VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

INFORMATICS MASTER STUDIES

MASTER‘S THESIS

Evaluating and Enhancing Semi-Supervised Learning Algorithms for Pancreatic Cancer Segmentation in CT Images

Kasos vėžio segmentavimo KT vaizduose įvertinimas ir tobulinimas pusiau prižiūrimų mokymosi algoritmų

Author:	Md Istiaque Ahmed	(parašas)
Supervisor:	Prof. Dr. Olga Kurasova	(parašas)
Reviewer:	Vytautas Valaitis, Assist. Prof., Dr.	(parašas)

Vilnius-2025

Acknowledgements

I would like to express my deepest gratitude to my esteemed supervisor, Prof. Dr. Olga Kurasova, for her invaluable guidance, unwavering support, and insightful feedback throughout the research and preparation of this Master's thesis. Her expertise and encouragement were instrumental in navigating the complexities of this project and bringing it to fruition.

The author is also thankful for the access to high-performance computing (HPC) resources provided by the Information Technology Research Center of Vilnius University, which were essential for conducting the computationally intensive experiments presented in this work.

Santrauka

Pusiau prižiūrimas mokymasis (SSL) siūlo perspektyvų būdą spręsti žymėtųjų duomenų trūkumo medicininių vaizdų segmentavime problemą – tai itin svarbus iššūkis atliekant tokias užduotis kaip kasos vėžio nustatymas kompiuterinės tomografijos (KT) tyrimuose. Šiame magistro darbe vertinamas dviejų žinomų SSL algoritmų – „Mean Teacher" ir „MixMatch" – efektyvumas 2D kasos KT vaizdų segmentavimui. Tyrime lyginamas jų našumas su patikimu prižiūrimu „U-Net" pradiniu lygiu, kai jie apmokyti naudojant labai ribotą žymėtųjų duomenų rinkinį ir didesnį nežymėtųjų duomenų telkinį, gautą iš „Medical Segmentation Decathlon Pancreas-CT" duomenų rinkinio.

Eksperimentai parodė, kad nors prižiūrimas pradinis lygis po išsamių mokymų pasiekė gerą našumą su ribotais žymėtais duomenimis, nei „Mean Teacher", nei „MixMatch", esant išbandytoms konfigūracijoms, neparodė reikšmingo pagerėjimo, palyginti su šiuo pradiniu lygiu. Buvo nustatyti pagrindiniai iššūkiai, įskaitant mokytojo modelio nestabilumą „Mean Teacher" sistemoje ir prastos kokybės pseudožymių generavimą, dėl kurio sumažėjo našumas naudojant „MixMatch" metodą. Šios problemos buvo ypač ryškios esant mažai duomenų.

Šis tyrimas prisideda prie griežto šių SSL metodų įvertinimo sudėtingoje klinikinėje praktikoje, išryškinant praktinius modelio stabilumo ir pseudoetikečių patikimumo apribojimus, kai trūksta anotacijų. Išvados pabrėžia, kaip sudėtinga efektyviai panaudoti nepažymėtus duomenis šiai užduočiai atlikti, ir suteikia įžvalgų apie būsimą patikimesnių pusiau prižiūrimų medicininio vaizdavimo metodų kūrimą.

Raktiniai žodžiai: pusiau prižiūrimas mokymasis, kasos vėžys, KT vaizdų segmentavimas, vidurkio nustatymas mokytoju, „MixMatch", gilusis mokymasis, medicininis vaizdavimas, riboti žymėti duomenys.

Abstract

Semi-supervised learning (SSL) offers a promising avenue to address the scarcity of labeled data in medical image segmentation, a critical challenge for tasks like detecting pancreatic cancer from Computed Tomography (CT) scans. This Master's thesis evaluates the efficacy of two prominent SSL algorithms, Mean Teacher and MixMatch, for 2D pancreatic CT image segmentation. The study compares their performance against a robust supervised U-Net baseline when trained with a severely limited set of labeled data and a larger pool of unlabeled data derived from the Medical Segmentation Decathlon Pancreas-CT dataset.

Experiments revealed that while the supervised baseline achieved strong performance with limited labeled data after extensive training, neither Mean Teacher nor MixMatch, under the tested configurations, demonstrated a significant improvement over this baseline. Key challenges were identified, including teacher model instability in the Mean Teacher framework and the generation of poor-quality pseudo-labels leading to performance degradation in the MixMatch approach. These issues were particularly pronounced in the low-data regime.

This research contributes a rigorous evaluation of these SSL techniques in a challenging clinical application, highlighting practical limitations concerning model stability and pseudo-label reliability with scarce annotations. The findings underscore the complexities of effectively leveraging unlabeled data for this task and provide insights for future development of more robust semi-supervised methods for medical imaging.

Keywords: Semi-Supervised Learning, Pancreatic Cancer, CT Image Segmentation, Mean Teacher, MixMatch, Deep Learning, Medical Imaging, Limited Labeled Data.

Contents

1. Introduction

Machine learning, particularly deep learning, has demonstrated significant potential in transforming medical image analysis. However, many state-of-the-art models rely on supervised learning, which necessitates large, meticulously annotated datasets. In the medical domain, acquiring such labeled data is often a formidable challenge due to the high cost, time-consuming nature of annotation, and the requisite involvement of clinical experts [Pei+18]. This data scarcity bottleneck can impede the development and deployment of robust diagnostic and prognostic tools.

