depth estimation approach designed specifically for endoscopic videos. It employs a DepthNet based on a U-Net [56] architecture with ResNet-18 encoder to predict dense disparity maps from single images. The PoseNet estimates six Degree of freedom relative camera poses between consecutive frames. Key innovations include a spatial attention module to focus on highly textured regions and an affine brightness transformer to handle illumination changes.

EndoDepth[2] [54] presents an approach for 3D reconstruction and camera motion estimation in monocular endoscopic sequences. It utilizes self-supervised depth networks, photometric tracking, and volumetric fusion to generate pseudo-RGBD frames, track camera pose, and fuse depth maps into a 3D model. The depth estimation uses the Monodepth2 [29] network architecture with a ResNet-18 [33] encoder. Camera poses are estimated by minimizing photometric error between frames.

Shao et al.[3] [71] propose a self-supervised framework for monocular depth and egomotion estimation in endoscopy that introduces the concept of appearance flow to handle brightness inconsistency. Their method, also known as AFSfMlearner, comprises four modules. The Structure module (DepthNet) uses a ResNet-18 encoder-decoder architecture to estimate dense depth maps. The Motion module (PoseNet) is a lightweight network that predicts six Degree of freedom relative camera poses between frames. The Appearance module predicts appearance flow and calibrates brightness, using a structure similar to DepthNet. The Correspondence module performs automatic registration to enhance performance, with architecture similar to the appearance module. The appearance flow allows relaxation of the brightness constancy assumption, addressing severe brightness fluctuations common in endoscopic scenarios.

OneSLAM[4] [77] presents a generalized SLAM approach for monocular endoscopic imaging based on tracking any point. It leverages a tracking any point foundation model to obtain sparse correspondences across multiple frames. These are used to perform local bundle adjustment to jointly optimize camera poses and 3D structure. OneSLAM demonstrates strong cross-domain performance on sinus endoscopy, colonoscopy, arthroscopy and

[1] https://github.com/CapsuleEndoscope/EndoSLAM
[2] https://github.com/UZ-SLAMLab/Endo-Depth-and-Motion
[3] https://github.com/ShuweiShao/AF-SfMLearner
[4] https://github.com/arcadelab/OneSLAM

laparoscopy data without retraining.

All competitor methods were used without modifications. This varied baseline enables thorough assessment of our approach's generalization capabilities across different endoscopic scenarios.

2.5.1 Performance Metrics and Statistical Analysis

We employed well-established metrics for depth and pose estimation evaluation. For depth estimation, we selected the Absolute Relative Difference, Root Mean Square Error (RMSE), RMSE log, Squared Relative Error, and accuracy thresholds of 1.25, 1.25^2, and 1.25^3, as described in [48] and presented in Table 2.2.

Table 2.2: Key metrics for evaluating depth prediction performance. d_i and d_i^* denote the predicted and ground-truth depth at pixel i, respectively, and N is the number of valid pixels.

Metric	**Formula**	**Description**
Absolute Relative Difference (AbsRel)	$\frac{1}{N}\sum_i \frac{\lvert d_i - d_i^*\rvert}{d_i^*}$	Mean of the per-pixel absolute errors normalised by the ground-truth depth; penalises errors on nearby structures.
Root Mean Square Error (RMSE)	$\sqrt{\frac{1}{N}\sum_i (d_i - d_i^*)^2}$	Standard quadratic error in metric units; emphasises large deviations.
RMSE (log)	$\sqrt{\frac{1}{N}\sum_i (\log d_i - \log d_i^*)^2}$	Quadratic error computed in log space; balances errors at near and far depths.
Squared Relative Error (SqRel)	$\frac{1}{N}\sum_i \frac{(d_i - d_i^*)^2}{d_i^*}$	Relative quadratic error; weights large relative deviations more heavily.
Accuracy threshold $\delta < 1.25^k$	$\frac{1}{N}\left\lvert\{\, i : \max\left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}\right) < 1.25^k \,\}\right\rvert$	Fraction of pixels whose predicted depth is within a factor 1.25^k of the ground truth ($k = 1, 2, 3$); higher is better.

To measure the pose estimation performance, we utilized the Relative Translation Error (RTE), Relative Rotation Error (RRE), and Absolute Translation Error (ATE), as shown in Table 2.3. These metrics are widely used in the literature for evaluating the accuracy of pose estimation in SLAM systems [67]. We also conducted statistical analysis to determine significance of performance differences between our MVSLAM approach and benchmark models. We verified that the data for each model meets the assumptions required for using

Table 2.3: Comparative overview of trajectory evaluation metrics used to assess the monocular pose estimation module. T_i and T_i^* denote the estimated and ground-truth poses at frame i, respectively, Δ_i is the relative transform between consecutive frames, and n is the number of frames.

Metric	**Formula**	**Description**
Absolute Trajectory Error (ATE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\| \operatorname{trans}(T_i^{*\,-1}\, S\, T_i) \right\|^2}$	Root-mean-square translational distance between estimated and ground-truth trajectories after global rigid alignment S; measures overall trajectory consistency.
Relative Translation Error (RTE)	$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n-1}\left\| \operatorname{trans}(\Delta_i^{*\,-1}\Delta_i) \right\|^2}$	Root-mean-square translational error between successive relative transforms; captures local drift in position.
Relative Rotation Error (RRE)	$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n-1}\left\| \operatorname{rot}(\Delta_i^{*\,-1}\Delta_i) \right\|^2}$	Root-mean-square rotational error between successive relative transforms; captures local drift in orientation.

an independent two-sample T-test, including normality and equal variances. We then performed pairwise comparisons between the MVSLAM approach and each benchmark model separately, using a significance level of $p = 0.05$.

2.6 Results

In this section, we present the results of our comprehensive evaluation aimed at validating the efficacy and robustness of our proposed MVSLAM approach. Initially, in section 2.6.1, we provide an analysis of the monocular pose estimation module. This is followed by section 2.6.2, where we explore the application and validation of the depth estimation module. It's worth noting that our preprocessing methods for certain datasets differ from some competitors, which may lead to variations in results. For instance, we used a version of the Hamlyn dataset that had been preprocessed using libellas, as this preprocessed version has been utilized in other notable works such as [54]. This choice ensures consistency and comparability with existing literature in the field. For the SCARED Dataset, we derived depth information through stereo rectification and disparity estimation, as opposed to extracting it directly from the point cloud as some other studies have done. We opted for this approach because it allows us to obtain a dense depth map, in contrast to the sparse depth

information typically yielded by point cloud extraction. It should be noted that the 3D Mapping Module (3DM) cannot be directly evaluated due to two key factors. First, the lack of ground truth 3D models for the tested environments hinders a comprehensive validation of the reconstruction quality. Second, even if ground truth 3D models were available, the scale ambiguity inherent in the monocular pose and depth estimation modules would render the reconstructed model incomparable due to differing scales. However, this does not limit our approach, as the 3DM employs state-of-the-art algorithms rigorously validated in prior work [21, 95]. Furthermore, the 3D mapping and localization capabilities critically depend on the performance of the pose estimation and depth estimation modules, which are thoroughly evaluated in the following sections. The successful functioning of the overall SLAM system thus implicitly validates the efficacy of the 3DM. While quantitative evaluation of the 3DM is not feasible, Figure 2.4 demonstrates the qualitative performance of the integrated BodySLAM system, showing consistent camera pose estimation across temporal frames on the Hamlyn dataset.

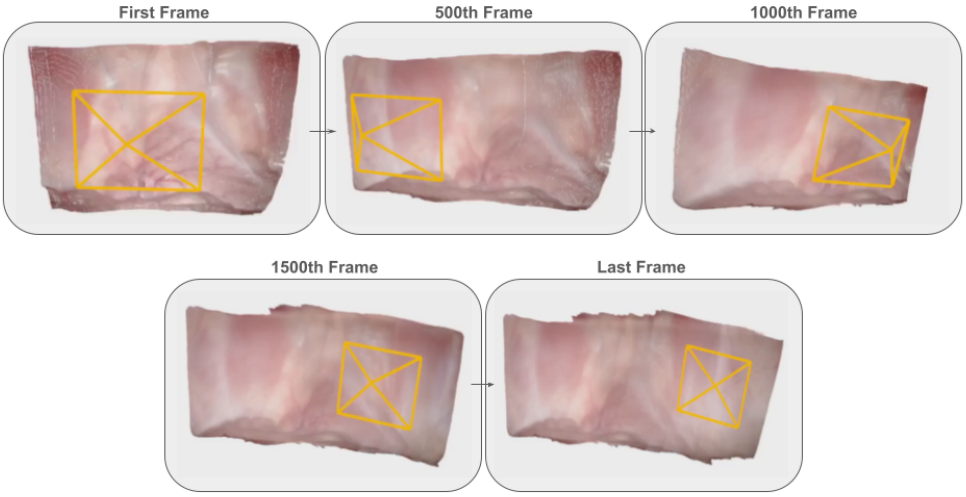

Figure 2.4: Qualitative results of BodySLAM on a sequence of the Hamlyn dataset. Yellow frustums represent the estimated camera pose and orientation at frames 1, 500, 1000, 1500, and the final frame, demonstrating consistent pose tracking throughout the endoscopic sequence.

2.6.1 Monocular Pose Estimation Module (MPEM)

Figure 2.5 presents the performances of the monocular pose estimation module measured in terms of ATE, RTE, and RRE on the two datasets (SCARED and EndoSLAM) when

adopting each of the five methods, i.e., EndoSfmLearner, EndoDepth, AFSfMlearner, OneSLAM and CycleVO. When interpreting these results, it's important to consider the training backgrounds of the methods. EndoSLAM was trained on the EndoSLAM dataset, while AFSfMLearner was trained on the SCARED dataset. This prior exposure to the test data or related datasets may introduce a potential bias in their performance, which should be taken into account when interpreting their results. In contrast, CycleVO has not been exposed to either dataset during its training.

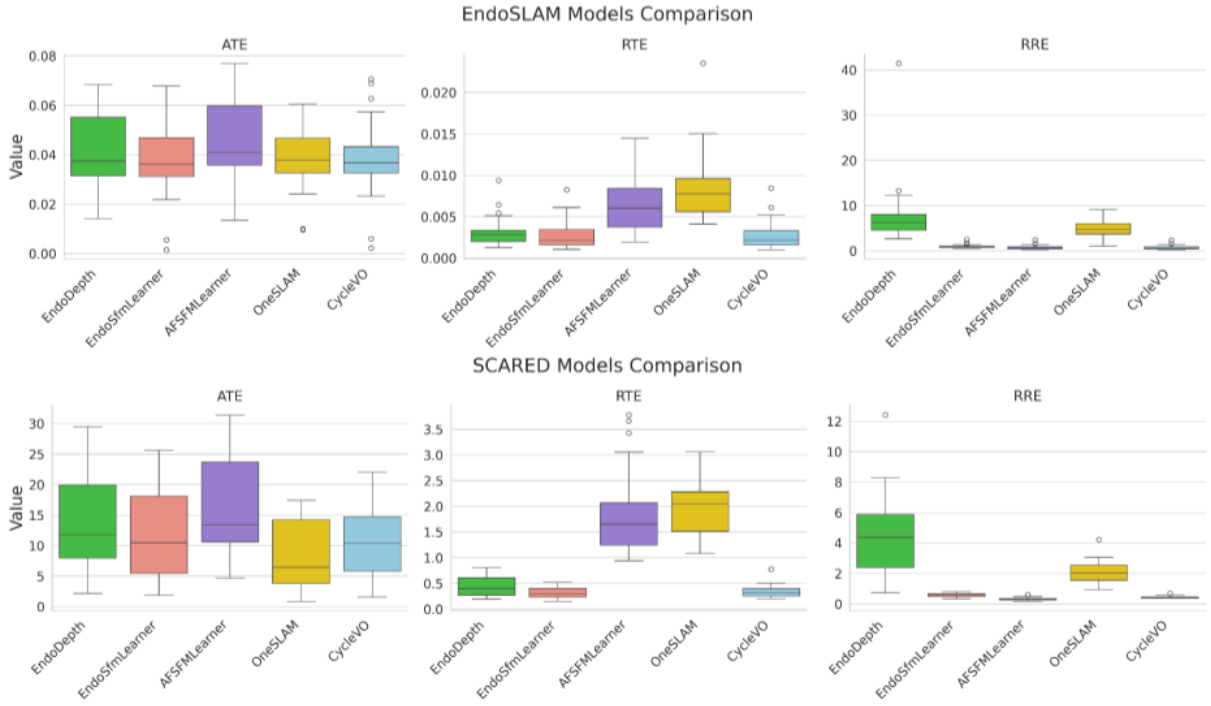

Figure 2.5: Comparison of the performance of EndoDepth, EndoSfmLearner, AFSfMlearner, OneSLAM and CycleVO algorithms on the SCARED and ENDOSLAM datasets. Metrics evaluated include Absolute Trajectory Error (ATE), Relative Trajectory Error (RTE), and Relative Rotation Error (RRE). Each box plot shows the distribution of the errors for the respective algorithms, highlighting the median, interquartile range, and outliers.

In the case of trajectory errors, being either an absolute or a relative error, both rows show that CycleVO and EndoSfmLearner have lower median scores and interquartile ranges than EndoDepth, AFSfMlearner and OneSLAM. The figure compares the three methods across the two datasets and three metrics. The rows represent the SCARED and ENDOSLAM datasets, while the columns show the ATE, RTE, and RRE metrics respectively. Regarding the statistical analysis, Table 2.4 shows the pairwise comparisons between CycleVO and the other methods. Notably, CycleVO achieves competitive performance without any prior exposure to either the EndoSLAM or SCARED datasets during its training, highlighting its strong generalization capabilities. It is also interesting to note that our CycleVO approach

	Metric	Comparison	T-Stat	P-Value
SCARED	ATE	BodySLAM vs EndoDepth	-1.12	0.2693
		BodySLAM vs EndoSfMLearner	-0.59	0.5593
		BodySLAM vs OneSLAM	1.18	0.2439
		BodySLAM vs AFSfMlearner	-2.44	**0.0205**
	RTE	BodySLAM vs EndoDepth	-1.65	0.1075
		BodySLAM vs EndoSfMLearner	0.63	0.5335
		BodySLAM vs OneSLAM	-11.79	**1.81e-10**
		BodySLAM vs AFSfMlearner	-7.01	**1.20e-06**
	RRE	BodySLAM vs EndoDepth	-6.23	**6.99e-06**
		BodySLAM vs EndoSfMLearner	-3.53	**0.0014**
		BodySLAM vs OneSLAM	-9.00	**3.70e-08**
		BodySLAM vs AFSfMlearner	4.17	**0.0002**
ENDOSLAM	ATE	BodySLAM vs EndoDepth	-1.02	0.3122
		BodySLAM vs EndoSfMLearner	-0.11	0.9087
		BodySLAM vs OneSLAM	-0.34	0.7343
		BodySLAM vs AFSfMlearner	-2.68	**0.0090**
	RTE	BodySLAM vs EndoDepth	-1.29	0.2012
		BodySLAM vs EndoSfMLearner	-0.01	0.9942
		BodySLAM vs OneSLAM	-9.79	**7.24e-14**
		BodySLAM vs AFSfMlearner	-7.60	**1.73e-10**
	RRE	BodySLAM vs EndoDepth	-7.59	**1.73e-09**
		BodySLAM vs EndoSfMLearner	-3.85	**0.0002**
		BodySLAM vs OneSLAM	-15.39	**5.06e-20**
		BodySLAM vs AFSfMlearner	-0.63	0.5299

Table 2.4: Comparative analysis of BodySLAM against EndoDepth, EndoSfMLearner, OneSLAM, and AFSfMlearner across different metrics (ATE, RTE, RRE) for SCARED and ENDOSLAM datasets. T-Statistic and P-Value are reported for each comparison. Bold P-Values indicate statistically significant differences (p ¡ 0.05).

not only performs as the others but it also has the lowest inference time (Table 2.5).

Turning our attention to RRE, i.e., a score measuring the angular deviation in the rotation component, we notice that CycleVO outperforms both EndoSfmLearner, EndoDepth, and OneSLAM on both datasets (both rows and last column of Figure 2.5). Furthermore, these performance differences are also statistically significant (last rows for both datasets of Figure 2.5). While CycleVO shows statistically significant superior RRE performance against AFSfMLearner on the SCARED dataset, but not on the ENDOSLAM dataset. To deepen the results we also perform an ablation study that aims to investigate the effectiveness of the new terms that we introduced for the cycle loss (section 2.3.1). To this end, we ran four experiments, two per dataset, setting $\lambda_2 \neq 0$, $\lambda_2 = 0$ (Figure 2.6). Both in the case

Model	Dataset	Average Time [s]
EndoDepth	EndoSLAM	1140.56
	SCARED	934.18
EndoSFMLearner	EndoSLAM	38.95
	SCARED	97.72
AFSfMlearner	EndoSLAM	14.63
	SCARED	59.95
OneSLAM	EndoSLAM	169.77
	SCARED	336.65
CycleVO	EndoSLAM	**3.32**
	SCARED	**38.57**

Table 2.5: Comparison of models, datasets, and execution times. This table presents the processing times for different models across two datasets, EndoSLAM and SCARED, highlighting the computational efficiency of each approach. All computations were performed on an NVIDIA RTX 3070 GPU.