Semi-Supervised Learning (SSL) emerges as a compelling paradigm to address this limitation. Operating at the confluence of supervised and unsupervised learning, SSL techniques are designed to leverage information from both a small set of labeled examples and a typically much larger pool of unlabeled data [Wu+22]. This dual approach is particularly advantageous in fields like medical diagnostics, where unlabeled medical images are often abundant but labeled counterparts are scarce [RVR18]. By effectively utilizing unlabeled data, SSL aims to enhance model performance, improve generalization, and reduce the dependency on extensive manual annotation efforts.

1.1. The Challenge of Pancreatic Cancer and the Role of CT Imaging

Pancreatic cancer remains one of the most lethal malignancies worldwide, primarily due to its often late-stage diagnosis and consequently low five-year survival rates [MKT+18]. Early and accurate detection is paramount for improving patient prognosis and enabling effective treatment strategies. Computed Tomography (CT) imaging plays a crucial role in the diagnostic pathway for pancreatic cancer, providing detailed cross-sectional anatomical information that can reveal the presence, extent, and characteristics of tumors.

However, the manual or even automated segmentation of pancreatic tumors from CT scans is a notably challenging task due to several intrinsic factors. The pancreas itself is an anatomically complex organ, situated deep within the abdomen and often exhibiting indistinct boundaries with adjacent organs and soft tissues. Furthermore, pancreatic tumors can present with significant variability in their appearance, including shape, size, texture, and contrast enhancement patterns, which complicates the development of universally robust segmentation algorithms. Compounding these issues is the limited availability of large, high-quality, publicly annotated CT datasets specifically for pancreatic cancer segmentation, a consequence of the expert-driven and labor-intensive nature of medical image annotation [Jia+23].

1.2. Semi-Supervised Learning for Pancreatic Cancer Segmentation

Given the challenges in data acquisition and the critical need for improved diagnostic tools, SSL offers a promising avenue for advancing pancreatic cancer segmentation in CT images. By learning from a combination of limited labeled data and more abundant unlabeled CT scans, SSL algorithms have the potential to develop models that are more accurate and robust than those trained solely on small labeled datasets. Enhanced segmentation accuracy can directly contribute to earlier and more reliable tumor detection, precise localization for treatment planning (such as surgical resection or radiotherapy), and better monitoring of treatment response. Ultimately, improvements in segmentation facilitated by SSL could play a vital role in improving clinical outcomes for patients with pancreatic cancer [SCH+24].

1.3. Aim and Objectives of This Thesis

The primary aim of this thesis is to **evaluate and enhance semi-supervised learning algorithms for pancreatic cancer segmentation in CT images**, particularly within a context of limited labeled data. This work seeks to understand the efficacy and practical challenges of applying established SSL techniques to this specific medical imaging problem.

The key objectives of this thesis are:

1. To establish robust supervised learning baselines for 2D pancreatic CT segmentation using a U-Net architecture with varying amounts of labeled data from the Medical Segmentation Decathlon (MSD) Pancreas-CT dataset [ARB+22].

2. To implement and evaluate two prominent SSL algorithms, namely Mean Teacher [TV17] and MixMatch [BCG+19], for the pancreatic segmentation task, utilizing a small set of labeled data supplemented by a larger pool of unlabeled data.

3. To conduct a comparative analysis of the supervised and semi-supervised approaches, focusing on segmentation accuracy (Dice Similarity Coefficient), training dynamics, and model behavior.

4. To identify key challenges and limitations encountered during the application of these SSL methods and to discuss potential reasons for their observed performance.

5. To propose future research directions for enhancing the effectiveness of SSL techniques for this and similar medical image segmentation tasks based on the insights gained.

1.4. Thesis Outline

This thesis is organized as follows: Chapter 2 provides a review of relevant literature on supervised learning for medical image segmentation, the limitations thereof, recent advances in SSL for medical image analysis, and specific SSL algorithms pertinent to tumor segmentation. Chapter 3 details the dataset used, the image preprocessing and data augmentation pipeline, the evaluation framework, and the specific implementation details of the supervised baseline, Mean Teacher, and MixMatch algorithms. Chapter 4 presents the empirical results from all conducted experiments, including performance metrics and analysis of training dynamics for each approach. Chapter 5.1 offers a comprehensive discussion of the findings, including a comparative analysis, challenges encountered, limitations of the study, and detailed future research directions. Finally, Chapter 6 summarizes the key contributions and conclusions of this thesis.

2. Scientific Literature Review

The accurate segmentation of medical images is a cornerstone of modern diagnostics and treatment planning, particularly in oncology. For challenging cases like pancreatic cancer, precise delineation of tumors and affected tissues from Computed Tomography (CT) scans is crucial for staging, surgical planning, and monitoring therapeutic response. While supervised deep learning models, especially U-Net architectures [RFB15], have achieved remarkable success in medical image segmentation [AAR+24], their performance is heavily contingent on the availability of large, expertly annotated datasets. The creation of such datasets is notoriously time-consuming, expensive, and requires specialized clinical expertise, posing a significant bottleneck in translating these advanced models into widespread clinical practice [Pei+18].

Semi-Supervised Learning (SSL) has emerged as a promising paradigm to mitigate this data scarcity problem. SSL methods aim to effectively utilize information from both a limited set of labeled samples and a more extensive collection of unlabeled data, thereby enhancing model performance, robustness, and generalization capabilities without a proportional increase in annotation effort [Wu+22]. This chapter reviews foundational SSL techniques, discusses their application and evolution within medical imaging, and highlights specific advancements and considerations relevant to the segmentation of pancreatic cancer from CT images, setting the context for the experimental evaluations conducted in this thesis.

2.1. Core SSL Paradigms and Their Relevance in Medical Imaging

Several core SSL strategies have been widely explored, with pseudo-labeling and consistency regularization being particularly prominent in medical image segmentation contexts [Lee13; SJT16; YTW+23].

2.1.1. Pseudo-Labeling and Self-Training

Pseudo-labeling, often used interchangeably with self-training, is an intuitive yet powerful SSL technique. In this approach, a model initially trained on available labeled data is used to generate predictions (pseudo-labels) for unlabeled samples. High-confidence pseudo-labels are then treated as if they were true labels and are incorporated into the training set for subsequent model iterations [CTQ+21]. This iterative process allows the model to progressively learn from the unlabeled data distribution, effectively expanding its training data.

The primary advantage of pseudo-labeling lies in its simplicity and its ability to leverage large amounts of unlabeled data. However, its main challenge is the risk of error propagation: if the initial model generates incorrect pseudo-labels with high confidence, these errors can be reinforced in subsequent training rounds, potentially leading to performance degradation or confirmation bias. Therefore, the quality and reliability of the generated pseudo-labels are critical.

To address this, various enhancements have been proposed. Chaitanya et al.[CEK+23] introduced an approach for medical image segmentation that integrates a local contrastive loss with pseudo-label-based self-training (Figure 1). This method encourages similar feature representations for pixels sharing the same pseudo-label or ground truth label while promoting dissimilarity for pixels with different labels. By focusing on learning discriminative pixel-level features, this technique aims to improve the precision of segmentation, particularly for delineating fine structures or tumor boundaries. The application of such refined pseudo-labeling techniques is highly relevant to pancreatic cancer segmentation, where accurate boundary definition is crucial.

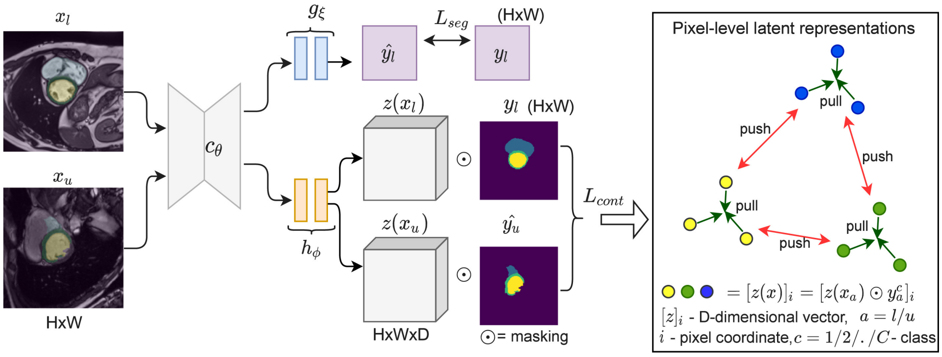
x_l
x_u
HxW
c_θ
g_ξ
h_ϕ
$\hat{y}_l$
L_{seg}
(HxW)
y_l
$z(x_l)$
$z(x_u)$
HxWxD
y_l (HxW)
$\hat{y}_u$
⊙
⊙
⊙= masking
L_{cont}
Pixel-level latent representations
pull
push
push
pull
pull
push
● ● ● $= [z(x)]_i = [z(x_a) \odot y_a^c]_i$
$[z]_i$ - D-dimensional vector, $a = l/u$
i - pixel coordinate, $c = 1/2/./C$ - class

Fig. 1. Conceptual overview of a semi-supervised method integrating pixel-level contrastive loss with pseudo-labels from unlabeled data and ground truth from limited labeled data, as proposed by Chaitanya et al. [CEK+23].

2.1.2. Consistency Regularization

Consistency regularization is a cornerstone of many successful SSL algorithms. It operates on the principle that a model's predictions should be invariant to small, semantics-preserving perturbations applied to its input or its internal structure. For an unlabeled sample, the model is encouraged to produce consistent outputs for different augmented versions of that sample, or consistent outputs between different model states (e.g., student vs. teacher models). This forces the decision boundary to lie in low-density regions of the input space, leading to smoother and more robust models that generalize better from limited labeled data [SJT16].

In medical imaging, where data can exhibit significant variability due to acquisition protocols, patient anatomy, and noise, consistency regularization is particularly valuable. For pancreatic cancer segmentation, it can help models become more resilient to variations in CT scan quality and subtle differences in tumor appearance. The Mean Teacher algorithm [TV17], which is evaluated in this thesis, is a prime example. It maintains two models: a student model that is trained conventionally, and a teacher model whose weights are an exponential moving average (EMA) of the student's weights. Consistency is enforced between the student's predictions on augmented unlabeled data and the more stable predictions from the teacher model on differently augmented versions of the same data. This EMA teacher provides more reliable targets than methods relying on the student's own past predictions. Enhancements often involve incorporating uncertainty estimation to weight the consistency loss, ensuring that the student learns more from confident teacher predictions [PBB+19; YYZ+21].

Lu et al. [LZY+23] (Figure 2) proposed a framework for semi-supervised medical image segmentation that refines consistency regularization by combining it with pseudo-labeling and dual consistency guided by uncertainty awareness. Their approach uses a cycle-loss mechanism for improved uncertainty estimation, leveraging these estimates to selectively apply pseudo-labels and enhance consistency between student and teacher networks (or multiple decoders within the student branch). Such sophisticated consistency mechanisms aim to improve the quality of learning from unlabeled data, which is critical for challenging tasks like pancreatic cancer segmentation where tumor boundaries can be ambiguous.