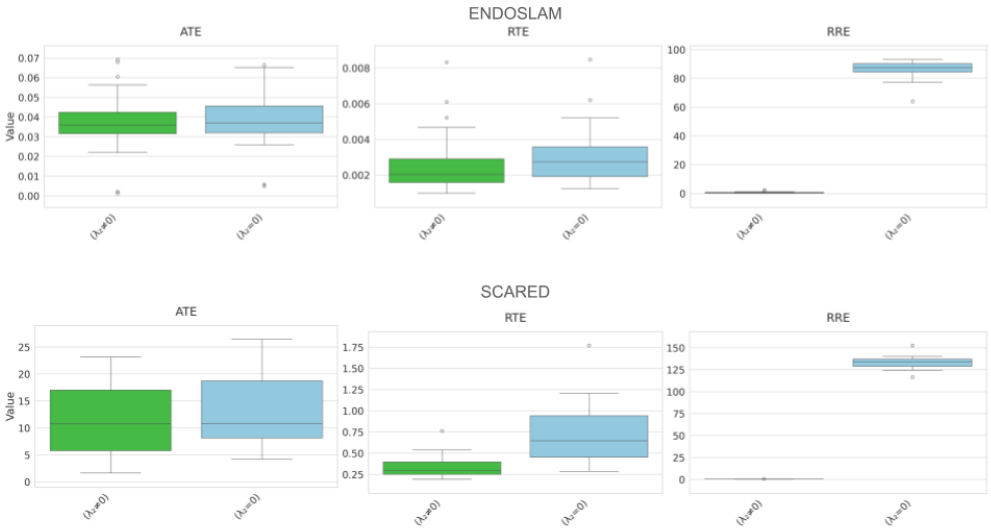

Figure 2.6: Ablation study of cycle loss term modifications. This figure presents the impact of different cycle loss term settings ($\lambda_2 \neq 0$ vs $\lambda_2 = 0$) on the performance of models evaluated on the SCARED and ENDOSLAM datasets. The metrics assessed include Absolute Trajectory Error (ATE), Relative Trajectory Error (RTE), and Relative Rotation Error (RRE). Each box plot illustrates the distribution of the errors for the respective models, highlighting the median, interquartile range, and outliers.

Metric	Comparison	T-Stat	P-Value
Absolute Relative Difference	BodySLAM vs EndoDepth	-2.18	**0.0366**
	BodySLAM vs EndoSfMLearner	-4.24	**0.0002**
	BodySLAM vs AFSfMLearner	-12.50	**3.55e-13**
RMSE	BodySLAM vs EndoDepth	-1.99	0.0550
	BodySLAM vs EndoSfMLearner	-2.86	**0.0069**
	BodySLAM vs AFSfMLearner	-6.26	**2.73e-06**
RMSE Log	BodySLAM vs EndoDepth	-2.13	**0.0400**
	BodySLAM vs EndoSfMLearner	-4.42	**9.94e-05**
	BodySLAM vs AFSfMLearner	-17.08	**1.29e-15**
Squared Relative Error	BodySLAM vs EndoDepth	-2.15	**0.0422**
	BodySLAM vs EndoSfMLearner	-2.71	**0.0105**
	BodySLAM vs AFSfMLearner	-4.80	**0.0001**
accuracy (1.25)	BodySLAM vs EndoDepth	5.14	**4.42e-05**
	BodySLAM vs EndoSfMLearner	2.55	**0.0151**
	BodySLAM vs AFSfMLearner	14.17	**2.55e-13**
accuracy $(1.25)^2$	BodySLAM vs EndoDepth	4.12	**0.0006**
	BodySLAM vs EndoSfMLearner	1.54	0.1396
	BodySLAM vs AFSfMLearner	10.38	**4.62e-09**
accuracy $(1.25)^3$	BodySLAM vs EndoDepth	3.56	**0.0022**
	BodySLAM vs EndoSfMLearner	1.37	0.1886
	BodySLAM vs AFSfMLearner	14.10	**3.62e-11**

Table 2.6: Statistical Performance Analysis of the Monocular Depth Estimation Module on the Hamlyn Dataset. The table details outcomes for Absolute Relative Difference, Root Mean Square Error (RMSE), RMSE Log, Squared Relative Error, and accuracy thresholds (accuracy ¡ 1.25, accuracy ¡ 1.25^2, accuracy ¡ 1.25^3). Results include pairwise comparisons of BodySLAM versus EndoDepth, EndoSfMLearner, and AFSfMLearner, assessed through t-tests with a significance level of p ¡ 0.05.

of trajectory errors, being either an absolute or a relative error, and the relative rotation error, both rows show that the model with the new term ($\lambda_2 \neq 0$) has lower median scores and interquartile ranges than the counterpart without the custom term ($\lambda_2 = 0$).

2.6.2 Monocular Depth Estimation Module (MDEM)

Figure 2.7, 2.8 present the performances of the monocular depth estimation module, BodySLAM, compared to EndoSfmLearner, EndoDepth and AFSfMLearner respectively on the Hamlyn and SCARED datasets. The figures compare the three methods across multiple evaluation metrics for each dataset. The x-axis represents the different methods (EndoSfmLearner, EndoDepth, AFSfMLearner and BodySLAM), while the y-axis shows the values for the evaluation metrics.

It's worth noting that OneSLAM is not included in this comparison. This is because

Metric	Comparison	T-Stat	P-Value
Absolute Relative Difference	BodySLAM vs EndoDepth	-9.69	**1.81e-09**
	BodySLAM vs EndoSfMLearner	-4.14	**1.99e-04**
	BodySLAM vs AFSfMlearner	-4.36	**3.39e-04**
RMSE	BodySLAM vs EndoDepth	-1.01	0.3212
	BodySLAM vs EndoSfMLearner	-0.25	0.8007
	BodySLAM vs AFSfMlearner	-0.78	0.4392
RMSE Log	BodySLAM vs EndoDepth	-14.99	**2.72e-13**
	BodySLAM vs EndoSfMLearner	-8.35	**1.05e-09**
	BodySLAM vs AFSfMlearner	-13.13	**4.65e-11**
Squared Relative Error	BodySLAM vs EndoDepth	-0.85	0.4069
	BodySLAM vs EndoSfMLearner	0.31	0.7601
	BodySLAM vs AFSfMlearner	-0.35	0.7305
accuracy (1.25)	BodySLAM vs EndoDepth	13.10	**3.61e-11**
	BodySLAM vs EndoSfMLearner	5.82	**6.53e-06**
	BodySLAM vs AFSfMlearner	9.36	**1.69e-08**
accuracy $(1.25)^{\mathbf{2}}$	BodySLAM vs EndoDepth	10.13	**7.29e-09**
	BodySLAM vs EndoSfMLearner	2.56	**0.0195**
	BodySLAM vs AFSfMlearner	6.26	**6.59e-06**
accuracy $(1.25)^{\mathbf{3}}$	BodySLAM vs EndoDepth	8.53	**9.75e-08**
	BodySLAM vs EndoSfMLearner	1.26	0.2228
	BodySLAM vs AFSfMlearner	5.47	**3.40e-05**

Table 2.7: Statistical Performance Analysis of the Monocular Depth Estimation Module on the SCARED Dataset. The table details outcomes for Absolute Relative Difference, Root Mean Square Error (RMSE), RMSE Log, Squared Relative Error, and accuracy thresholds (accuracy ¡ 1.25, accuracy ¡ $1.25^{\mathbf{2}}$, accuracy ¡ $1.25^{\mathbf{3}}$). Results include pairwise comparisons of BodySLAM versus EndoDepth, EndoSfMLearner, and AFSfMlearner, assessed through t-tests with a significance level of p ¡ 0.05.

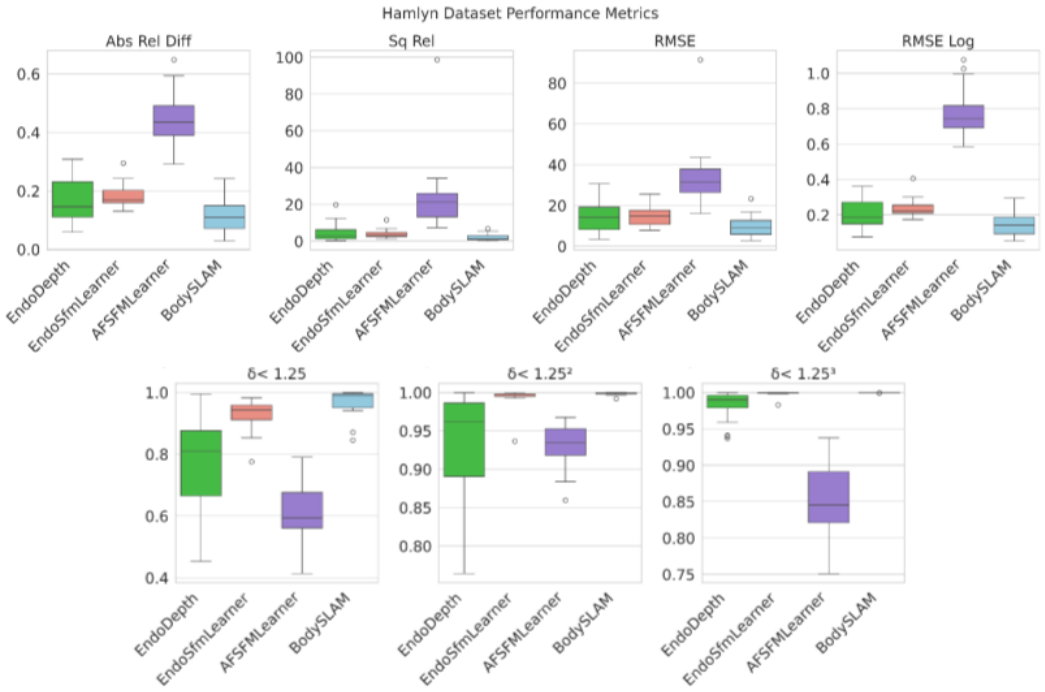

Figure 2.7: Comparison of depth estimation performance for EndoDepth, EndoSfmLearner, AFSfMlearner and BodySLAM across various metrics on Hamlyn Dataset. Box plots show that BodySLAM achieves the best performance with the lowest median values for absolute relative difference, RMSE, RMSE log, and squared relative difference, and the highest median values for accuracy thresholds $\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$. EndoSfmLearner demonstrates the second best performance, while EndoDepth shows the poorest results among the three methods compared.

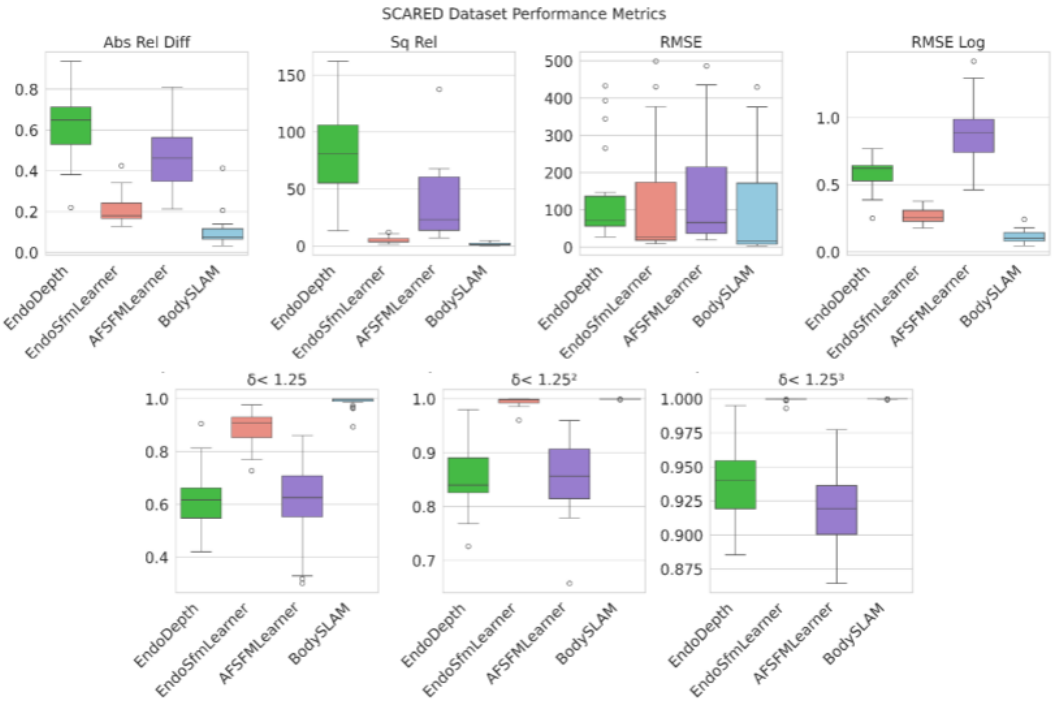

Figure 2.8: Comparison of depth estimation performance for EndoDepth, EndoSfmLearner, AFSfMlearner and BodySLAM across various metrics on SCARED Dataset. Box plots show that BodySLAM achieves the best performance with the lowest median values for absolute relative difference, RMSE log, and squared relative difference, and the highest median values for accuracy thresholds $\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$. EndoSfmLearner demonstrates the second best performance, while EndoDepth shows the poorest results among the three methods compared. Outliers beyond 1.5 times the interquartile range from the quartiles have been excluded to enhance clarity and focus on the central trends of the data.

OneSLAM produces sparse depth maps, while the other methods generate dense depth maps. The evaluation metrics used in this study are designed for dense depth maps, making a direct comparison with OneSLAM's sparse output impossible. When interpreting these results, it's important to consider the training backgrounds of the methods. EndoDepth was trained on the Hamlyn dataset, while AFSfMLearner was trained on the SCARED dataset. This prior exposure to the test data or related datasets may introduce a potential bias in their performance, which should be taken into account when interpreting their results. In contrast, BodySLAM has not been exposed to either dataset during its training, which highlights its generalization capabilities.

In the case of the Hamlyn dataset (Figure 2.7), BodySLAM outperforms all competitors across all metrics. For Absolute relative difference, RMSE, RMSE log and Squared Relative error (first row, Figure 2.7), BodySLAM exhibits lower median values and smaller interquartile ranges compared to the other methods. Regarding the threshold accuracy (second row, Figure 2.7), where values close to or equal to 1 are considered optimal, BodySLAM shows a median closest to this optimal value and a smaller interquartile range compared to the others. Similarly, for the SCARED dataset (Figure 2.8), BodySLAM outperforms all competitors across all metrics. As with the Hamlyn dataset, BodySLAM demonstrates lower median values and smaller interquartile ranges for Absolute relative difference, RMSE, RMSE log and Squared Relative error (first row, Figure 2.8). For the threshold accuracy (second row, Figure 2.8), BodySLAM median is closest to the optimal value of 1, with a smaller interquartile range compared to the other methods. To deepen the results we also performed a statistical analysis presented in Table 2.6 for the Hamlyn dataset and in Table 2.7 for the SCARED Dataset. The statistical analysis presented on the Hamlyn dataset shows that BodySLAM outperforms the other depth estimation models, except for the threshold accuracy $(1.25)^2$ and $(1.25)^3$, where BodySLAM does not significantly outperform EndoSfmLearner. For the SCARED dataset, BodySLAM outperforms the other depth estimation models in most metrics. However, there are a few exceptions: in RMSE, BodySLAM does not statistically outperform EndoSfmLearner; in Squared Relative Error, BodySLAM does not statistically outperform both EndoDepth and EndoSfmLearner; and in Accuracy $(1.25)^3$, BodySLAM does not statistically outperform EndoSfmLearner.

2.7 Conclusions

This chapter presented a robust deep learning-based SLAM approach that can effectively operate across various endoscopic surgical settings, including laparoscopy, gastroscopy, and colonoscopy. We addressed the challenges posed by hardware limitations and environmental

variations by integrating deep learning models with strong generalization capabilities into the framework. Our results demonstrate that our novel unsupervised pose estimation method, CycleVO, exhibits robust generalization capabilities and competitive performance compared to existing state-of-the-art pose estimation techniques, while maintaining the lowest inference time. This makes CycleVO a promising solution for real-time surgical applications. An ablation study conducted on the pose estimation module highlighted the effectiveness of new terms introduced for cycle loss, with the model incorporating the custom term consistently outperforming its counterpart without the term. These findings were confirmed by statistical analysis revealing that the CycleVO approach performed comparably to other methods while having the lowest inference time. Furthermore, we demonstrated that Zoe's exceptional generalization abilities, previously showcased in various imaging contexts, can be effectively adapted to the endoscopic field. When compared to current state-of-the-art depth estimation algorithms in endoscopy, Zoe exhibited superior performance, confirming its potential for SLAM applications in surgical environments. These findings are consistent with the results reported by [31]. Statistical analysis revealed that Zoe significantly outperformed other depth estimation models in most metrics, with only a few exceptions where the differences were not statistically significant.

Despite these promising results, our study has four limitations. First, although we validated our approach on three relevant endoscopic scenarios, its performance in other settings remains to be investigated. Future research should focus on collecting diverse datasets from a wider range of endoscopic procedures and fine-tuning the deep learning models to adapt to the unique characteristics of each domain, which could help extend the framework's applicability. Beyond broadening the range of procedures, future data collection could also target a more direct, quantitative validation of the 3D reconstruction stage, which could not be performed in this study due to the absence of ground truth surface reconstructions. Two complementary strategies could be pursued: surrogate geometric-consistency measures that do not require external ground truth, such as multi-view reprojection error, photometric re-rendering consistency, and surface-to-surface distance between reconstructions obtained from overlapping sub-sequences, which quantify the internal coherence and repeatability of the recovered geometry; and phantom-based experiments, in which anatomically realistic models with known geometry are imaged endoscopically and reconstructed offline, supplying a controlled ground truth against which absolute reconstruction accuracy and geometric consistency could be assessed once the scale factor is resolved. Together, the reporting of scale-invariant surrogate metrics and the design of a phantom-based validation protocol constitute a natural next step toward a fully quantitative assessment of the reconstruction stage. Second, the impact of varying illumination conditions and the presence of fluids or debris

on the approach's performance needs further examination. Investigating domain adaptation techniques and developing specialized modules for handling procedure-specific challenges, such as varying illumination and the presence of fluids, could improve the approach's robustness. Third, the scale ambiguity problem inherent to monocular cameras may affect both the depth estimation and pose estimation modules, resulting in a 3D model that has a scale that differs from the real scale. Exploring methods to mitigate the scale ambiguity problem, such as incorporating additional sensors or prior knowledge, could enhance the accuracy of depth estimation and pose estimation, and enable more reliable 3D reconstruction validation. Fourth, we have in-silico validated our approach and we are aware that the practical application of the approach in real-world surgical scenarios requires extensive validation and collaboration with medical experts. Conducting rigorous clinical trials involving diverse patient populations and surgical conditions, in close collaboration with medical experts, will be essential for validating the approach's performance and ensuring its seamless integration into the surgical workflow. Future research will focus on optimizing the computational efficiency of our framework, particularly through model compression techniques such as knowledge distillation, quantization, and pruning of the Zoe architecture. Additionally, we plan to explore lightweight alternatives that maintain the generalization capabilities while reducing memory footprint and inference time, making the system suitable for real-time deployment on embedded surgical platforms with limited computational resources. In spite of these limitations, the proposed SLAM approach has the potential to significantly improve the accuracy and efficiency of endoscopic procedures by providing surgeons with enhanced depth perception and 3D reconstruction capabilities. By addressing the challenges posed by hardware limitations and environmental variations, our deep learning-based framework represents a promising step towards overcoming the limitations of current endoscopic surgical practices. Future research efforts, guided by the identified limitations and potential improvements, can further refine and validate this approach, ultimately leading to its successful integration into clinical practice and improved surgical outcomes for patients.

Chapter 3

SPARSE: Few-Shot Semi-Supervised Learning via Class-Conditioned Image Translation

The contents of this chapter are partly or entirely adapted from our work [47], to which the reader is referred for further details. The chapter has been reformulated to fit the overall narrative of this thesis, with minor editorial changes and additional context connecting the material to the surgical robotics scenario outlined in Chapter 1 and to the companion contribution presented in Chapter 2.

3.1 Introduction

Deep learning has demonstrated remarkable potential in revolutionizing medical imaging [19]. However, insufficient labeled data for model training is one of the main challenges hindering its effectiveness that arises from several constraints such as: the stringent privacy regulations and ethical guidelines governing patient data access and distribution [20], and the necessity of specialized medical expertise for data annotation, a resource limited by healthcare professionals' primary commitment to patient care [12]. These constraints often result in what is known as the low-data regime - a situation where the number of labeled medical images falls below the threshold needed for reliable convergence of deep networks, typically ranging from dozens to a few hundred samples depending on the complexity of the task and model architecture. To address this issue, researchers have explored unsupervised, supervised and semi-supervised learning.

The approaches in the first category tackle the low-data regime through unsupervised learn-

ing, which exploits the abundant unlabeled medical images to learn meaningful representations. Models are trained to capture underlying data patterns through tasks like image reconstruction, anomaly detection, or feature learning. This approach is particularly valuable in medical imaging where unlabeled data is available, but a fundamental challenge remains: without sufficient labeled validation data, it is difficult to ensure that the extracted features are clinically relevant rather than merely statistically significant in the data distribution. In the second case of supervised learning, researchers have explored various strategies to mitigate the limited availability of labeled samples, such as transfer learning, data augmentation and synthetic data generation. Transfer learning leverages models pre-trained on large datasets by fine-tuning them for specific tasks with limited data [59], but its effectiveness diminishes when the target domain differs significantly from the source domain [5]. Data augmentation techniques artificially expand training datasets through various transformations [17], but cannot introduce new information. Synthetic data generation through simulation or generative models has emerged as another strategy [65, 3], though generating fully synthetic datasets that realistically capture real-world data distributions remains challenging [28, 62]. The third approach is semi-supervised learning (SSL) that simultaneously leverages labeled and unlabeled data. Traditional SSL methods relied on machine learning techniques such as self-training and co-training. However, recent advances in generative models, particularly Generative Adversarial Networks (GANs) based methods, have improved SSL performance, not by generating completely synthetic data but by learning to extract meaningful features from the real unlabeled data distribution to enhance the learning process [61]. Despite these advances, existing SSL methods often struggle in low-data regimes as it happens in medical imaging. In this chapter we tackle this issue and we introduce the following contributions:

- A novel GAN-based semi-supervised learning method specifically designed for medical image classification in low labeled-data regimes.
- A dynamic training schedule that alternates between supervised phases and unsupervised phases to optimize learning efficiency.
- An image-to-image translation mechanism employed as a secondary task that, unlike purely generative approaches that create images de novo from noise vectors, modifies existing real unlabeled images to preserve authentic anatomical features while enriching feature representations beyond what traditional generative approaches provide.
- A confidence-weighted temporal ensemble technique that combines predictions from multiple model components and previous training iterations, significantly improving

pseudo-labeling reliability in low-data scenarios.

- A comprehensive empirical evaluation demonstrating competitive performance against six state-of-the-art SSL methods across eleven benchmark datasets for medical image classification tasks.

The remainder of this chapter is organized as follows: section 3.2 reviews the related works in the field, providing context for our research contributions. Section 3.3 details our proposed method. Section 3.4 describes the experimental configuration, including datasets, parameters, and evaluation metrics. Section 3.5 presents our results and provides comprehensive analysis. Finally, section 3.6 concludes the chapter with a summary of our findings and suggestions for future research directions.

3.2 Related Works

Semi-supervised learning methods aim to leverage both labeled and unlabeled data to improve model performance, particularly in scenarios where labeled data is scarce or expensive to obtain. A recent survey [**?**] has established a clear taxonomy of SSL approaches, distinguishing between two main classes: inductive methods, which construct classifiers that can generate predictions for any input, and transductive methods, which optimize directly over predictions for a given set of unlabeled data points. Inductive methods can be further subdivided into three categories based on how they incorporate unlabeled data: (1) wrapper methods, which iteratively train classifiers on labeled data and use their predictions to generate pseudo-labels for unlabeled samples; (2) unsupervised preprocessing methods, which extract features or determine initial parameters from unlabeled data before supervised training; and (3) intrinsically semi-supervised methods, which directly incorporate unlabeled data into the objective function or optimization procedure. Since our proposed approach falls within this third category, and it exploits GANs, the remainder of this section reviews the GAN-based SSL methods across various domains, focusing on approaches that have introduced key architectural innovations, while the interested readers can refer to [61] for a comprehensive review of approaches within this category. Semi-supervised learning is built on the fundamental assumption that the data distribution in the input space contains substantial information about label distribution in the output space. Within this context, GANs are particularly suitable candidates for SSL applications, given their inherent ability to model underlying data distributions and reveal patterns in the input space. The first significant work in this context introduced the SGAN model [50], which expands the traditional GAN architecture by augmenting the discriminator to perform dual functions that distinguish

between real and synthetic samples while simultaneously predicting class labels for input data. This dual-purpose approach represented an important extension of the framework through pseudo-labeling, where the discriminator/classifier is trained on both labeled data and generated samples with known class labels. SGAN exemplifies what is known as a two-player model in GAN-based SSL, where the traditional generator-discriminator architecture is maintained but the discriminator is extended to perform both adversarial discrimination and classification tasks simultaneously. Building on these foundations, MatchGAN [75] introduced an innovative approach that leveraged the Wasserstein distance and conditional generation. As a semi-supervised conditional GAN, MatchGAN utilizes the label space in the target domain along with unlabeled samples to generate additional labeled training data. The framework assigns labels from the pool of labeled samples to unlabeled samples, then passes these through the generator to create synthetic versions of images based on the target labels. This work also introduces a match loss term that compares the generated images to the original labeled images from which the target labels are sampled. A breakthrough in GAN-based SSL came with the introduction of TripleGAN [36], which addressed the difficulty of simultaneously optimizing both generator and discriminator performance. TripleGAN pioneered the three-player model architecture by incorporating an additional classifier that works independently from the discriminator, creating a tripartite interaction between generator, discriminator, and classifier. In this three-player setup, the classifier works in conjunction with the generator to characterize conditional distributions between images while limiting the discriminator's role to identifying fake image-label pairs. This separation of concerns allows each component to specialize in its primary task, potentially leading to improved overall performance compared to two-player models where the discriminator must balance competing objectives. Recent developments have significantly expanded upon the TripleGAN framework [32, 94, 85]. EC-GAN [32] proposed a mechanism where generated images are immediately processed by a classifier to produce pseudo-labels. This classifier-generator interaction is regulated through a hyperparameter-weighted loss function that precisely controls the influence of generated samples on classifier training. SEC-CGAN [94] introduced a co-supervised learning paradigm where a conditional GAN is trained alongside the classifier, providing semantics-conditioned, confidence-aware synthesized examples during training. CISSL-GAN [85] then extended the Triple-GAN framework to address semi-supervised learning with class-imbalanced data through a dynamic class-rebalancing sampler that strategically selects pseudo-labeled samples from unlabeled data. Beyond GAN-based approaches, recent SSL methods have explored diverse strategies including multi-task unified learning [88], masked image modeling [87], differential feature learning [89], and loss-based pseudo-label selection [90], demonstrating the value of learning from multiple complemen-

tary signals—principles that align with our ensemble-based approach. The analysis of the literature reported so far shows that several open issues still exist in current GAN-based SSL approaches:

- None of the existing methods specifically addresses the challenges of extremely low labeled data regimes, such as the case where only 5-10 labeled samples per class are available.
- Current approaches like SGAN [50], TripleGAN [36], and EC-GAN [32] primarily rely on generation-based paradigms for unsupervised learning, lacking effective mechanisms to integrate supervised and unsupervised signals.
- Existing methods rely on a single discriminator or on a single classifier, while none has investigated possible advantages given by the use of an ensemble of models that, in other domains, has proven to provide complementary outputs that enhance model robustness [14, 30].

The next section introduces our methodology that addresses these limitations.

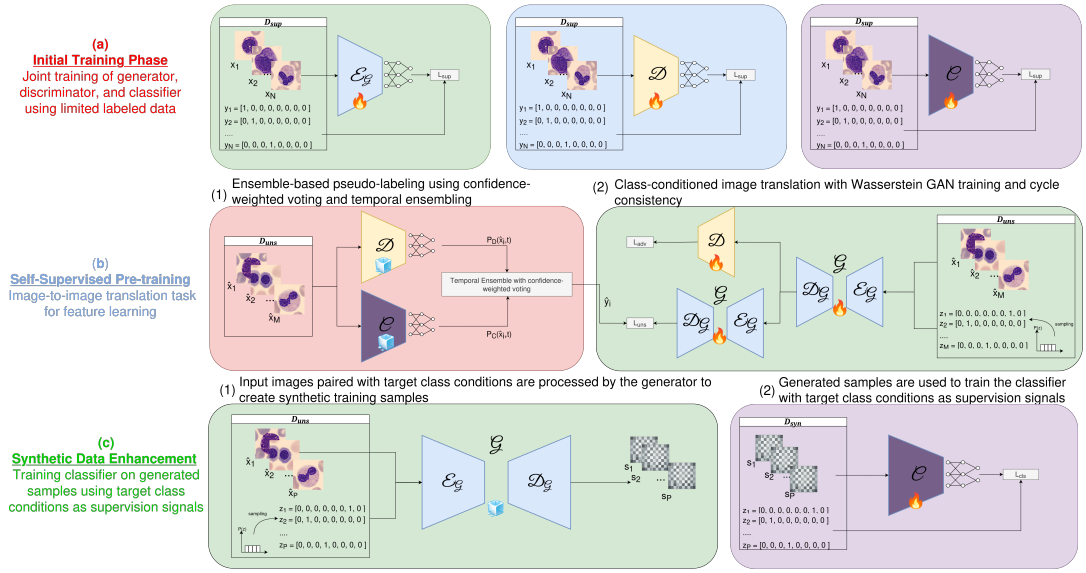

Figure 3.1: Three-phase framework for semi-supervised learning with limited labeled data. Our approach integrates three specialized networks: a generator ($\mathscr{G}$) for class-conditioned image synthesis, a discriminator ($\mathscr{D}$) for authenticity assessment and classification signaling, and a dedicated classifier ($\mathscr{C}$) for the primary classification task. The generator comprises an encoder ($\mathscr{E}_{\mathscr{G}}$) and decoder ($\mathscr{D}_{\mathscr{G}}$) for image-to-image translation.(a) Initial Training Phase: Joint training of these three networks using limited labeled data. (b) Self-Supervised Pre-training: Two-part approach combining (1) ensemble-based pseudo-labeling using confidence-weighted voting and temporal ensembling, and (2) class-conditioned image translation with Wasserstein GAN training and cycle consistency. (c) Synthetic Data Enhancement Phase: Training classifier on generated samples using target class conditions as supervision signals, where (1) input images paired with target class conditions are processed by the generator to create synthetic training samples, and (2) generated samples are used to train the classifier with target class conditions as supervision signals. Fire symbols indicate trainable networks while ice symbols represent frozen weights during respective training phases.

3.3 Methods

We propose a novel semi-supervised learning strategy called SPARSE (**S**emi-supervised **P**seudo-labeling via **A**dversarial **R**epresentation tran**S**lation **E**nhancement), designed to achieve robust classification performance in extremely low labeled-data regimes. Our approach integrates three specialized neural networks: a generator ($\mathscr{G}$) that performs class-conditioned image translation, a discriminator ($\mathscr{D}$) that assesses image authenticity while providing classification signals, and a dedicated classifier ($\mathscr{C}$) that focuses exclusively on the classification task. Our approach consists of three main phases:

1. Supervised training phase (Figure 3.1a): it jointly trains the three aforementioned

networks, $\mathscr{G}, \mathscr{D}$ and $\mathscr{C}$. Each model is trained with a supervised dataset $D_{sup} = \{(x_i, y_i)\}_{i=1}^{N}$, where N is the number of few-shot samples, x_i is the input sample - an image in our experiments, and y_i is the corresponding one-hot encoded ground truth vector, with $y_i \in \{0,1\}^{K\times 1}$ where K is the total number of classes. This initial phase is crucial for maintaining classification accuracy and preventing drift in the unsupervised learning process, represented in panel b of the same figure, by providing supervised signals from the limited labeled data.

2. Self-supervised pre-training phase (Figure 3.1b), which consists of two components. The first is an ensemble-based pseudo-labeling block that combines the confidence-weighted scores provided by $\mathscr{D}$ and $\mathscr{C}$ given a set $D_{uns} = \{\hat{x}_i\}_{i=1}^{M}$ of M unsupervised samples, with $M >> N$. This ensemble outputs the pseudo-labels $\{\hat{y}_i\}_{i=1}^{M}$, with $\hat{y}_i \in \{0,1\}^{K\times 1}$, assigned to each sample in D_{uns}. The second component of the self-supervised pretraining phase introduces a class-conditioned image translation task that uses randomly sampled class conditions $\{z_i\}_{i=1}^{M}$, with $z_i \in \{0,1\}^{K\times 1}$. Hence, this phase leverages the abundant unlabeled data to improve feature representations and model generalization capabilities through image-to-image translation tasks.

3. Synthetic data enhancement phase (Figure 3.1c): it receives as input D_{uns} paired with one-hot encoded class vectors $\{z_i\}_{i=1}^{P}$ randomly sampled from a uniform distribution, which are processed by the generator $\mathscr{G}$ to create synthetic training samples $D_{syn} = \{s_i\}_{i=1}^{P}$, with $P >> N$. Subsequently, D_{syn} is used to train the classifier $\mathscr{C}$ with these same class vectors $\{\hat{z}_i\}_{i=1}^{P}$ serving as supervision signals. This final phase aims to expand the effective training set by creating synthetic samples that augment the limited labeled data.

The training schedule alternates between two phases: the supervised phase (executed at every epoch) and the combined self-supervised and synthetic data enhancement phase (executed every μ epochs, where μ is a hyperparameter), ensuring stable and effective utilization of the entire dataset.

It is worth noting that our approach addresses two primary issues in semi-supervised learning in an extremely low labeled-data regime. The first is the availability of insufficient labeled samples for effective supervised learning, which may affect panel (a) of Figure 3.1: the self-supervised pretraining in panel (b) synthesizes new samples that are used in panel (c) to train the classifier with a large amount of samples. The second issue concerns the integration of supervised and unsupervised learning signals. While the paradigm in the literature [50, 36, 32] defines pre-tasks where a generative model conditionally generates new

samples that enhance the downstream classification task, our approach integrates an image-to-image translation pre-task that enriches the model with more semantic information than a standard generative step (Figure 3.1b).

The rest of this section details such three phases: next subsection 3.3.1 presents panel (a) of Figure 3.1, whilst subsection 3.3.2 describes both panels (b) and (c) of the same f

3.3.1 Supervised Training Phase

The supervised phase in (Figure 3.1a) is crucial for maintaining classification accuracy and preventing drift in the unsupervised learning process. It occurs at every epoch, utilizing the limited labeled data to train simultaneously the encoder of the generator $E_{\mathscr{G}}$, $\mathscr{D}$ and $\mathscr{C}$ to perform classification. We leverage deep supervision by adding a classification tail to the bottleneck of the generator's encoder $E_{\mathscr{G}}$ denoted as a set of interconnected neurons in panel (a) of Figure 3.1; it is located at the bottleneck because it serves as the information compression point between the encoder and decoder paths. By applying classification supervision at this critical juncture, we ensure that the most compact representation in the network encodes both structural information needed for generation and semantic information required for classification.

The supervised loss function used by all the three networks ($\mathcal{L}_{sup}$) combines four specialized loss terms, each addressing a specific challenge in few-shot learning and weighted by coefficients to balance their contribution.

$$\mathcal{L}_{sup} = \mathcal{L}_{prototype} + \alpha\mathcal{L}_{mutual} + \beta\mathcal{L}_{entropy} + \gamma\mathcal{L}_{mixup} \tag{3.1}$$

The prototype loss ($\mathcal{L}_{prototype}$) helps create discriminative class-specific features by learning robust prototypical representations for each class. The mutual learning loss ($\mathcal{L}_{mutual}$) enables knowledge sharing between the three models, leveraging their complementary perspectives on the data. The entropy minimization loss ($\mathcal{L}_{entropy}$) encourages the models to make confident predictions, helping combat the uncertainty inherent in limited-data scenarios. Finally, the mixup loss ($\mathcal{L}_{mixup}$) provides regularization through data augmentation, helping prevent overfitting which is particularly crucial when training with few samples. The rest of this section details the computation of each of these four loss functions.