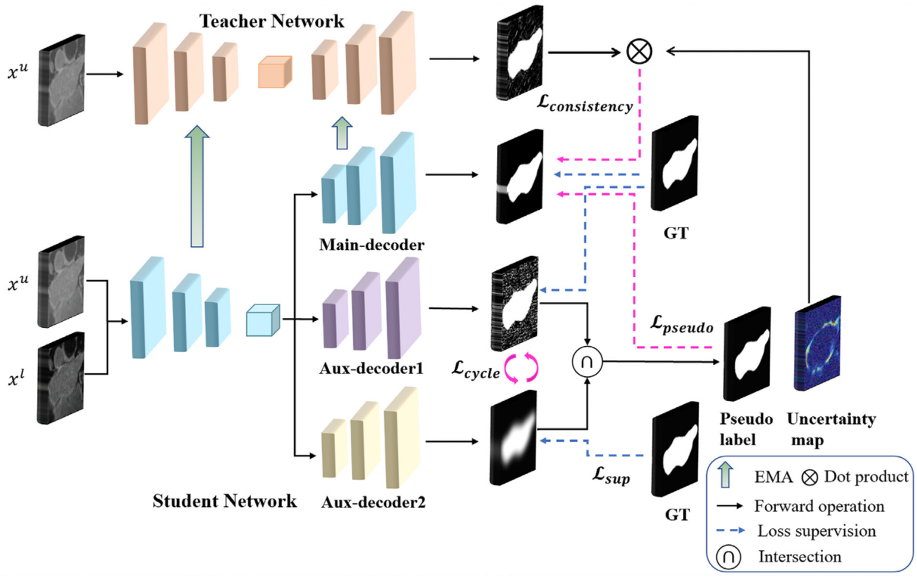

Fig. 2. Overview of the framework proposed by Lu et al. [LZY⁺23] for semi-supervised medical image segmentation, employing a variant of the Mean Teacher architecture with dual consistency regularization and uncertainty awareness.

2.1.3. Graph-Based SSL

Graph-based SSL techniques construct a graph where nodes represent data samples (labeled and unlabeled) and edges encode their relationships or similarities. Label information is then propagated from labeled nodes to unlabeled nodes through this graph structure [CDY⁺20]. Graph Convolutional Networks (GCNs) have extended this concept by enabling deep learning directly on graph-structured data, allowing for the learning of node representations that incorporate information from their neighborhood.

For medical image segmentation, GCNs can model complex spatial and contextual relationships within images. For instance, Liu et al. [LLH⁺22] proposed the Graph-Enhanced Pancreas Segmentation Network (GEPS-Net), which integrates a 3D U-Net with GCNs for pancreatic tumor segmentation from CT scans. This model leverages an uncertainty-guided iterative refinement strategy for pseudo-label generation. By explicitly modeling relationships between image regions or superpixels as a graph, GCNs can capture long-range dependencies and contextual cues that might be missed by standard CNNs operating on local receptive fields. While promising for tasks requiring understanding of intricate anatomical structures, graph-based methods can be computationally intensive and the construction of meaningful graphs from image data remains a key challenge. As this thesis focuses on consistency-based and pseudo-label/mixing approaches, graph-based SSL is noted here for completeness but not experimentally evaluated.

2.2. Advanced SSL Frameworks and Applications in Oncological Imaging

Building upon core SSL paradigms, more advanced frameworks have been developed, often combining multiple techniques or tailoring them for specific challenges in oncological imaging.

2.2.1. Adaptive and Hybrid Pseudo-Labeling Strategies

To improve the reliability of pseudo-labels, adaptive strategies have been explored. Zhang et al. [ZZH+23] introduced the BoostMIS framework (Figure 3), which synergizes adaptive pseudo-labeling with active learning for medical image segmentation. BoostMIS dynamically adjusts thresholds for accepting pseudo-labels based on model confidence and incorporates consistency regularization. Its active learning component identifies informative unlabeled samples characterized by low confidence or high uncertainty (estimated via virtual adversarial perturbation and density-aware entropy). These selected samples are then prioritized for expert annotation and integrated into the training set, creating a closed-loop system that iteratively refines the model by focusing annotation efforts where they are most needed. While demonstrated on MESCC and COVIDx datasets, the principle of adaptive pseudo-label refinement and targeted annotation is highly relevant for improving SSL in data-scarce scenarios like pancreatic cancer segmentation.

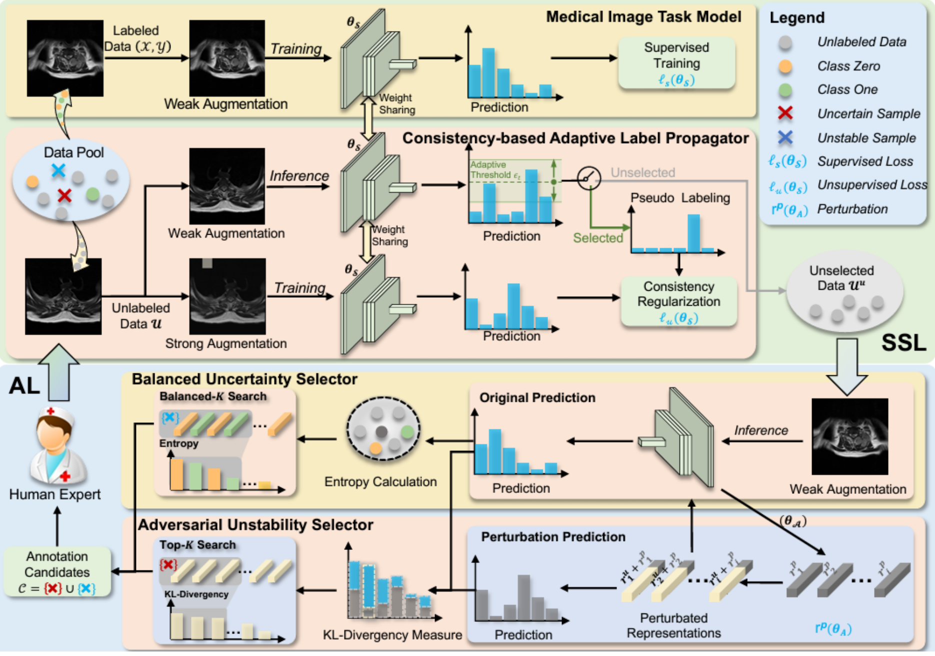

Fig. 3. Overview of the BoostMIS framework by Zhang et al. [ZZH+23], illustrating modules for task model training, consistency-based adaptive label propagation, and active learning via adversarial unstability and balanced uncertainty selection.