The prototype loss $\mathcal{L}_{prototype}$ creates discriminative class-specific representations by com-

paring softmax probabilities with class prototypes:

$$\mathcal{L}_{prototype} = -\sum_{i=1}^{N} \log \frac{\exp(-d(p_T(x_i), c_{y_i}))}{\sum_{k=1}^{K} \exp(-d(p_T(x_i), c_k))} \tag{3.2}$$

where $p_T(x_i)$ represents the softmax probabilities of input x_i with temperature T, c_k is the prototype of class k computed as the mean of the class probabilities:

$$c_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} p_T(x_i) \tag{3.3}$$

where S_k is the set of samples from class k. The distance function $d(,)$ is defined as the negative sum of squared differences:

$$d(p_T(x), c_k) = -\sum_{j=1}^{K} (p_{T,j}(x) - c_{k,j})^2 \tag{3.4}$$

where K is the total number of classes.

The mutual learning loss $\mathcal{L}_{mutual}$ facilitates knowledge transfer between models through a combination of supervised cross-entropy and KL divergence:

$$\mathcal{L}_{mutual} = \mathcal{L}_{ce} + \lambda_{kl}\mathcal{L}_{kl} \tag{3.5}$$

where λ_{kl} is the weight coefficient for the KL divergence term, $\mathcal{L}_{ce}$ is the sum of cross-entropy losses for each model:

$$\mathcal{L}_{ce} = \sum_{m \in \{E_{\mathscr{G}}, \mathscr{D}, \mathscr{C}\}} \mathrm{CE}(p_m(x), y) \tag{3.6}$$

and $\mathcal{L}_{kl}$ is the symmetric KL divergence between each model's probabilities and the average of other models' probabilities:

$$\mathcal{L}_{kl} = \sum_{m \in \{E_{\mathscr{G}}, \mathscr{D}, \mathscr{C}\}} \mathrm{KL}(p_m(x) | \frac{1}{2} \sum_{n \neq m} p_n(x)) \tag{3.7}$$

The entropy minimization loss $\mathcal{L}_{entropy}$ promotes confident predictions:

$$\mathcal{L}_{entropy} = -\sum_{k=1}^{K} p_k(x) \log(p_k(x)) \tag{3.8}$$

Finally, the mixup loss $\mathcal{L}_{mixup}$ provides regularization by training on interpolated samples and labels. For each pair of samples (x_i, y_i) and (x_j, y_j), the interpolation weight λ_{mix} is sampled from a Beta distribution:

$$\lambda_{mix} \sim \text{Beta}(\alpha_{mix}, \alpha_{mix}) \tag{3.9}$$

The Beta distribution is particularly suitable for generating interpolation weights as it is bounded between [0,1] and can be symmetric around 0.5, ensuring a balanced mixing of samples while maintaining their relative contributions. The hyperparameter α_{mix} in the Beta distribution controls the strength of interpolation - higher values of α_{mix} lead to interpolation weights closer to 0.5, while lower values favor weights closer to 0 or 1. This weight is then used to create interpolated samples and labels:

$$\tilde{x} = \lambda_{mix} x_i + (1 - \lambda_{mix}) x_j \tag{3.10}$$

$$\tilde{y} = \lambda_{mix} y_i + (1 - \lambda_{mix}) y_j \tag{3.11}$$

The mixup loss is then computed as:

$$\mathcal{L}_{mixup} = \mathcal{L}_{ce}(p(\tilde{x}), \tilde{y}) \tag{3.12}$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss.

3.3.2 Unsupervised Training Phase

The unsupervised training phase, illustrated in panel (b) of Figure 3.1, executes every μ epochs. This phase leverages unlabeled data through an image-to-image translation framework with three distinct stages.

First, we use an ensemble-based pseudo-labeling mechanism to estimate class probabilities for unlabeled samples (subsection 3.3.2). These probability estimates then guide a class-conditioned image translation process, which learns to generate class-specific variations of input images (subsection 3.3.2). Finally, we employ these generated samples in a synthetic data enhancement phase. Here, the synthetic images serve as additional training data to strengthen the classifier's ability to distinguish between classes (subsection 3.3.2).

Ensemble-based Pseudo-labeling

To effectively utilize unlabeled samples, we require reliable class probability estimates. Our approach addresses three key challenges: (1) quantifying model uncertainty, (2) aggregating predictions from multiple models, and (3) maintaining temporal stability throughout training. We tackle these challenges through an ensemble mechanism (Figure 3.1b) that integrates confidence-weighted voting, temporal ensembling, and adaptive thresholding.

At epoch t, each unlabeled image $\hat{x}_i$ from $D_{uns} = \{\hat{x}_i\}_{i=1}^{M}$ is processed by the models trained during the previous initial training phase. We apply temperature scaling to obtain calibrated class probabilities:

$$p_m(\hat{x}_i, t) = \text{softmax}\left(\frac{l_m(\hat{x}_i, t)}{T}\right) \tag{3.13}$$

where $l_m(\hat{x}_i, t)$ represents the **logits** (raw pre-softmax outputs) from model $m \in \{\mathscr{D}, \mathscr{C}\}$ for input $\hat{x}_i$ at epoch t. The temperature parameter T controls prediction sharpness: lower values ($T < 1$) yield more confident predictions, while higher values ($T > 1$) produce smoother probability distributions.

Each model's prediction reliability is quantified using an entropy-based confidence measure:

$$c_m(\hat{x}_i, t) = 1 - \frac{H(p_m(\hat{x}_i, t))}{H_{max}} \tag{3.14}$$

where the entropy $H(p_m(\hat{x}_i, t))$ is computed as:

$$H(p_m(\hat{x}_i, t)) = -\sum_{k=1}^{K} p_m^{(k)}(\hat{x}_i, t) \log p_m^{(k)}(\hat{x}_i, t) \tag{3.15}$$

and $H_{max} = \log K$ represents the maximum possible entropy for K classes. The term $p_m^{(k)}(\hat{x}_i, t)$ denotes model m's predicted probability for class k. This confidence score ranges from 0 (complete uncertainty) to 1 (complete certainty). Together, the temperature-scaled probabilities and this entropy-based confidence measure make each model's predictions uncertainty-aware, following the rationale of modern uncertainty-quantification techniques, in which confidence estimates and their aggregation across models are used to obtain more trustworthy predictions under limited supervision.

We combine predictions from both models by weighting each according to its confidence score:

$$p_{weighted}(\hat{x}_i, t) = \frac{\sum_{m \in \{\mathscr{D}, \mathscr{C}\}} c_m(\hat{x}_i, t) \cdot p_m(\hat{x}_i, t)}{\sum_{m \in \{\mathscr{D}, \mathscr{C}\}} c_m(\hat{x}_i, t)} \tag{3.16}$$

The denominator ensures proper normalization. This weighting scheme allows more confident

models to contribute more strongly to the final prediction.

To enhance prediction stability, we incorporate historical information by blending current predictions with past predictions:

$$p_{ens}(\hat{x}_i, t) = \alpha \cdot p_{weighted}(\hat{x}_i, t) + (1 - \alpha) \cdot p_{ema}(\hat{x}_i, t - 1) \tag{3.17}$$

Here, $p_{ema}(\hat{x}_i, t-1)$ captures the temporal history through an exponential moving average (EMA) of past ensemble predictions. The parameter $\alpha \in [0, 1]$ balances current information (higher α) against historical stability (lower α). This temporal smoothing prevents abrupt prediction changes that could destabilize the training process.

After computing the current ensemble prediction, we update the EMA:

$$p_{ema}(\hat{x}_i, t) = \beta \cdot p_{ema}(\hat{x}_i, t - 1) + (1 - \beta) \cdot p_{ens}(\hat{x}_i, t) \tag{3.18}$$

The momentum parameter $\beta \in [0, 1]$ determines the temporal memory span. Large values (e.g., $\beta = 0.99$) maintain longer memory for stable predictions, while smaller values allow faster adaptation to recent changes.

With reliable probability estimates in hand, we select only the most confident predictions for pseudo-labeling:

$$\mathcal{S}(t) = \{\hat{x}_i \in D_{uns} : \max_k(p_{ens}^{(k)}(\hat{x}_i, t)) > \tau(t)\} \tag{3.19}$$

where $\max_k(p_{ens}^{(k)}(\hat{x}_i, t))$ is the highest class probability from the ensemble prediction. The subset $\mathcal{S}(t)$ contains selected samples at epoch t, and the adaptive threshold $\tau(t)$ is defined as:

$$\tau(t) = Q_\rho(\{\max_k(p_{ens}^{(k)}(\hat{x}_i, t)) : \hat{x}_i \in D_{uns}\}) \tag{3.20}$$

Here, Q_ρ denotes the ρ-th percentile of maximum probabilities across all unlabeled samples in the current batch, where $\rho \in [0, 1]$ is the percentile threshold parameter.

This percentile-based approach automatically adjusts to the model's current performance level. It selects $(1 - \rho) \times 100\%$ of the most confident samples. For instance, setting $\rho = 0.8$ selects the top 20% most confident predictions. This adaptive mechanism prevents error accumulation from unreliable pseudo-labels while naturally accommodating the model's improving performance.

For each selected confident sample $\hat{x}_i \in \mathcal{S}(t)$, we create a discrete pseudo-label:

$$\hat{y}_i = \text{one-hot}(\arg\max_k p_{ens}^{(k)}(\hat{x}_i, t)) \tag{3.21}$$

This produces a one-hot encoded pseudo-label $\hat{y}_i \in \{0,1\}^{K\times 1}$ for each unlabeled sample $\hat{x}_i$. The complete set of pseudo-labels $\{\hat{y}_i\}_{i=1}^{|\mathcal{S}(t)|}$ for all selected samples then feeds into the subsequent class-conditioned image translation process.

Class-conditioned Image Translation

Using the pseudo-labels $\{\hat{y}_i\}_{i=1}^{M}$ obtained from our ensemble mechanism, we implement a class-conditioned image translation process that leverages $\mathscr{G}$ and $\mathscr{D}$ working in tandem (Figure 3.1b right). Now the generator $\mathscr{G}$, which consists of a complete U-Net architecture, not just the encoder as in the supervised phase, learns to perform class-conditioned image translation while preserving semantic features relevant to classification. Its training objective $\mathcal{L}_{uns}$ combines four components:

$$\mathcal{L}_{uns} = \mathcal{L}_{adv} + \mathcal{L}_{cls}^{\mathscr{D}} + \mathcal{L}_{cls}^{\mathscr{G}} + \lambda_{rec}\mathcal{L}_{rec} \tag{3.22}$$

where λ_{rec} is the weight coefficient for the reconstruction loss, and the adversarial loss $\mathcal{L}_{adv}$ uses the Wasserstein distance metric to assess image realism:

$$\mathcal{L}_{adv} = -\mathbb{E}_{\hat{x}\sim D_{uns}}[\mathscr{D}(\mathscr{G}(\hat{x}, z_{target}))], \tag{3.23}$$

Next, the classification losses $\mathcal{L}_{cls}^{\mathscr{D}}$ and $\mathcal{L}_{cls}^{\mathscr{G}}$ ensure accurate conditioning on target classes using cross-entropy from both the discriminator and generator classifiers:

$$\mathcal{L}_{cls}^{\mathscr{D}} = -\mathbb{E}_{\hat{x}\sim D_{uns}} \sum_{k=1}^{K} z_{target}^{k} \log(p_k^{\mathscr{D}}(\mathscr{G}(\hat{x}, z_{target}))) \tag{3.24}$$

$$\mathcal{L}_{cls}^{\mathscr{G}} = -\mathbb{E}_{\hat{x}\sim D_{uns}} \sum_{k=1}^{K} z_{target}^{k} \log(p_k^{\mathscr{G}}(\mathscr{G}(\hat{x}, z_{target}))) \tag{3.25}$$

where z_{target}^{k} is the k-th element of the target class one-hot encoding, $p_k^{\mathscr{D}}$ represents the probability for class k from the discriminator, and $p_k^{\mathscr{G}}$ represents the probability for class k from the generator's classifier. The reconstruction loss $\mathcal{L}_{rec}$ maintains content consistency using L1 distance:

$$\mathcal{L}_{rec} = \mathbb{E}_{\hat{x}\sim D_{uns}}[|\mathscr{G}(\mathscr{G}(\hat{x}, z_{target}), z_{source}) - \hat{x}|_1] \tag{3.26}$$

where z_{source} is the one-hot encoding of the most probable class according to $p_{ens}(\hat{x}, t)$, connecting this translation process with our previous pseudo-labeling step.

The discriminator $\mathscr{D}$ serves a dual role: it assesses image realism through a Wasserstein

distance metric while also providing classification signals. Its objective function reflects these dual tasks:

$$\mathcal{L}_{\mathscr{D}} = -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{gp}\mathcal{L}_{gp} \tag{3.27}$$

where λ_{cls} and λ_{gp} are weight coefficients for the classification loss and gradient penalty terms, respectively. Its adversarial component implements the Wasserstein distance as follows:

$$-\mathcal{L}_{adv} = -\mathbb{E}_{\hat{x}\sim D_{uns}}[\mathscr{D}(\hat{x})] + \mathbb{E}_{\hat{x}\sim D_{uns}}[\mathscr{D}(\mathscr{G}(\hat{x}, z_{target}))] \tag{3.28}$$

The gradient penalty term $\mathcal{L}_{gp}$ enforces the Lipschitz constraint:

$$\mathcal{L}_{gp} = \mathbb{E}_{\hat{x}_{interp}}[(|\nabla_{\hat{x}_{interp}}\mathscr{D}(\hat{x}_{interp})|_2 - 1)^2] \tag{3.29}$$

where $\hat{x}_{interp}$ is sampled uniformly along straight lines between pairs of real and generated images.

Synthetic Data Enhancement

After establishing the image translation process, we leverage translated images as a form of synthetic labeled data to enhance the classifier's performance (Figure 3.1c). The key process involves taking unlabeled input images $\{\hat{x}_i\}_{i=1}^{M}$ and sampling random target classes $\{z_i\}_{i=1}^{P}$ from a uniform distribution over the K classes in one-hot format, where P is the number of synthetic samples generated. We then use these randomly sampled classes to condition the generator to translate the original images of unknown classes. The resulting translated images have known target classes $\{z_i\}_{i=1}^{P}$.

These translated images with known target classes can then be used to train the classifier in a supervised manner:

$$\mathcal{L}_{\mathscr{C}} = \mathcal{L}_{cls}(\mathscr{C}(\mathscr{G}(\hat{x}, z)), z) \tag{3.30}$$

where z represents the target class condition provided to the generator for synthetic data generation. This approach provides us with effectively labeled training samples, since we know exactly what class condition was used to generate each image. Importantly, as demonstrated in [22], the visual quality of generated images is not critical for our classification objective: instead, we focus on ensuring that the translation process captures and preserves discriminative features that are useful for classification.

3.3.3 Inference configuration

Our approach in the inference phase can be deployed in different configurations that exploit the model trained as reported in section 3.3. In particular, in the rest of this chapter we consider the following two setups for inference:

- SPARSE: it uses only the classifier $\mathscr{C}$.
- SPARSE$_{\text{ens}}$: it exploits both the discriminator $\mathscr{D}$ and the classifier $\mathscr{C}$, which are combined in late fusion by averaging the estimates of posterior probabilities per class.

It is worth noting that in our ensemble configuration, we should also consider the potential use of the generator's encoder $\mathscr{E}_{\mathscr{G}}$. However, since it is trained to minimize a loss function that balances both generation and classification objectives, this results in a performance degradation - a finding we experimentally verified, though omitted here for conciseness.

3.4 Experimental Configuration

This section describes our experimental methodology: it details the materials (subsection 3.4.1), followed by our training configuration (subsection 3.4.2).

3.4.1 Materials

We conducted experiments using eleven datasets from MedMNIST repository [86]: BloodMNIST, BreastMNIST, ChestMNIST, DermaMNIST, OCTMNIST, OrganAMNIST, OrganCMNIST, OrganSMNIST, PathMNIST, PneumoniaMNIST and TissueMNIST. As shown in Table 3.1, these datasets span different medical imaging modalities and classification tasks, with varying number of classes (2-11) and dataset sizes (from hundreds to hundreds of thousands of samples).

To evaluate our method's effectiveness in extremely low labeled-data regimes, we conducted experiments across four few-shot settings: 5-shot, 10-shot, 20-shot, and 50-shot per class, with 5-shot representing the most challenging scenario. For each N-shot setting, we constructed the training set using N labeled samples per class, with the remaining samples treated as unlabeled data. For data preprocessing, we applied a transformation pipeline consisting of random horizontal flipping for data augmentation and tensor conversion, with input images maintaining their original 128×128 pixel resolution. To ensure reproducibility, we utilized the original validation/test splits provided by the MedMNIST authors.

Table 3.1: Characteristics of MedMNIST datasets used in our experiments.