Another approach to enhance pseudo-labeling is reference-guided generation. Seibold et al. [SRK+22] proposed a method for semi-supervised semantic segmentation where a small set of labeled reference images guides the pseudo-labeling of unlabeled images by finding best-matching pixels or regions in the reference set. This technique showed strong performance, achieving results comparable to fully supervised models with substantially fewer labels on X-ray and retinal datasets. Adapting such exemplar-based strategies to the complex intensity profiles and textural variations in pancreatic CT scans could offer a way to generate more reliable pseudo-labels, although the computational cost of dense matching across reference sets needs careful consideration.

2.2.2. Robust Consistency and Auxiliary Task Learning

Enhancing consistency regularization often involves designing more robust perturbation strategies or incorporating auxiliary tasks to improve feature representation. The work by Lu et al. [LYF+23] (Figure 4), previously mentioned for its dual consistency, also highlights the use of uncertainty-aware consistency loss (via KL-divergence) to dynamically adapt to pseudo-label quality, particularly in multi-modal settings (e.g., CT and MRI). Their framework, evaluated on the NIH Pancreas-CT dataset [RFT+16], demonstrated the benefits of such robust consistency for improving segmentation accuracy.

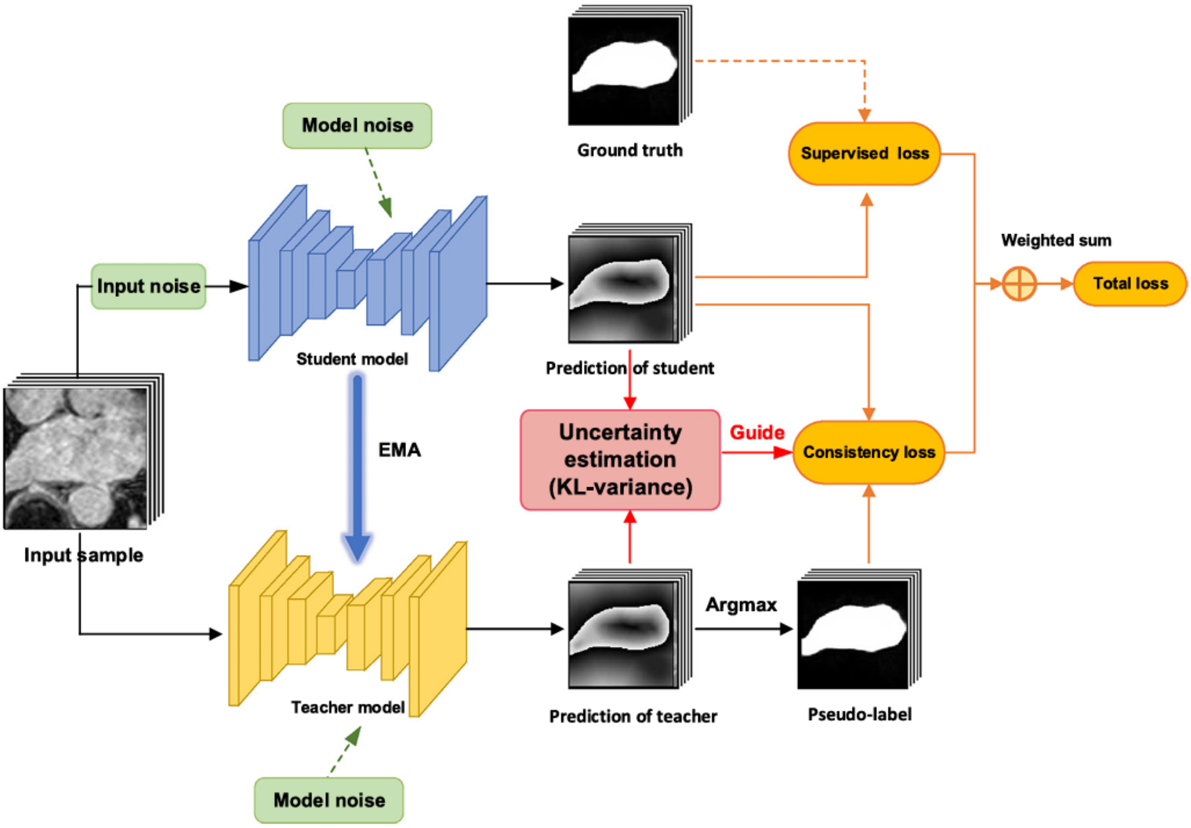

Fig. 4. Detailed overview of the uncertainty-aware pseudo-label and consistency framework proposed by Lu et al. [LYF+23], utilizing V-Net based student-teacher models and KL-divergence for semi-supervised medical image segmentation.