Dataset	Modality	Task Type	# Samples		
			Total	Training	Val/Test
BloodMNIST	Blood Cell Microscope	Multi-Class (8)	17,092	11,959	1,712/3,421
BreastMNIST	Breast Ultrasound	Binary-Class (2)	780	546	78/156
ChestMNIST	Chest X-Ray	Binary-Class (2)	112,120	78,468	11,219/22,433
DermaMNIST	Dermatoscope	Multi-Class (7)	10,015	7,007	1,003/2,005
OCTMNIST	Retinal OCT	Multi-Class (4)	109,309	97,477	10,832/1,000
OrganAMNIST	Abdominal CT	Multi-Class (11)	58,830	34,561	6,491/17,778
OrganCMNIST	Abdominal CT	Multi-Class (11)	23,583	12,975	2,392/8,216
OrganSMNIST	Abdominal CT	Multi-Class (11)	25,211	13,932	2,452/8,827
PathMNIST	Colon Pathology	Multi-Class (9)	107,180	89,996	10,004/7,180
PneumoniaMNIST	Chest X-Ray	Binary-Class (2)	5,856	4,708	524/624
TissueMNIST	Kidney Cortex Microscope	Multi-Class (8)	236,386	165,466	23,640/47,280

3.4.2 Training Configuration

The training schedule alternates between supervised and unsupervised learning phases. The supervised phase occurs at every epoch, while the unsupervised phase is executed every μ epochs as already described in section 3.3. All models were trained for 1000 epochs using the AdamW optimizer, maintaining identical configurations throughout all experiments which are reported in table 3.2. During training, we implemented a model checkpoint strategy that saved the model state achieving the best validation accuracy, which was then used for final evaluation. For all the models, we did not investigate any hyperparameter configuration since their tuning is out of the scope of this chapter. Nevertheless, because the 'No Free Lunch' Theorem for optimization [84] states that no universal set of hyperparameters will optimize a model's performance across all possible datasets, this approach ensures a fair comparison among all the approaches.

3.5 Results

This section presents the experimental results evaluating our framework's effectiveness in extremely low labeled data regimes in medical image classification. We compare our approach against the six state-of-the-art semi-supervised GAN architectures which are already presented in section 3.2, using their original implementations without additional optimization or modifications to ensure fair comparison with our method. To assess model performance, we employ classification accuracy per class as our primary evaluation metric. For each dataset, we first compute the accuracy across all available classes and then the final score

Table 3.2: Training hyperparameters used in our experiments

Parameter	**Value**
General Training Parameters	
Training Epochs	1000
Base Learning Rate	0.0002
Optimizer	AdamW
μ (unsupervised phase frequency)	10
T (temperature parameter)	2.0
Loss Weights for Supervised Objective	
α (mutual learning loss weight)	0.1
β (entropy minimization loss weight)	0.01
γ (mixup loss weight)	0.5
λ_{kl} (KL divergence weight)	0.5
mixup parameters	
α_{mix} (beta distribution parameter)	0.2
loss weights for generator	
λ_{rec} (reconstruction loss weight)	10.0
loss weights for discriminator	
λ_{cls} (classification loss weight)	1.0
λ_{gp} (gradient penalty weight)	10.0
Ensemble Parameters	
α (temporal ensemble weight)	0.6
β (EMA momentum)	0.99
ρ (percentile threshold)	0.75

is obtained by averaging these accuracies across all datasets, providing a robust evaluation of the model's generalization capabilities across varying medical imaging tasks, modalities, and data distributions. Individual per-dataset results are reported in Appendix A.

Table 3.3 presents the performance of all models across different few-shot settings, where we systematically vary the number of samples per class from extremely limited (5-shot) to more moderate (50-shot) situations. The first two rows show the results of our approach using only the classifier $\mathscr{C}$ (SPARSE) or using the ensemble between the discriminator $\mathscr{D}$ and $\mathscr{C}$ ($\text{SPARSE}_{\text{ens}}$). The following six rows display the results of the six competitors. We note that the two variants of our approach consistently outperform existing competitors across all settings, and the ensemble setting ($\text{SPARSE}_{\text{ens}}$) achieves the best performance in all scenarios. As the number of labeled samples increases from 5 to 50 shots per class, we observe a steady improvement in performance across all models, though the relative advantage of our approach remains consistent.

To further investigate these findings, the rest of this section examines our improvements in the 5-shot setting (section 3.5.1), which represents the most challenging scenario where

Table 3.3: Model Performance Across Different Few-Shot Settings

Model	Average Accuracy per class			
	5-shot	10-shot	20-shot	50-shot
SPARSE	63.21	68.50	73.44	77.15
$\text{SPARSE}_{\text{ens}}$	**66.22**	**70.95**	**75.71**	**78.28**
SGAN [50]	25.80	25.96	24.92	27.28
MatchGAN [75]	39.88	48.73	51.90	54.15
EC-GAN [32]	35.09	34.66	34.29	47.41
TripleGAN [36]	64.23	68.79	71.40	76.25
SEC-GAN [94]	58.73	63.79	66.95	73.30
CISSL [85]	45.84	46.80	49.19	51.20

the extreme scarcity of labeled data tests the true effectiveness of semi-supervised learning. We then complement the results with a detailed analysis in the 50-shot setting (section 3.5.2), which helps us understand how our method scales when more labeled data becomes available. Statistical significance testing of the performance differences is presented within these analyses. Finally, section 3.5.2 investigates how the frequency of unsupervised training affects model performance.

3.5.1 Performance in 5-shot

We now deepen our analysis on the most challenging 5-shot learning scenario, where only 5 labeled samples per class are available for training while the remaining samples are treated as unlabeled. To provide robust statistical evaluation, we employ the Wilcoxon signed-rank test, a non-parametric method that compares paired accuracy values across our 11 datasets, with Benjamini-Hochberg FDR correction for multiple comparisons. Table 3.4 presents the results of this statistical comparison, with the upper triangular part showing corrected p-values with effect sizes (r) and their interpretations, where significant results ($p < 0.05$) are highlighted in bold, and the lower triangular part displaying Win-Tie-Loss (W-T-L) statistics from the row model's perspective when compared against the column model.

The primary finding is that our ensemble model, $\text{SPARSE}_{\text{ens}}$, achieves statistically significant superiority over most competing approaches, also with a number of wins larger than losses, demonstrating robust performance across diverse medical imaging datasets.
We now discuss the effectiveness of the ensemble approach by comparing $\text{SPARSE}_{\text{ens}}$ to its base model, SPARSE. The results show that the ensemble model demonstrates a statistically significant improvement ($p = 4.1 \times 10^{-2}, r = 0.818$), securing wins on 10 out of 11 datasets with only a single loss. This marked improvement is achieved through ensemble averag-

Table 3.4: 5-Shot statistical comparison of model performance. The upper triangular part shows p-values from statistical comparison (significant values $p < 0.05$ highlighted in bold), while the lower triangular part shows Win-Tie-Loss (W-T-L) statistics. The r-value represents the effect size (Pearson's correlation coefficient) with interpretations: small ($r < 0.3$), medium ($0.3 \leq r < 0.5$), large ($0.5 \leq r < 0.7$), and very large ($r \geq 0.7$). Best performing model is highlighted in bold.

Model	Statistical Comparison (5 shot)							
	SPARSE	**SPARSE$_{ens}$**	SGAN	MatchGAN	CISSL	SEC-GAN	TripleGAN	ECGAN
SPARSE	-	**4.1e-02** r=0.818 (very large)	**2.1e-02** r=0.818 (very large)	6.0e-02 r=0.636 (very large)	8.7e-02 r=0.455 (large)	2.4e-01 r=0.455 (large)	7.7e-01 r=0.091 (small)	**4.4e-02** r=0.636 (very large)
SPARSE$_{ens}$	(10-0-1)	-	**2.1e-02** r=1.000 (very large)	**4.4e-02** r=0.636 (very large)	**2.2e-02** r=0.818 (very large)	5.3e-02 r=0.636 (very large)	1.5e-01 r=0.455 (large)	**3.5e-02** r=0.800 (very large)
SGAN	(1-0-10)	(0-1-10)	-	**2.2e-02** r=0.636 (very large)	**2.2e-02** r=0.636 (very large)	**2.1e-02** r=0.818 (very large)	**2.1e-02** r=0.818 (very large)	**4.4e-02** r=0.600 (very large)
MatchGAN	(2-0-9)	(2-0-9)	(9-0-2)	-	2.6e-01 r=0.273 (medium)	6.0e-02 r=0.455 (large)	6.0e-02 r=0.455 (large)	1.0e-01 r=0.636 (very large)
CISSL	(3-0-8)	(1-0-10)	(9-0-2)	(7-0-4)	-	**4.1e-02** r=0.636 (very large)	**4.4e-02** r=0.455 (large)	**4.4e-02** r=0.455 (large)
SEC-GAN	(3-0-8)	(2-0-9)	(10-0-1)	(8-0-3)	(9-0-2)	-	8.7e-02 r=0.455 (large)	**2.7e-02** r=0.818 (very large)
TripleGAN	(5-0-6)	(3-0-8)	(10-0-1)	(8-0-3)	(8-0-3)	(8-0-3)	-	**4.4e-02** r=0.455 (large)
ECGAN	(2-0-9)	(1-1-9)	(8-1-2)	(2-0-9)	(3-0-8)	(1-0-10)	(3-0-8)	-

ing at inference time. Although both configurations share an identical training procedure, the classifier $\mathscr{C}$ and discriminator $\mathscr{D}$ networks develop complementary decision boundaries due to their architectural differences. Averaging their predictions effectively reduces prediction variance, a critical advantage in the extreme 5-shot regime. Let us now focus on how $\text{SPARSE}_{\text{ens}}$ performs with respect to the external competitors. The analysis shows that $\text{SPARSE}_{\text{ens}}$ obtains statistically significant improvements over a range of generative models, including SGAN ($p = 2.1 \times 10^{-2}, r = 1.000$), MatchGAN ($p = 4.4 \times 10^{-2}, r = 0.636$), CISSL ($p = 2.2 \times 10^{-2}, r = 0.818$) and EC-GAN ($p = 3.5 \times 10^{-2}, r = 0.800$). The largest performance gap is observed against SGAN, where our method was superior across all 11 datasets. The closest competition came from TripleGAN and SEC-GAN; while the differences did not reach statistical significance ($p = 1.5 \times 10^{-1}$ and $p = 5.3 \times 10^{-2}$, respectively), $\text{SPARSE}_{\text{ens}}$ still outperformed them on the majority of datasets with Win-Tie-Loss equal to 8-0-3 and 9-0-2. The consistent outperformance of $\text{SPARSE}_{\text{ens}}$ over this diverse set of competitors stems from a crucial methodological distinction. All competing methods are purely generative, creating images de novo from a noise vector. In contrast, our framework employs image-to-image translation, which modifies existing real, unlabeled images. This strategy preserves the authentic and complex anatomical features of the medical data, providing a more robust training signal. This fundamental advantage is further amplified when compared to two-player models like SGAN and MatchGAN, which suffer from an internal optimization conflict by tasking a single network with both discrimination and classification. Our three-player design avoids this issue. Even when compared to advanced three-player models like TripleGAN that also separate these tasks, our method's reliance on translating real images, rather than generating them from noise, appears to be the decisive factor for success in this data-scarce context.

3.5.2 Performance in 50-shot

We extend our analysis to the 50-shot setting to examine how model performance scales with increased labeled data availability, while still remaining within the low-labeled data regime. This configuration provides ten times more labeled samples per class compared to the 5-shot scenario, allowing us to assess the scaling properties of our methods. Table 3.5 presents the statistical comparison results using the Wilcoxon signed-rank test with Benjamini-Hochberg FDR correction. The table follows the same format as the 5-shot analysis, with p-values and effect sizes in the upper triangular portion and Win-Tie-Loss statistics in the lower triangular portion. The primary finding reveals that $\text{SPARSE}_{\text{ens}}$ maintains its statistical superiority over most competing approaches even as more labeled data becomes available. When com-

Table 3.5: 50-Shot statistical comparison of model performance. The upper triangular part shows p-values from statistical comparison (significant values $p < 0.05$ highlighted in bold), while the lower triangular part shows Win-Tie-Loss (W-T-L) statistics. The r-value represents the effect size (Pearson's correlation coefficient) with interpretations: small ($r < 0.3$), medium ($0.3 \leq r < 0.5$), large ($0.5 \leq r < 0.7$), and very large ($r \geq 0.7$). Best performing model is highlighted in bold.

Model	Statistical Comparison (50 shot)							
	SPARSE	**SPARSE$_{ens}$**	SGAN	MatchGAN	CISSL	SEC-GAN	TripleGAN	ECGAN
SPARSE	-	1.0e-01 r=0.636 (very large)	**2.0e-03** r=1.000 (very large)	**2.0e-03** r=1.000 (very large)	**5.0e-03** r=0.818 (very large)	**1.5e-02** r=0.636 (very large)	3.5e-01 r=0.455 (large)	**4.0e-03** r=0.818 (very large)
SPARSE$_{ens}$	(9-0-2)	-	**2.0e-03** r=1.000 (very large)	**2.0e-03** r=1.000 (very large)	**2.0e-03** r=1.000 (very large)	**2.0e-03** r=1.000 (very large)	1.0e-01 r=0.636 (very large)	**2.0e-03** r=1.000 (very large)
SGAN	(0-0-11)	(0-0-11)	-	**8.0e-03** r=0.818 (very large)	**2.0e-03** r=1.000 (very large)	**2.0e-03** r=1.000 (very large)	**2.0e-03** r=1.000 (very large)	**1.1e-02** r=0.818 (very large)
MatchGAN	(0-0-11)	(0-0-11)	(10-0-1)	-	9.0e-01 r=0.091 (small)	**4.0e-03** r=0.818 (very large)	**2.0e-03** r=1.000 (very large)	1.7e-01 r=0.455 (large)
CISSL	(1-0-10)	(0-0-11)	(11-0-0)	(6-0-5)	-	**1.9e-02** r=0.455 (large)	**8.0e-03** r=0.636 (very large)	5.5e-01 r=0.091 (small)
SEC-GAN	(2-0-9)	(0-0-11)	(11-0-0)	(10-0-1)	(8-0-3)	-	**3.1e-02** r=0.818 (very large)	**4.0e-03** r=0.818 (very large)
TripleGAN	(3-0-8)	(2-0-9)	(11-0-0)	(11-0-0)	(9-0-2)	(10-0-1)	-	**2.0e-03** r=1.000 (very large)
ECGAN	(1-0-10)	(0-0-11)	(10-0-1)	(3-0-8)	(5-0-6)	(1-0-10)	(0-0-11)	-

paring the two variants of our approach, $\text{SPARSE}_{\text{ens}}$ achieves a Win-Tie-Loss equal to 9-0-2 against the base SPARSE model. However, the p-value of 1.0×10^{-1} ($r = 0.636$) indicates no statistical significance: this narrowing gap can be attributed to the changing role of ensemble averaging in different data regimes. With only 5 shots per class, individual model predictions are inherently less reliable and more prone to overfitting, making the ensemble approach particularly valuable for reducing prediction variance. The discriminator and classifier develop highly complementary decision boundaries due to the scarcity of supervised signals. However, with 50 labeled samples per class, the classifier receives sufficient supervision to develop more stable and generalizable representations independently, reducing its reliance on the discriminator's complementary perspective. Examining $\text{SPARSE}_{\text{ens}}$'s performance against external competitors we note that it achieves statistically significant improvements over SGAN ($p = 2.0 \times 10^{-3}, r = 1.000$), MatchGAN ($p = 2.0 \times 10^{-3}, r = 1.000$), CISSL ($p = 2.0 \times 10^{-3}, r = 1.000$), and EC-GAN ($p = 2.0 \times 10^{-3}, r = 1.000$), with Win-Tie-Loss equal to 11-0-0 against each. These results represent stronger statistical evidence compared to the 5-shot setting, where p-values ranged from 2.1×10^{-2} to 4.4×10^{-2}. Furthermore $\text{SPARSE}_{\text{ens}}$ achieves statistical significance against SEC-GAN ($p = 2.0 \times 10^{-3}, r = 1.000$, W-T-L: 0-0-11), similarly to what happens in the 5-shot setting. When comparing against TripleGAN, we notice that the Win-Tie-Loss ratio is 2-0-9 in favor of our approach, similar to the 5-shot settings, but now the performance differences are not statistically significant at $p = 0.1$. These results suggest that performance differences become more consistent across datasets as labeled data increases. The patterns observed in this comparison can be attributed to differences in how methods utilize additional supervision. Our image-to-image translation approach leverages the increased labeled data to learn more accurate class-conditional transformations. With 50 labeled samples per class, the generator receives stronger supervision signals, enabling better preservation of discriminative features during translation. Additionally, the quality of pseudo-labels generated in the unsupervised phase improves due to more reliable initial classifiers.