Auxiliary task learning can also regularize the shared encoder in an SSL framework, encouraging it to learn more generalizable features beneficial for the primary segmentation task. Myronenko [Myr19], in the context of brain tumor segmentation (BraTS 2018 challenge), augmented an encoder-decoder network with a variational autoencoder (VAE) branch. This VAE branch, sharing the encoder, reconstructed the input MRI scans, acting as a regularizer that helped the encoder learn robust features from limited data and mitigate overfitting, leading to state-of-the-art results. While not strictly SSL in the consistency/pseudo-labeling sense, the principle of using reconstruction or other self-supervised auxiliary tasks to improve the encoder's representations is a valuable strategy that can be combined with SSL approaches for tasks like pancreatic cancer segmentation.

2.2.3. SSL Specifically for Pancreatic Cancer Imaging

The unique challenges of pancreatic cancer imaging—such as the organ's complex anatomy, variable tumor appearance, subtle early signs, and often indistinct boundaries on CT—necessitate SSL approaches that are particularly robust and effective at extracting meaningful features from limited labeled data. While many general SSL techniques are applicable, their direct translation requires careful consideration of these domain-specific issues.

The studies by Lu et al. [LYF+23] using the NIH Pancreas-CT dataset and Liu et al. [LLH+22] with their GEPS-Net for pancreatic tumor segmentation directly address this application area, highlighting the potential of uncertainty-aware consistency and graph-based modeling, respectively. Shao et al. [SCH+24] developed a semi-supervised 3D segmentation framework specifically for pancreatic tumors using PET/CT images (Figure 5). Their MIM-CMFNet leverages mutual information minimization and cross-modal fusion, achieving a Dice coefficient of 73.14% on their pancreatic cancer dataset. This work underscores the benefit of multi-modal data and sophisticated feature fusion in SSL for this challenging anatomy, though it also notes the increased computational demands.

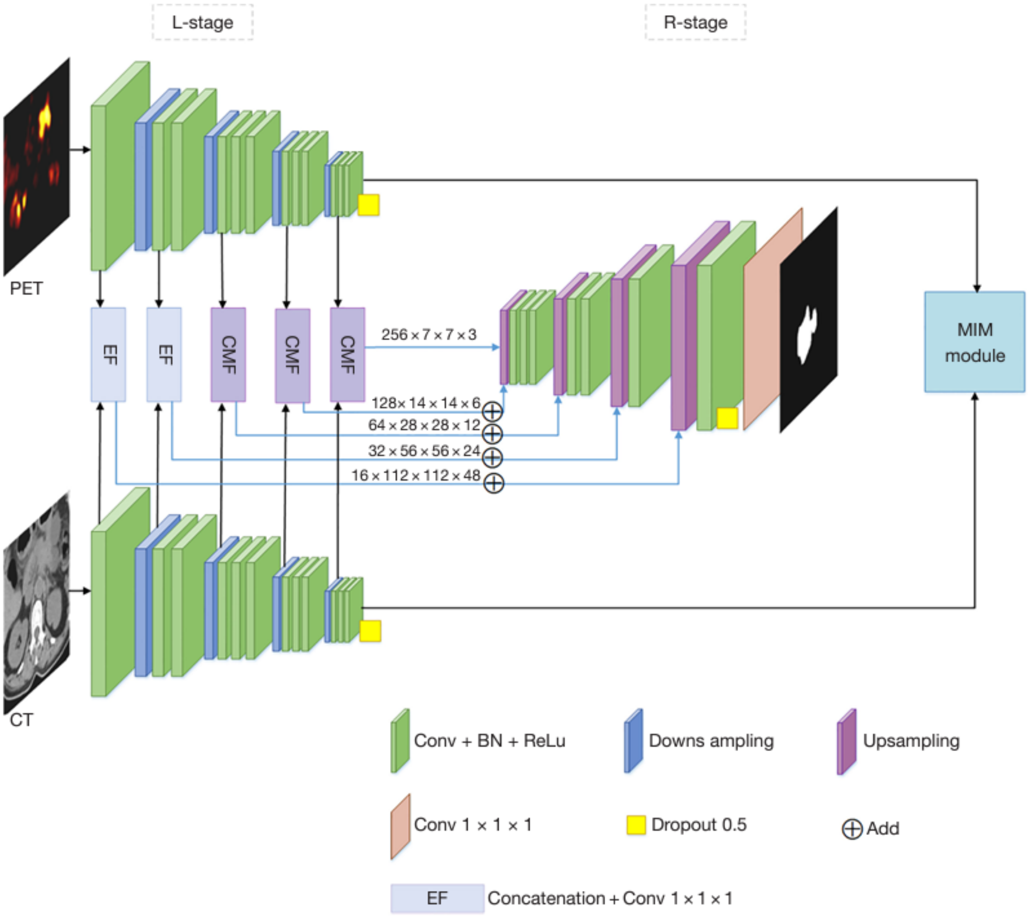

Fig. 5. Structure of the MIM-CMFNet for semi-supervised 3D segmentation of pancreatic tumors from PET/CT images, proposed by Shao et al. [SCH+24]. It incorporates modules for EasyFusion (EF), Cross-Modal Fusion (CMF), and Mutual Information Minimization (MIM).

The Mean Teacher [TV17] and MixMatch [BCG+19] algorithms, which form the core of the experimental investigation in this thesis, represent two distinct and influential SSL paradigms. Mean Teacher focuses on temporal ensembling and consistency between a student and an EMA-updated teacher, while MixMatch provides a holistic approach combining data augmentation, pseudo-label sharpening, and MixUp for consistency. Evaluating their performance and limitations on 2D pancreatic CT slice segmentation provides critical insights into their applicability and potential areas for enhancement in this specific, challenging context. The study by Kurasova et al. [KMŠ+23], which implemented a framework combining pseudo-labeling with consistency regularization for pancreatic cancer detection on CT scans (including data from Vilnius University Hospital Santaros Klinikos), further reinforces the interest in these hybrid SSL strategies for this clinical problem, reporting high F1 scores and accuracy.