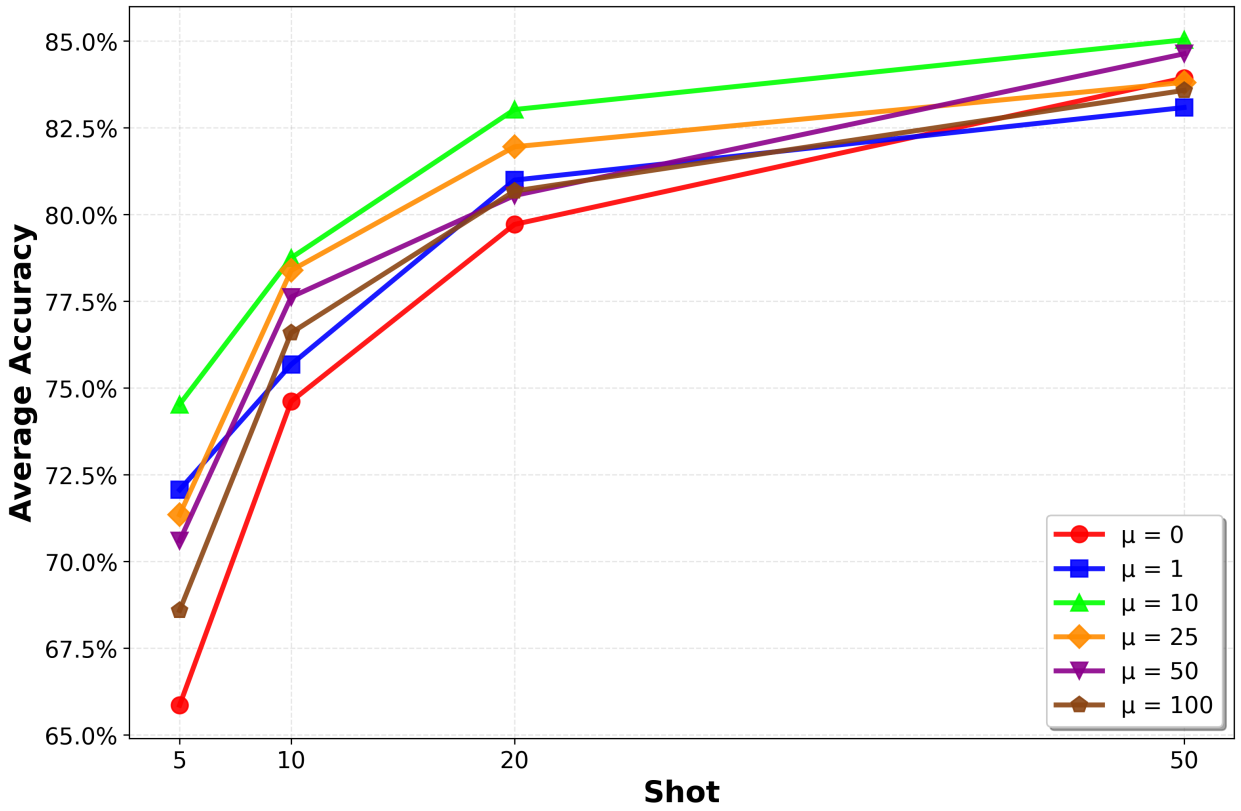

Figure 3.2: Average classification accuracy as a function of the number of labeled samples per class (shots) for different unsupervised training frequencies (μ). Results are averaged across all eleven MedMNIST datasets and computed on the validation set. The parameter μ controls how frequently the unsupervised training phase is executed, with $\mu = 0$ representing supervised learning only (no unsupervised phase), $\mu = 1$ indicating unsupervised training at every epoch, and higher values ($\mu \in \{10, 25, 50, 100\}$) representing unsupervised training every μ epochs.

3.5.3 Framework Component Analysis

To understand the source of our framework's performance gains, we conduct a systematic analysis of key design decisions and architectural components. This section examines two critical aspects: first, we analyze how the frequency of unsupervised training phases affects model performance, demonstrating the importance of properly balancing supervised and unsupervised learning signals. Second, we perform controlled ablation experiments to quantify the contribution of individual components—specifically the generator and exponential moving average mechanism—isolating their impact on classification performance across different data availability regimes

Impact of Unsupervised Training Frequency

In our approach, the hyperparameter μ controls how frequently the unsupervised training phase is executed, with the unsupervised phase running every μ epochs. We investigate how varying this parameter affects our model's performance across different few-shot settings, as it represents the balance between supervised and unsupervised learning signals in our semi-supervised framework. Figure 3.2 presents the average accuracy across all datasets as a function of the number of shots per class for different values of μ. These results were obtained using the validation set to avoid any bias that could arise from hyperparameter selection on the test set. The key finding is that $\mu = 10$ consistently achieves the highest performance across all shot settings. In the 5-shot setting, $\mu = 10$ reaches 74.5% accuracy, representing an 8.5% point improvement over supervised-only learning ($\mu = 0$ at 66.0%). This performance advantage narrows as labeled data increases, reducing to approximately 2% points in the 50-shot setting (85.1% vs 83.1%). The figure reveals distinct patterns in how unsupervised learning frequency affects performance. Moving from supervised-only learning to any incorporation of unsupervised learning produces improvements, particularly evident in low-shot scenarios. Moderate frequencies ($\mu \in \{1, 10, 25\}$) achieve optimal performance, with $\mu = 10$ emerging as optimum. Very high frequencies ($\mu \in \{50, 100\}$) lead to performance degradation compared to the optimum, though they still outperform supervised-only learning in data-scarce settings. These findings validate our design choice of alternating between supervised and unsupervised learning phases, demonstrating that the optimal frequency ($\mu = 10$) remains consistent across different data availability scenarios and that properly scheduled unsupervised learning is particularly crucial in extreme low-data regimes. Additionally, these results shows the collaborative dynamics between the supervised phase and the combined self-supervised pre-training and synthetic data enhancement phases. The supervised-only baseline ($\mu = 0$) establishes the performance ceiling achievable without leveraging unlabeled data, reaching 66.0% in the 5-shot setting. The substantial improvement when incorporating the unsupervised phases ($\mu = 10$ achieving 74.5%, an 8.5 percentage point gain) demonstrates that the self-supervised pre-training and synthetic data enhancement phases contribute complementary learning signals that supervised training alone cannot provide. However, the performance degradation at very high frequencies ($\mu \in [50, 100]$) reveals that this collaboration requires careful balance—excessive unsupervised training can destabilize the classifier before supervised signals adequately anchor its learning. The optimal frequency ($\mu = 10$) thus creates a productive feedback loop: the supervised phase provides stable models that enable reliable pseudo-labeling in the self-supervised pre-training phase, while the synthetic data enhancement phase expands the effective training set that the supervised phase leverages in subsequent iterations. This interdependence explains why the alternating

schedule proves more effective than either continuous supervised training or more frequent unsupervised updates.

Ablation Study of Core Components

To isolate the contribution of individual components within our framework, we conduct controlled ablation experiments by systematically removing key architectural elements. Specifically, we evaluate two critical components: (1) the exponential moving average (EMA) mechanism in the pseudo-labeling ensemble, which provides temporal stability to pseudo-label predictions, and (2) the generator network, which performs class-conditioned image translation in the unsupervised learning phase. These ablations are evaluated across all four few-shot settings (5, 10, 20, and 50 shots per class), averaging results across both SPARSE (classifier-only) and $\text{SPARSE}_{\text{ens}}$ configurations to assess overall framework performance. To quantify the magnitude of each component's contribution, we employ Cohen's d effect sizes, which provide a standardized measure that accounts for variance across the eleven diverse datasets. This approach enables meaningful comparison of component importance through interpretable thresholds (negligible, small, medium, large effects), assessing not just whether differences exist, but whether they are practically significant.

Table 3.6: Component ablation analysis across few-shot settings. Average accuracy and Cohen's d effect sizes when removing individual components from the full SPARSE framework. Results averaged over SPARSE and SPARSEens configurations across all eleven datasets. Positive d indicates full model outperforms ablated variant. Effect size interpretation: $|d| < 0.2$ (negligible), $0.2 \leq |d| < 0.5$ (small), $0.5 \leq |d| < 0.8$ (medium), $|d| \geq 0.8$ (large).

Model Variant	**Component Removed**	**5-shot**	**10-shot**	**20-shot**	**50-shot**
Full SPARSE	None	66.22	70.95	75.70	78.28
Ablation 1	EMA	65.09 d=0.21 (small)	69.73 d=0.32 (small)	75.70 d=-0.00 (neg.)	78.78 d=-0.09 (neg.)
Ablation 2	Generator	59.17 d=0.70 (medium)	65.92 d=0.66 (medium)	70.98 d=0.72 (medium)	72.48 d=0.85 (large)

Table 3.6 presents the ablation results, showing the mean performance across both inference configurations (SPARSE and $\text{SPARSE}_{\text{ens}}$) when averaged over all eleven datasets. Cohen's d effect sizes quantify the magnitude of each component's contribution. The analysis reveals that the generator component is the most critical element of the framework across all data availability regimes. Removing the generator produces medium effect sizes in the first three settings (5-shot: d = 0.702, 10-shot: d = 0.663, 20-shot: d = 0.715) and escalates to large magnitude in the 50-shot setting (d = 0.853). This increasing pattern demonstrates that the generator provides fundamental value through its image-to-image translation mechanism that enriches feature representations synergistically with supervised learning signals,

rather than merely compensating for data scarcity. The modification of real unlabeled images preserves authentic anatomical features while introducing semantic variations that enhance classification performance even when moderate amounts of labeled data are available. In contrast, the EMA mechanism shows negligible to small effect sizes across all configurations (5-shot: d = 0.208, 10-shot: d = 0.318, 20-shot: d = -0.001, 50-shot: d = -0.085), with none achieving statistical significance. This indicates that the confidence-weighted ensemble mechanism already provides substantial stability by combining predictions from multiple models, rendering additional temporal smoothing non-essential for the framework's success. Additionally, the generator ablation reveals the specific contribution of the synthetic data enhancement phase to the overall framework collaboration. Removing the generator eliminates this phase while preserving the supervised and self-supervised pre-training phases, resulting in performance drops of 7.1 percentage points in 5-shot and 5.8 percentage points in 50-shot settings. The escalating effect size demonstrates that the synthetic data enhancement phase amplifies the benefits of supervised learning through a specific mechanism: the self-supervised pre-training phase learns discriminative class-conditional transformations from unlabeled data, which the synthetic data enhancement phase exploits to generate high-quality samples with known labels. These synthetic samples subsequently augment the supervised phase in later training epochs, creating a three-way interaction where each phase enhances the effectiveness of the others. This collaborative mechanism—operating iteratively across the alternating training schedule—explains why our framework substantially outperforms methods employing only supervised learning with generation or two-phase approaches that lack this complete feedback loop.

3.5.4 Computational Analysis

The use of three specialized networks in our framework introduces computational overhead during both training and inference. This section provides a quantitative analysis of these requirements and examines and examines the trade-offs between computational cost and classification performance.

Table 3.7 presents the computational complexity breakdown for each network component. During training, all three networks participate simultaneously across both learning phases. The supervised phase requires 8.767 GFLOPs per sample as the generator's encoder $E_{\mathscr{G}}$, discriminator $\mathscr{D}$ and classifier $\mathscr{C}$ jointly process the limited labeled data. The unsupervised phase is substantially more expensive at 25.170 GFLOPs per sample, with the complete generator architecture performing class-conditioned image translation accounting for 23.135 GFLOPs (92% of the total). The remaining computational cost is distributed between the

Table 3.7: Computational complexity analysis of SPARSE framework components. Measurements are reported for input images of size 128×128×3 pixels with 8 output classes.

Model Component	**GFLOPs**	**Params (M)**
Generator (image translation)	23.135	8.140
Generator (classifier mode)	6.732	4.693
Discriminator (PatchGAN)	1.368	44.787
Classifier (EfficientNet-B3)	0.667	10.709
TOTAL (Supervised phase)	**8.767**	**60.188**
TOTAL (Unsupervised phase)	**25.170**	**60.188**
TOTAL (SPARSE inference)	**0.667**	**10.709**
TOTAL ($\text{SPARSE}_{\text{ens}}$ inference)	**2.035**	**55.496**

discriminator (1.368 GFLOPs) and classifier (0.667 GFLOPs). The overall training cost depends critically on the frequency parameter μ. With $\mu = 10$, the configuration that achieves optimal performance across all few-shot settings as demonstrated in Section 3.5.2 — the expensive unsupervised phase occurs once every 10 epochs. This results in an average computational cost per epoch that is dominated by the unsupervised phase when considering the substantially larger unlabeled dataset ($M >> M$). The parameter μ provides a direct mechanism to modulate computational cost: increasing μ 50 or 100 reduces the frequency of unsupervised training, thereby decreasing overall training time. However, as shown in Figure 3.2, this comes at the cost of reduced classification accuracy, illustrating the fundamental trade-off between computational efficiency and model performance. The ensemble-based pseudo-labeling mechanism adds minimal overhead beyond the forward passes required for prediction. Computing confidence-weighted ensemble predictions involves averaging probability distributions and selecting samples above a percentile-based threshold—operations with complexity $O(K)$ per sample, where K is the number of classes. The temporal ensemble with exponential moving average maintains running statistics without additional forward passes, adding only $O(K)$ memory overhead per unlabeled sample.

At inference, computational requirements reduce substantially. The generator, responsible for the majority of training overhead, is discarded entirely. The SPARSE configuration requires only 0.667 GFLOPs per image using the classifier alone, while $\text{SPARSE}_{\text{ens}}$ requires 2.035 GFLOPs through ensemble averaging of the discriminator and classifier predictions. This represents a 3 times increase in inference cost for $\text{SPARSE}_{\text{ens}}$, which must be weighed against its performance gains. The empirical results demonstrate that these computational investments yield substantial performance improvements in data-scarce regimes. As shown in Table 3.3, $\text{SPARSE}_{\text{ens}}$ achieves 66.22% average accuracy with only 5 labeled samples per

class. When compared to two-player architectures—SGAN (25.80%), MatchGAN (39.88%), and CISSL (45.84%)—the performance gap demonstrates the value of the three-network design. This advantage persists even when comparing against methods with 10 times more labeled data: SGAN reaches only 27.28% with 50 samples per class, indicating that architectural sophistication in the three-network design provides learning capabilities that simpler models cannot achieve through moderate increases in labeled data alone. The $\text{SPARSE}_{\text{ens}}$ configuration further justifies its 3 times inference cost through statistically significant improvements over SPARSE as detailed in Table 3.4.

3.6 Conclusions

This chapter has introduced a novel GAN-based semi-supervised learning framework specifically designed for extremely low labeled-data regimes. Our approach addresses the fundamental challenge of insufficient labeled data through three key innovations: (1) a dynamic training schedule that alternates between supervised and unsupervised phases, (2) an image-to-image translation mechanism that enriches feature representations by learning class-conditional transformations from real unlabeled images, and (3) a confidence-weighted temporal ensemble technique for reliable pseudo-labeling. By leveraging these components within a three-player GAN architecture, our method effectively combines the complementary strengths of a generator, discriminator, and dedicated classifier. The comprehensive empirical evaluation across eleven MedMNIST datasets demonstrates the effectiveness of our approach. In the extreme 5-shot setting, our ensemble configuration achieved statistically significant improvements over six state-of-the-art semi-supervised methods, with effect sizes ranging from large to very large. The method's superiority stems from its fundamental design choice of performing image-to-image translation rather than generating images from noise, which leverages real medical images as the foundation for learning discriminative features crucial for classification tasks. This advantage is further amplified by our temporal ensemble mechanism, which aggregates predictions across training epochs to produce more reliable pseudo-labels in data-scarce scenarios. Our analysis of the unsupervised training frequency revealed that moderate alternation between supervised and unsupervised phases ($\mu = 10$) balances both learning signals. This finding provides practical guidance for deployment, as the optimal frequency remained consistent across different data availability scenarios, eliminating the need for extensive hyperparameter tuning in clinical applications. Despite these promising results, several avenues for future research remain. The computational requirements of maintaining multiple networks may pose challenges in resource-constrained clinical settings, motivating the development of more efficient architectures.

Another promising direction builds on the uncertainty-aware components already embedded in the framework: calibrating the predicted probabilities and complementing them with explicit uncertainty estimates would allow low-confidence predictions to be deferred to a clinician, improving reliability in decision-support applications, while interpretability tools that highlight the image regions driving each prediction would further support transparent clinical adoption. Additionally, extending the framework to incorporate domain-specific medical knowledge and multi-modal imaging data could further enhance its clinical applicability.
In conclusion, our GAN-based semi-supervised learning framework represents a significant advancement in addressing the labeled data scarcity challenge in medical imaging. By effectively leveraging both labeled and unlabeled data through image translation and ensemble techniques, our method enables robust classification performance even with as few as five labeled samples per class, offering a practical solution for medical imaging applications where annotation costs are prohibitive.

Chapter 4

Conclusions and Future Directions

This thesis has addressed the pressing need for accurate and data-efficient artificial intelligence in endoscopic and robotic surgery, with a particular focus on restoring the three-dimensional perceptual information lost through the monocular two-dimensional camera feed that underlies both paradigms [25, 10, 15]. Recognizing the limitations of current surgical visualization, including the loss of depth perception, the presence of blind spots, and the substantial expertise required to compensate for them, the work presented here leverages deep learning to recover this missing perceptual information and to enable reliable downstream scene-understanding tasks. To address this objective, the thesis was structured around two research questions (RQ1–RQ2), each targeting a specific challenge in deploying AI within the operating room: the scarcity of labeled surgical and medical data, and the strict real-time inference budget imposed by clinical use. The key findings corresponding to each research question are summarized below.

RQ1: Can deep learning methods learn accurate and generalizable representations for surgical perception and scene understanding while operating under severe labeled data constraints? Both core contributions of this thesis provided convergent evidence that such learning is possible. BodySLAM's CycleVO module learned pose representations entirely from unlabeled endoscopic video through a GAN-based cyclic image-translation pretext task, generalizing across laparoscopy, gastroscopy, and colonoscopy without domain-specific retraining and matching or surpassing four state-of-the-art methods trained on the target datasets; the integration of the Zoe model further demonstrated that strong cross-domain depth estimation is attainable without any fine-tuning on endoscopic data. SPARSE addressed the same label-scarcity challenge in the downstream classification setting through three coordinated innovations within a three-player GAN architecture: a dynamic training schedule that alternates supervised phases on the few labeled samples with unsupervised phases that exploit the abundant unlabeled pool; a class-conditioned image-to-

image translation mechanism that enriches feature representations by modifying real medical images rather than generating synthetic ones from noise, thereby preserving the authentic anatomical structure of the input; and a confidence-weighted temporal ensemble that stabilizes pseudo-labeling across training iterations by aggregating predictions from multiple models and from historical training states. Evaluated across eleven MedMNIST benchmarks with as few as five labeled samples per class, SPARSE achieved statistically significant gains over six state-of-the-art semi-supervised methods, with effect sizes ranging from large to very large. These findings confirm that RQ1 admits a positive answer, and that accurate and generalizable representations for surgical perception and medical image classification can in fact be learned under extreme label scarcity, provided that the learning objective is carefully decoupled from the availability of annotations.

RQ2: Can AI systems for surgical perception and scene understanding satisfy the strict real-time inference requirements of clinical deployment without sacrificing accuracy or generalizability? Computational efficiency was treated as a first-class design constraint throughout the thesis, rather than as a post-hoc concern. Within BodySLAM, CycleVO achieved the lowest inference time among all compared pose estimation methods on both the EndoSLAM and SCARED datasets while matching or exceeding their accuracy, directly demonstrating that data-efficient and real-time objectives need not be in tension. Within SPARSE, the training cost of maintaining three coordinated networks is substantially amortized at deployment: the generator, which accounts for the overwhelming majority of the computational load during the unsupervised phase, is discarded entirely at inference, so the deployed system reduces to the lightweight classifier alone or, in the ensemble configuration, to the classifier paired with the discriminator. The ensemble variant remains considerably lighter than the training-time footprint while still delivering statistically significant accuracy improvements over the classifier alone, demonstrating that the architectural sophistication required for robust few-shot learning does not need to be paid for at inference time. These results provide a clear and affirmative answer to RQ2, showing that computational viability and generalizable accuracy can be engineered to coexist within a single framework. Across all contributions, this thesis has proposed and validated a coherent methodological line that addresses the critical barriers to deploying AI in endoscopic and robotic surgery: the scarcity of labeled medical data, the need for cross-domain generalization across diverse surgical modalities, and the strict computational budgets imposed by real-time use in the operating room. The proposed solutions, namely unsupervised and self-supervised pretext tasks, image-to-image translation operating on real unlabeled data rather than on noise vectors, confidence-weighted temporal ensembling for reliable pseudo-labeling, and the strategic discarding of expensive components at deployment time, decouple

the learning signal from the label supply while keeping inference-time complexity low. Taken together, these strategies concretely demonstrate that AI-enhanced surgical perception can be made accurate, generalizable, and computationally tractable within a single, clinically-oriented research programme that addresses the compounding barriers identified at the outset of this thesis rather than treating them as isolated concerns.

Looking ahead, several avenues for future work open up at the level of the thesis as a whole, beyond the per-contribution refinements already discussed in Chapters 2 and 3. First, the two contributions were developed as independent building blocks, each targeting a distinct sub-task of surgical perception; their integration into a unified pipeline is the most natural next step. The three-dimensional scene reconstructed by BodySLAM could provide spatial and geometric context for SPARSE-style few-shot classification of anatomical regions, lesions, or surgical phases, while the semi-supervised representations learned by SPARSE could in turn inform the feature extraction stages of the SLAM pipeline, moving the research programme from isolated perception modules toward a coherent surgical scene-understanding system. Second, both contributions independently converge on image-to-image translation as an effective pretext task for learning under label scarcity, with CycleVO relying on cyclic translation between consecutive frames and SPARSE on class-conditioned translation between real medical images. Abstracting this convergence into a general image-translation-as-supervision paradigm, and extending it to dense surgical prediction tasks such as instrument segmentation, tissue boundary delineation, and workflow recognition, would broaden the clinical reach of the methods developed here while preserving their data efficiency. Third, the data-efficient strategies developed here could be integrated with the rapidly maturing ecosystem of foundation models. Vision transformers and large multimodal models, which jointly reason over images and additional sources of clinical context, offer broadly pretrained representations that could serve as backbones for both CycleVO-style pose learning and SPARSE-style few-shot classification. Combining the annotation efficiency of the pretext tasks proposed in this thesis with the transferable priors of such models is a promising route toward surgical perception systems that are at once data-efficient, generalizable, and readily adaptable to new procedures and imaging devices. Finally, all of the validation reported in this thesis has been performed *in silico* on publicly available benchmarks. The natural thesis-level trajectory is therefore toward prospective *in vivo* validation of an integrated BodySLAM and SPARSE system on real surgical workflows, conducted in close collaboration with clinical partners, with outcomes measured not only through benchmark metrics but also through clinically meaningful endpoints such as surgeon cognitive load, procedural accuracy, and ultimately patient outcomes. Ultimately, the integration of these strategies moves us closer to realizing non-invasive, accurate, data-efficient, and computationally viable

AI systems that can be translated into routine clinical practice for endoscopic and robotic surgical care.

4.1 Toward Clinical Translation

The methods proposed in this thesis have been developed and validated as research contributions, and several research directions remain before their benefits can be realized in clinical practice, where a range of technical and clinical barriers still separates promising results from routine use [15, 96]. A first direction concerns the feasibility of operating within the practical constraints of the operating room. Because both frameworks work on the monocular video stream already produced by existing endoscopic and robotic platforms, they require no additional sensing hardware, which makes them well suited to integration studies; nonetheless, future work should characterize their end-to-end latency within realistic intra-operative pipelines and study how several perception modules can be combined without exceeding the real-time budget, since the inference of deep models does not always satisfy real-time constraints and combining multiple functionalities entails a trade-off between tractability, robustness, and accuracy [15]. The emphasis on computational efficiency adopted throughout this thesis is a first step in this direction.

A second direction concerns how the outputs of these systems would be used by the surgical team. Rather than acting autonomously, perception outputs such as the three-dimensional reconstructions produced by BodySLAM or the tissue classifications produced by SPARSE are best framed as decision support that augments the surgeon's judgment, which raises research questions about how such information should be presented and how it influences clinical decision-making [15]. Investigating these questions calls for user studies involving surgeons, in which the usability and perceived trustworthiness of the system are assessed directly, and it further motivates the interpretability and calibrated-uncertainty mechanisms discussed in Chapter 3, which would help clinicians judge when a prediction can be relied upon.

The most important direction, however, is broader empirical validation. Beyond the prospective *in vivo* evaluation already outlined above, establishing the generalizability of the proposed methods will require testing them on multicentre datasets acquired with different imaging systems and under heterogeneous clinical protocols, moving beyond the single public benchmarks used in this work. Studies of this kind, progressively complemented by prospective clinical evidence of the sort that has begun to accompany AI systems in colonoscopy [83], would both probe the generalization claims made here and provide the evidence on which the translation of surgical data science into practice ultimately depends [44].

Pursuing these directions is the natural next step toward turning the contributions of this thesis into tools that can meaningfully support surgical care.

Bibliography

[1] Stereo correspondence and reconstruction of endoscopic data sub-challenge. https://endovissub2019-scared.grand-challenge.org/, 2019. Accessed: 2024-06-17.

[2] M. Areia, Y. Mori, L. Correale, A. Repici, M. Bretthauer, P. Sharma, et al. Cost-effectiveness of artificial intelligence for screening colonoscopy: a modelling study. *Lancet Digital Health*, 4(6):e436–e444, 2022.

[3] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification, 2023.

[4] Sophia Bano, Francisco Vasconcelos, Anna L. David, Jan Deprest, and Danail Stoyanov. Placental vessel-guided hybrid framework for fetoscopic mosaicking. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 11(4):1166–1171, 2023.

[5] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3319–3327, 2017.

[6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In Aleš Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision – ECCV 2006*, pages 404–417, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[7] Tobias Bergen and Thomas Wittenberg. Stitching and surface reconstruction from endoscopic image sequences: A review of applications and methods. *IEEE Journal of Biomedical and Health Informatics*, 20(1):304–321, 2016.

[8] Shariq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Mueller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 02 2023.

[9] Lorenzo Bianchi, Francesco Chessa, Andrea Angiolini, Laura Cercenelli, Simonetta Lodi, Barbara Bortolani, Elena Molinaroli, Carlo Casablanca, Matteo Droghetti, Caterina Gaudiano, Francesca Giunchi, Riccardo Schiavina, Michelangelo Fiorentino, Emanuela Marcelli, Stefano Diciotti, and Eugenio Brunocilla. The use of augmented reality to guide robot-assisted partial nephrectomy: preliminary experience with Hugo RAS system. *European Urology Open Science*, 61:98–101, 2024.

[10] R. Bogdanova, P. Boulanger, and B. Zheng. Depth perception of surgeons in minimally invasive surgery. *Surgical Innovation*, 23(5):515–524, 2016.

[11] A. Buia, F. Stockhausen, and E. Hanisch. Laparoscopic surgery: A qualified systematic review. *World Journal of Methodology*, 5(4):238–254, 2015.

[12] S. Bull, P. Y. Cheah, S. Denny, I. Jao, V. Marsh, L. Merson, N. Shah More, L. T. Nhan, D. Osrin, D. Tangseefa, D. Wassenaar, and M. Parker. Best practices for ethical sharing of individual-level health research data from low- and middle-income settings. *Journal of Empirical Research on Human Research Ethics*, 10(3):302–313, 2015.

[13] Carlos Campos, Richard Elvira, Juan J. Gomez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

[14] Yue Cao, Thomas Andrew Geddes, Jean Yee Hwa Yang, and Pengyi Yang. Ensemble deep learning in bioinformatics. *Nature Machine Intelligence*, 2(9):500–508, September 2020.

[15] F. Chadebecq, L. B. Lovat, and D. Stoyanov. Artificial intelligence and automation in endoscopy and surgery. *Nature Reviews Gastroenterology & Hepatology*, 20(3):171–182, 2023.

[16] Muhammad Nadeem Cheema, Anam Nazir, Bin Sheng, Ping Li, Jing Qin, Jinman Kim, and David Dagan Feng. Image-aligned dynamic liver reconstruction using intra-operative field of views for minimal invasive surgery. *IEEE Transactions on Biomedical Engineering*, 66(8):2163–2173, 2019.

[17] Jianguo Chen, Nan Yang, Yuhui Pan, Hailing Liu, and Zhaolei Zhang. Synchronous medical image augmentation framework for deep learning-based image segmentation. *Computerized Medical Imaging and Graphics*, 104:102161, 2023.

[18] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative

adversarial nets. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[19] Xuxin Chen, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C. Thai, Kathleen Moore, Robert S. Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis*, 79:102444, 2022.

[20] V. Chiruvella and A. K. Guddati. Ethical issues in patient data ownership. *Interactive Journal of Medical Research*, 10(2):e22269, 2021.

[21] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996.

[22] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Russ R Salakhutdinov. Good semi-supervised learning that requires a bad gan. *Advances in neural information processing systems*, 30, 2017.

[23] Sian Davies, Mohammed Ghallab, Shahab Hajibandeh, Shahin Hajibandeh, and Sarah Addison. Three-dimensional versus two-dimensional imaging during laparoscopic cholecystectomy: a systematic review and meta-analysis of randomised controlled trials. *Langenbeck's Archives of Surgery*, 405(5):563–572, Aug 2020.

[24] Simon DiMaio, Mike Hanuschik, and Usha Kreaden. The da Vinci surgical system. *Surgical Robotics: Systems Applications and Visions*, pages 199–217, 2011.

[25] A. F. Durrani and G. M. Preminger. Three-dimensional video imaging for endoscopic surgery. *Computers in Biology and Medicine*, 25(2):237–247, 1995.

[26] Asim F. Durrani and Glenn M. Preminger. Three-dimensional video imaging for endoscopic surgery. *Computers in Biology and Medicine*, 25(2):237–247, 1995. Virtual Reality for Medicine.

[27] Itai Fried, Douglas P. Hsu, Daniel Zamora, Ashish Chiluwal, David Hamilton, Christine T. Lau, Jamee Klaseus, David M. Mahvi, and Naveen Agrawal. Ion endoluminal system: a robotic-assisted bronchoscopy platform for peripheral pulmonary nodule biopsy. *Journal of Bronchology & Interventional Pulmonology*, 27(4):e63–e66, 2020.

[28] Scott Geng, Cheng-Yu Hsieh, Vivek Ramanujan, Matthew Wallingford, Chun-Liang Li, Pang Wei Koh, and Ranjay Krishna. The unmet promise of synthetic training images: Using retrieved real images performs better, 2025.

[29] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3827–3837, 2019.

[30] Valerio Guarrasi, Natascha Claudia D'Amico, Rosa Sicilia, Ermanno Cordelli, and Paolo Soda. Pareto optimization of deep networks for covid-19 diagnosis from chest x-rays. *Pattern Recognition*, 121:108242, 2022.

[31] John J. Han, Ayberk Acar, Callahan Henry, and Jie Ying Wu. Depth anything in medical images: A comparative study, 2024.

[32] Ayaan Haque. Ec-gan: Low-sample classification using semi-supervised algorithms and gans (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 15797–15798, 2021.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[34] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 225–234, 2007.

[35] A. R. Lanfranco, A. E. Castellanos, J. P. Desai, and W. C. Meyers. Robotic surgery: a current perspective. *Annals of Surgery*, 239(1):14–21, 2004.

[36] Chongxuan Li, Kun Xu, Jun Zhu, Jiashuo Liu, and Bo Zhang. Triple generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9629–9640, 2022.

[37] Liang Li, Sophia Bano, Jan Deprest, Anna L. David, Danail Stoyanov, and Francisco Vasconcelos. Globally optimal fetoscopic mosaicking based on pose graph optimisation with affine constraints. *IEEE Robotics and Automation Letters*, 6(4):7831–7838, 2021.

[38] Yang Li, Florian Richter, Jingpei Lu, Emily K. Funk, Ryan K. Orosco, Jianke Zhu, and Michael C. Yip. Super: A surgical perception framework for endoscopic tissue

manipulation with surgical robotics. *IEEE Robotics and Automation Letters*, 5(2):2294–2301, 2020.

[39] Ziteng Liu, Wenpeng Gao, Jiahua Zhu, Zhi Yu, and Yili Fu. Surface deformation tracking in monocular laparoscopic video. *Medical Image Analysis*, 86:102775, 2023.

[40] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(5828):133–135, 1981.

[41] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov 2004.

[42] X. Luo, K. Mori, and T. M. Peters. Advanced endoscopic navigation: surgical big data, methodology, and applications. *Annual Review of Biomedical Engineering*, 20:221–251, 2018.

[43] Nader Mahmoud, Alexandre Hostettler, Toby Collins, Luc Soler, Christophe Doignon, and J. Montiel. Slam based quasi dense reconstruction for minimally invasive surgery scenes. 05 2017.

[44] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, et al. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696, 2017.

[45] Karim Makki, Kilian Chandelon, and Adrien Bartoli. Elliptical specularity detection in endoscopy with application to normal reconstruction. *International Journal of Computer Assisted Radiology and Surgery*, 18(7):1323–1328, Jul 2023.

[46] Guido Manni, Clemente Lauretti, Francesco Prata, Rocco Papalia, Loredana Zollo, and Paolo Soda. BodySLAM: A generalized monocular visual SLAM framework for surgical applications. *Image and Vision Computing*, 174:106094, 2026.

[47] Guido Manni, Clemente Lauretti, Loredana Zollo, and Paolo Soda. SPARSE data, rich results: Few-shot semi-supervised learning via class-conditioned image translation. *Computerized Medical Imaging and Graphics*, 128:102705, 2026.

[48] Armin Masoumian, Hatem A. Rashwan, Julián Cristiano, M. Salman Asif, and Domenec Puig. Monocular depth estimation using deep learning: A review. *Sensors*, 22(14), 2022.

[49] A. Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

[50] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

[51] Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Taylor L. Bobrow, Gulfize Coskun, Kagan Incetan, Yasin Almalioglu, Faisal Mahmood, Eva Curto, Luis Perdigoto, Marina Oliveira, Hasan Sahin, Helder Araujo, Henrique Alexandrino, Nicholas J. Durr, Hunter B. Gilbert, and Mehmet Turan. Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Medical Image Analysis*, 71:102058, 2021.

[52] Brusabhanu S. Peters, Priscila R. Armijo, Crystal Krause, Songita A. Choudhury, and Dmitry Oleynikov. Review of emerging surgical robotic technology. *Surgical Endoscopy*, 32(4):1636–1655, 2018.

[53] Rene Ranftl, Katrin Lasinger, David Hafner, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 08 2020.

[54] David Recasens, José Lamarca, José M. Fácil, J. M. M. Montiel, and Javier Civera. Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *IEEE Robotics and Automation Letters*, 6(4):7225–7232, 2021.

[55] Edgar Rojas-Muñoz, Daniel Andersen, Maria Eugenia Cabrera, Voicu Popescu, Sherri Marley, Brian Mullis, Ben Zarzaur, and Juan Wachs. Evaluation of an augmented reality platform for austere surgical telementoring: a randomized controlled crossover study in cricothyroidotomies. *NPJ Digital Medicine*, 3(1):75, 2020.

[56] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[57] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, 2011.

[58] R. Rudiman. Minimally invasive gastrointestinal surgery: from past to the future. *Annals of Medicine and Surgery*, 71:102922, 2021.

[59] Filippo Ruffini, Elena Mulero Ayllon, Linlin Shen, Paolo Soda, and Valerio Guarrasi. Benchmarking foundation models and parameter-efficient fine-tuning for prognosis prediction in medical imaging, 2025.

[60] A. R. Sajun and I. Zualkernan. Survey on implementations of generative adversarial networks for semi-supervised learning. *Applied Sciences*, 12(3):1718, 2022.

[61] Ali Reza Sajun and Imran Zualkernan. Survey on implementations of generative adversarial networks for semi-supervised learning. *Applied Sciences*, 12(3), 2022.

[62] Marco Salmè, Lorenzo Tronchin, Rosa Sicilia, Paolo Soda, and Valerio Guarrasi. Beyond the generative learning trilemma: Generative model assessment in data scarcity domains, 2025.

[63] Narimantas E. Samalavicius, Vinsas Janusonis, Robertas Siaulys, Marius Jasenas, Olegas Deduchovas, Raimondas Venckus, Aurika Karbonskiene, Giedre Smailyte, Zilvinas Saladzinskas, Algimantas Tamelis, et al. Robotic surgery using Senhance robotic platform: single center experience with first 100 cases. *Journal of Robotic Surgery*, 14(2):371–376, 2020.

[64] L. F. Sánchez-Peralta, L. Bote-Curiel, A. Picón, F. M. Sánchez-Margallo, and J. B. Pagador. Deep learning to find colorectal polyps in colonoscopy: a systematic literature review. *Artificial Intelligence in Medicine*, 108:101823, 2020.

[65] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8011–8021, 2023.