2.3. Relevant Datasets for Pancreatic Cancer Imaging Research

The advancement of SSL algorithms for pancreatic cancer segmentation is critically dependent on the availability of suitable public datasets. Key resources include The Cancer Imaging Archive (TCIA) Pancreas-CT dataset [RFT+16], which provides contrast-enhanced CT scans from 82 patients with detailed annotations, making it valuable for SSL due to its mix of annotated and unannotated data potential. The Medical Segmentation Decathlon (MSD) - Task07 Pancreas dataset [ARB+22], utilized in this thesis (forming the basis of the `preprocessed` dataset), offers a larger cohort of 281 CT scans and has facilitated standardized evaluations. These datasets, among others, are instrumental for developing, validating, and comparing SSL techniques aimed at improving pancreatic cancer diagnosis and treatment planning.

2.4. Common Strategies for Optimizing and Enhancing SSL Algorithms

Beyond the core SSL paradigms, various strategies are employed in the literature to optimize their performance and enhance their robustness, particularly when applied to challenging medical imaging tasks. These often involve algorithmic refinements, meticulous hyperparameter tuning, specialized loss functions, and leveraging external knowledge through transfer learning.

2.4.1. Algorithmic Adjustments and Hybrid Approaches

Standard SSL algorithms are frequently adapted or combined to better address specific dataset characteristics or task requirements. As discussed previously, **adaptive pseudo-labeling** strategies, such as in BoostMIS [ZZH+23] (Section 2.2.1), refine the basic pseudo-labeling concept by dynamically adjusting label acceptance criteria, thereby improving label quality and mitigating error propagation.

Similarly, **enhanced consistency regularization** techniques seek to generate more potent consistency signals. This can involve designing domain-specific data augmentations or enforcing consistency across multiple scales or representations. For instance, Li et al. [DL24] proposed a curriculum consistency learning scheme (Figure 6) that enforces consistency between predictions derived from the full input image and those from dynamically cropped patches. This multi-scale approach guides the model to learn features robust to variations in object scale and context, which can be beneficial in semi-supervised medical image segmentation. The integration of graph-based methods with CNNs, as seen in GEPS-Net [LLH+22], also represents a hybrid approach to better capture spatial context.

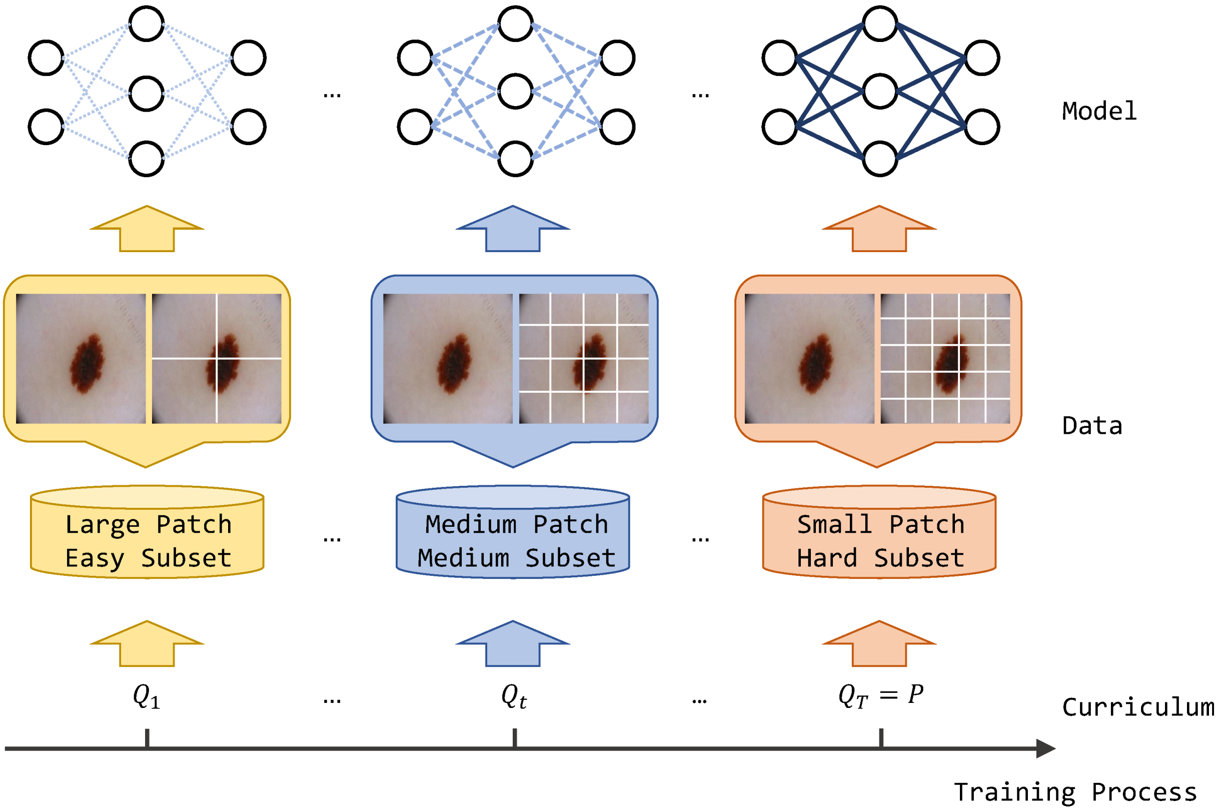

Fig. 6. Illustration of a curriculum consistency learning scheme, adapted from Li et al. [DL24], enforcing consistency between global image predictions and local patch predictions to guide semi-supervised learning.