[66] A. Schmidt, O. Mohareri, S. DiMaio, M. C. Yip, and S. E. Salcudean. Tracking and mapping in medical computer vision: A review. *Medical Image Analysis*, 94:103131, 2024.

[67] Adam Schmidt, Omid Mohareri, Simon DiMaio, Michael C. Yip, and Septimiu E. Salcudean. Tracking and mapping in medical computer vision: A review. *Medical Image Analysis*, 94:103131, 2024.

[68] Karin Schwab, Ross Smith, Vanessa Brown, Michael Whyte, and Isabella Jourdan. Evolution of stereoscopic imaging in surgery and recent advances. *World Journal of Gastrointestinal Endoscopy*, 9(8):368–377, 2017.

[69] Johannes Schüle, Jonathan Haag, Peter Somers, Carina Veil, Christina Tarín, and Oliver Sawodny. A model-based simultaneous localization and mapping approach for deformable bodies. In *2022 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 607–612, 2022.

[70] J. Shah, A. Vyas, and D. Vyas. The history of robotics in surgical specialties. *American Journal of Robotic Surgery*, 1(1):12–20, 2014.

[71] Shao, Shuwei, Pei, Zhongcai, Chen, Weihai, Zhu, Wentao, Wu, Xingming, Sun, Dianmin, and Baochang Zhang. Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. *Medical image analysis*, 77:102338, 2022.

[72] R Smith, K Schwab, A Day, T Rockall, K Ballard, M Bailey, and I Jourdan. Effect of passive polarizing three-dimensional displays on surgical performance for experienced laparoscopic surgeons. *British Journal of Surgery*, 101(11):1453–1459, 08 2014.

[73] L. Stewart and L. W. Way. The prevention of laparoscopic bile duct injuries: An analysis of 300 cases from a human factors and cognitive psychology perspective. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 51, pages 617–620, 2007.

[74] Lygia Stewart and Lawrence W. Way. The prevention of laparoscopic bile duct injuries: An analysis of 300 cases of from a human factors and cognitive psychology perspective. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 51(11):617–620, 2007.

[75] Jiaze Sun, Binod Bhattarai, and Tae-Kyun Kim. Matchgan: A self-supervised semi-supervised conditional generative adversarial network. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[76] Stine Maya Dreier Sørensen, Mona Meral Savran, Lars Konge, and Flemming Bjerrum. Three-dimensional versus two-dimensional vision in laparoscopy: a systematic review. *Surgical Endoscopy*, 30(1):11–23, Jan 2016.

[77] Timo Teufel, Hongchao Shu, Roger D. Soberanis-Mukul, Jan Emily Mangulabnan, Manish Sahu, S. Swaroop Vedula, Masaru Ishii, Gregory Hager, Russell H. Taylor, and Mathias Unberath. Oneslam to map them all: a generalized approach to slam for monocular endoscopic imaging based on tracking any point. *International Journal of Computer Assisted Radiology and Surgery*, 19(7):1259–1266, 2024.

[78] Brian C. Thomas, Mark Slack, Muthusamy Hussain, Neil Barber, Amit Pradhan, Edward Dinneen, and Gerry D. Stewart. Preclinical evaluation of the Versius surgical system, a new robot-assisted surgical device for use in minimally invasive general surgery. In *Journal of Laparoendoscopic & Advanced Surgical Techniques*, volume 31, pages 68–73, 2021.

[79] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment — a modern synthesis. In Bill Triggs, Andrew Zisserman, and Richard Szeliski, editors, *Vision Algorithms: Theory and Practice*, pages 298–372, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[80] G. Urban, P. Tripathi, T. Alkayali, M. Mittal, F. Jalali, W. Karnes, and P. Baldi. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*, 155(4):1069–1078, 2018.

[81] J. E. van Engelen and H. H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

[82] Nereo Vettoretto, Emanuela Foglia, Lucrezia Ferrario, Alberto Arezzo, Roberto Cirocchi, Gianfranco Cocorullo, Giuseppe Currò, Domenico Marchi, Giuseppe Portale, Chiara Gerardi, Umberto Nocco, Michele Tringali, Gabriele Anania, Micaela Piccoli, Gianfranco Silecchia, Mario Morino, Andrea Valeri, and Emauele Lettieri. Why laparoscopists may opt for three-dimensional view: a summary of the full hta report on 3d versus 2d laparoscopy by s.i.c.e. (società italiana di chirurgia endoscopica e nuove tecnologie). *Surgical Endoscopy*, 32(6):2986–2993, Jun 2018.

[83] P. Wang, T. M. Berzin, J. R. Glissen Brown, S. Bharadwaj, A. Becq, X. Xiao, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut*, 68(10):1813–1819, 2019.

[84] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[85] Yingpeng Xie, Qiwei Wan, Hai Xie, Yanwu Xu, Tianfu Wang, Shuqiang Wang, and Baiying Lei. Fundus image-label pairs synthesis and retinopathy screening via gans with class-imbalanced semi-supervised learning. *IEEE Transactions on Medical Imaging*, 42(9):2714–2725, 2023.

[86] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[87] Q. Zeng, Z. Lu, Y. Xie, et al. Pick: Predict and mask for semi-supervised medical image segmentation. *International Journal of Computer Vision*, 133:3296–3311, 2025.

[88] Q. Zeng, Y. Xie, Z. Lu, M. Lu, Y. Wu, and Y. Xia. Segment together: A versatile paradigm for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(7):2948–2959, July 2025.

[89] Q. Zeng, Y. Xie, Z. Lu, M. Lu, J. Zhang, and Y. Xia. Consistency-guided differential decoding for enhancing semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(1):44–56, January 2025.

[90] Q. Zeng, Y. Xie, Z. Lu, and Y. Xia. Pefat: Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15671–15680, Vancouver, BC, Canada, 2023.

[91] Shuai Zhang, Liang Zhao, Shoudong Huang, Ruibin Ma, Boni Hu, and Qi Hao. 3d reconstruction of deformable colon structures based on preoperative model and deep neural network. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1875–1881, 2021.

[92] Shuai Zhang, Liang Zhao, Shoudong Huang, Hua Wang, Qi Luo, and Qi Hao. Slam-tka: Real-time intra-operative measurement of tibial resection plane in conventional total knee arthroplasty. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 126–135, Cham, 2022. Springer Nature Switzerland.

[93] Qingyu Zhao, True Price, Stephen Pizer, Marc Niethammer, Ron Alterovitz, and Julian Rosenman. The endoscopogram: A 3d model reconstructed from endoscopic video frames. In Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 439–447, Cham, 2016. Springer International Publishing.

[94] Hao Zhen, Yucheng Shi, Jidong J. Yang, and Javad Mohammadpour Vehni. Co-supervised learning paradigm with conditional generative adversarial networks for sample-efficient classification. *Applied Computing and Intelligence*, 3(1):13–26, 2023.

[95] Qian-Yi Zhou and Vladlen Koltun. Dense scene reconstruction with points of interest. *ACM Transactions on Graphics*, 32, 07 2013.

[96] X. Y. Zhou, Y. Guo, M. Shen, and G. Z. Yang. Application of artificial intelligence in surgery. *Frontiers of Medicine*, 14:417–430, 2020.

[97] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

Appendices

Appendix A

Per-Dataset Results and Dataset-Characteristic Analysis for SPARSE

This appendix reports the full per-dataset accuracy breakdown of the SPARSE framework under all four few-shot regimes considered in Chapter 3, together with an analysis of how the characteristics of each MedMNIST dataset relate to the observed performance patterns.

Table A.1: Comparison of accuracy scores across different methods on medical image datasets using 5-shot learning. Best performance for each dataset is highlighted in bold.

Dataset	SPARSE	$\text{SPARSE}_{\text{ens}}$	SGAN	MatchGAN	CISSL	SEC-GAN	TripleGAN	ECGAN
bloodmnist	0.868	**0.887**	0.173	0.302	0.463	0.813	0.862	0.241
breastmnist	0.542	0.667	0.667	**0.792**	0.615	0.719	0.677	0.667
chestmnist	0.545	0.571	0.539	0.535	**0.610**	0.533	0.510	0.532
dermamnist	0.511	0.576	0.124	**0.626**	0.566	0.535	0.563	0.626
octmnist	0.488	0.465	0.263	0.333	0.254	0.363	**0.528**	0.265
organamnist	0.780	**0.801**	0.180	0.114	0.422	0.683	0.780	0.198
organcmnist	0.754	**0.792**	0.104	0.268	0.436	0.683	0.782	0.200
organsmnist	0.567	**0.588**	0.051	0.216	0.215	0.487	0.561	0.141
pathmnist	0.751	**0.778**	0.100	0.187	0.367	0.453	0.676	0.186
pneumoniamnist	0.796	0.806	0.585	0.710	0.785	**0.854**	0.762	0.767
tissuemnist	0.351	0.353	0.052	0.302	0.308	0.339	**0.363**	0.040

A.1 Dataset Characteristics and Performance Patterns

We analyze the relationship between dataset characteristics and SPARSE performance to identify which datasets benefit most from our approach and why. To understand when and

Table A.2: Comparison of accuracy scores across different methods on medical image datasets using 10-shot learning. Best performance for each dataset is highlighted in bold.

Dataset	SPARSE	$\text{SPARSE}_{\text{ens}}$	SGAN	MatchGAN	CISSL	SEC-GAN	TripleGAN	ECGAN
bloodmnist	0.895	**0.907**	0.171	0.407	0.505	0.851	0.890	0.266
breastmnist	0.604	0.698	0.688	0.760	0.615	0.781	0.677	**0.792**
chestmnist	0.563	**0.568**	0.517	0.532	0.559	0.536	0.549	0.532
dermamnist	**0.629**	0.626	0.114	0.626	0.626	0.527	0.567	0.625
octmnist	0.490	0.515	0.246	0.240	0.316	0.481	**0.606**	0.247
organamnist	0.820	**0.834**	0.194	0.108	0.393	0.795	0.797	0.103
organcmnist	0.765	**0.808**	0.148	0.744	0.405	0.676	0.793	0.167
organsmnist	0.662	**0.694**	0.050	0.495	0.230	0.521	0.605	0.188
pathmnist	0.835	**0.873**	0.099	0.372	0.332	0.657	0.792	0.050
pneumoniamnist	0.854	0.848	0.577	0.775	0.831	0.838	**0.881**	0.802
tissuemnist	0.417	**0.434**	0.052	0.302	0.336	0.355	0.409	0.040

Table A.3: Comparison of accuracy scores across different methods on medical image datasets using 20-shot learning. Best performance for each dataset is highlighted in bold.

Dataset	SPARSE	$\text{SPARSE}_{\text{ens}}$	SGAN	MatchGAN	CISSL	SEC-GAN	TripleGAN	ECGAN
bloodmnist	0.925	**0.939**	0.173	0.649	0.556	0.913	0.917	0.311
breastmnist	0.698	**0.792**	0.646	0.667	0.708	0.750	0.688	0.646
chestmnist	0.563	0.574	0.525	0.562	0.574	0.565	**0.582**	0.532
dermamnist	0.652	**0.654**	0.111	0.623	0.623	0.612	0.627	0.622
octmnist	0.609	**0.667**	0.248	0.334	0.292	0.463	0.642	0.259
organamnist	0.880	**0.885**	0.161	0.098	0.441	0.822	0.843	0.134
organcmnist	0.859	**0.881**	0.081	0.784	0.423	0.771	0.809	0.171
organsmnist	0.724	**0.737**	0.064	0.573	0.212	0.587	0.654	0.181
pathmnist	0.894	**0.909**	0.095	0.335	0.439	0.649	0.801	0.084
pneumoniamnist	**0.823**	0.817	0.585	0.781	0.815	0.792	0.821	0.756
tissuemnist	0.450	**0.472**	0.052	0.302	0.326	0.440	0.470	0.075

why our method achieves strong performance, we investigate two potential factors that relate to the core mechanisms of our approach. First, we examine whether discriminative feature content, the degree to which classes exhibit visually distinguishable characteristics, affects the effectiveness of our image-to-image translation mechanism. Since the generator must learn meaningful class-conditional transformations, we hypothesize that performance may depend on the presence of clear visual differences between classes. Second, we investigate how unlabeled dataset size impacts semi-supervised learning performance. Given that our method relies on pseudo-labeling and translation of unlabeled samples during the unsupervised training phase, we explore whether the quantity of available unlabeled data influences the model's ability to extract additional information beyond the limited labeled samples.

Table A.4: Comparison of accuracy scores across different methods on medical image datasets using 50-shot learning. Best performance for each dataset is highlighted in bold.

Dataset	SPARSE	$\text{SPARSE}_{\text{ens}}$	SGAN	MatchGAN	CISSL	SEC-GAN	TripleGAN	ECGAN
bloodmnist	0.966	**0.968**	0.172	0.819	0.612	0.936	0.956	0.807
breastmnist	**0.844**	0.813	0.646	0.708	0.698	0.781	0.792	0.750
chestmnist	0.613	**0.616**	0.522	0.532	0.604	0.568	0.577	0.532
dermamnist	**0.708**	0.701	0.133	0.625	0.619	0.617	0.660	0.624
octmnist	0.648	0.683	0.241	0.278	0.339	0.643	**0.826**	0.247
organamnist	0.928	**0.930**	0.178	0.141	0.485	0.867	0.887	0.155
organcmnist	0.884	**0.888**	0.183	0.797	0.429	0.837	0.841	0.619
organsmnist	0.746	**0.763**	0.023	0.623	0.290	0.682	0.686	0.499
pathmnist	0.906	**0.917**	0.069	0.333	0.369	0.841	0.856	0.116
pneumoniamnist	0.813	**0.860**	0.783	0.798	0.850	0.842	0.821	0.815
tissuemnist	0.431	0.472	0.052	0.302	0.335	0.449	**0.485**	0.052

A.1.1 Discriminative Features and Image-to-Image Translation

We first investigate whether the presence of discriminative features, visual characteristics that distinguish between classes, affects SPARSE's performance, given that our method relies on image-to-image translation as its core semi-supervised learning mechanism. Unlike purely generative approaches that synthesize images from noise, our method modifies real unlabeled images to preserve authentic features while learning class-conditional transformations. We hypothesize that the generator can only learn meaningful class-conditional transformations when classes exhibit visually distinguishable characteristics. Our analysis of the results in Tables A.1–A.4 reveals a clear performance stratification based on discriminative feature strength. Datasets where classes exhibit distinct morphological or structural characteristics achieve superior performance: BloodMNIST (96.8% in 50-shot, Table A.4) contains 8 blood cell types with fundamentally different cellular structures and nuclear; PathMNIST (91.7%) comprises 9 tissue types with distinct histological patterns; OrganAMNIST (93.0%) represents 11 abdominal organs with distinct anatomical shapes and densities. In these cases, the generator appears to learn meaningful class-conditional transformations that emphasize distinctive characteristics, while the discriminator recognizes authentic class-specific features. Conversely, datasets with limited discriminative features show substantially reduced performance. TissueMNIST (47.2%) contains cells from the same anatomical region sharing similar morphologies; ChestMNIST (61.6%) presents subtle and overlapping disease manifestations. This pattern suggests that when biological similarity or imaging constraints limit discriminative features, image translation cannot create information absent in the original data. Datasets with intermediate performance present moderate discriminative features: DermaMNIST (70.1%) contains skin lesions with some overlapping visual characteristics;

OCTMNIST (68.3%) captures subtle retinal tissue changes that may not present dramatically distinct patterns; PneumoniaMNIST (86.0%) shows variable pneumonia manifestations that may not always differ obviously from normal chest X-rays. These findings suggest that discriminative feature strength is a primary determinant of SPARSE performance, with the image-to-image translation mechanism requiring visually distinguishable class characteristics to be effective.

A.1.2 Unlabeled Dataset Size Impact

We next investigate how unlabeled data quantity affects semi-supervised learning performance. The MedMNIST datasets vary significantly in training set size (546 to 165,466 samples, Table 3.1), providing a natural experiment for this analysis. Since SPARSE alternates between supervised and unsupervised phases, where the unsupervised phase leverages all available unlabeled samples for pseudo-labeling and image translation, we examine whether the size of the unlabeled dataset impacts how much additional information the model can extract beyond the limited labeled samples. Our analysis reveals that unlabeled dataset size alone does not guarantee performance. Examining large unlabeled datasets, we observe that TissueMNIST (165,466 training samples) achieves only 47.2% despite having the largest unlabeled dataset, while PathMNIST (89,996 samples) reaches 91.7%. This demonstrates that unlabeled data quantity cannot compensate for absent discriminative features. However, when discriminative features are present, we find that larger unlabeled datasets amplify performance gains. PathMNIST, OrganAMNIST (34,561 samples, 93.0%), and OCTMNIST (97,477 samples, 68.3%) all show substantial improvements from 5-shot to 50-shot settings (+13.9%, +12.9%, +21.8% respectively, comparing Tables A.1 and A.4), indicating effective utilization of unlabeled data for pseudo-labeling and translation. Conversely, datasets with smaller unlabeled sets demonstrate that strong discriminative features can compensate for limited unlabeled data. BloodMNIST (11,959 samples) achieves 96.8%, the highest accuracy among all datasets, while PneumoniaMNIST (4,708 samples) reaches 86.0%. Both show more modest 5-to-50-shot improvements (+8.1%, +5.4%), suggesting performance depends more on labeled sample quality and discriminative features than unlabeled data quantity when the unlabeled set is small. These findings reveal that the relationship between discriminative feature strength and unlabeled dataset size is multiplicative rather than additive. Strong discriminative features enable effective learning even with limited unlabeled data (BloodMNIST, PneumoniaMNIST), while weak discriminative features limit performance regardless of unlabeled data abundance (TissueMNIST, ChestMNIST). The optimal scenario combines both strong discriminative features and large unlabeled datasets (PathMNIST, OrganAM-

NIST), where SPARSE achieves performance exceeding 90%.