2.4.2. Hyperparameter Optimization and Training Strategies

The performance of SSL algorithms is often sensitive to the choice of hyperparameters. **Learning rate optimization** is critical; while adaptive optimizers like Adam are common, advanced schedules (e.g., cosine decay with restarts) or methods like LipschitzLR [MCK⁺23; YSP21], which calculates learning rates based on the Lipschitz constant for faster and more stable training, are being explored. **Batch size selection** also impacts learning dynamics, with trade-offs between update stability (larger batches) and finer detail capture or resource constraints (smaller batches) [SK22].

Standard **regularization techniques** such as L2 regularization (weight decay) and dropout remain important in SSL to prevent overfitting to the limited labeled data and to improve generalization from unlabeled data [ZG23]. The choice of **optimizer** itself (e.g., Adam, SGD with momentum, RMSprop) can influence convergence and final performance. Rigorous hyperparameter selection, often guided by **k-fold cross-validation** on the available labeled data, is essential for finding optimal configurations, though this can be challenging with very small labeled sets [Ben12].

2.4.3. Loss Function Engineering

Customizing or weighting loss functions can significantly improve SSL performance by directing the model's attention to clinically relevant regions or by addressing class imbalance. For instance, in segmentation tasks, assigning higher weights in the loss calculation to tumor regions or difficult boundary areas can enhance the model's sensitivity and delineation accuracy. Specialized loss functions like the Tversky loss [SEG17] or boundary-focused losses [KBD+19] have been developed to better handle imbalances and improve performance on specific aspects of segmentation, and their integration into SSL frameworks is an active area of research.

2.4.4. Leveraging Transfer Learning

Transfer learning, where a model is pre-trained on a large, often general-purpose dataset (e.g., ImageNet for natural images, or a large diverse medical imaging dataset) and then fine-tuned on a smaller, task-specific dataset, is a powerful technique. In the context of SSL for medical imaging, pre-training can provide a strong feature extractor [KCS+22]. The SSL phase can then fine-tune this pre-trained model using a combination of limited task-specific labeled data and abundant unlabeled data from the target domain. This approach allows the model to benefit from generalized features learned from the source domain while adapting to the nuances of the specific medical task, potentially leading to better performance than training from scratch, especially when labeled data is extremely scarce.

The various strategies discussed for enhancing SSL performance underscore the complexity of optimizing these algorithms for specific applications like pancreatic cancer segmentation. The effectiveness of any given enhancement is often task-dependent and typically requires careful empirical evaluation, which forms a core motivation for the experimental work undertaken in this thesis.

3. Methodology

This chapter details the dataset utilized, the comprehensive image preprocessing pipeline, data augmentation strategies, the experimental framework including data partitioning and evaluation metrics, and the common model architecture. Subsequently, it elaborates on the specific implementations of the supervised baseline, Mean Teacher, and MixMatch algorithms, including their distinct configurations and training protocols.

3.1. Dataset Description and Challenges

This study utilizes the publicly available Medical Segmentation Decathlon (MSD) Pancreas-CT dataset [ARB+22]. This dataset comprises 281 contrast-enhanced abdominal CT scans from multiple institutions, representing a diverse patient population. Each CT scan consists of a series of axial slices with varying in-plane resolution and slice thicknesses (typically 1.5–2.5 mm). The primary task is the binary segmentation of pancreatic tissue (including tumors, consolidated as a single positive class) against the background. Pancreatic segmentation from CT images presents inherent challenges (Table 1), including the organ's small and variable size, anatomical variability, poorly defined boundaries, pathological alterations, and image artifacts [JSD24; RFL+15].

Table 1. Inherent Challenges in Pancreas Segmentation from CT Images.

Challenge	**Description**
Small and Variable Size	The pancreas typically occupies a small fraction (e.g., 1–2%) of the abdominal cavity volume, with considerable inter-patient variation in size and morphology.
Anatomical Variability	Significant differences in pancreatic shape, orientation, and position relative to other abdominal organs exist across individuals.
Poorly Defined Boundaries	The pancreas is often isodense with adjacent soft tissues (e.g., duodenum, spleen, blood vessels) and surrounding adipose tissue, making its boundaries ambiguous on CT imaging.
Pathological Alterations	The presence of tumors, cysts, or inflammation can further distort the normal pancreatic anatomy and alter tissue appearance, complicating segmentation.
Image Artifacts	CT scans can be susceptible to motion artifacts, beam hardening, and noise, which can degrade image quality and obscure tumor boundaries.

3.2. Image Preprocessing and Data Augmentation

3.2.1. Standardized Preprocessing Pipeline

A standardized preprocessing pipeline was applied to all 3D CT volumes from the MSD dataset to ensure data consistency and prepare inputs for the 2D segmentation models. This pipeline, implemented in Python (using NiBabel, SimpleITK, NumPy, TensorFlow), involved:

1. **Orientation and Voxel Spacing Standardization**: Each NIfTI volume was reoriented to a canonical RAS (Right-Anterior-Superior) coordinate system and resampled to a uniform isotropic in-plane voxel spacing of 1.0×1.0 mm and an axial slice thickness of 2.5 mm, using B-spline interpolation for images and nearest-neighbor for masks.

2. **Intensity Normalization**: Pixel intensities within each 3D volume were clipped to the 1st and 99th percentiles and then linearly scaled to the range [0, 1].

3. **Slice Extraction and 2D Formatting**: After 3D normalization and resampling, each processed 3D patient CT volume (image and corresponding segmentation mask) was saved as a distinct NumPy (`.npy`) file. The image data within each file was stored as a stack of 2D axial slices with dimensions $(D \times H \times W)$, where D represents the number of slices for that specific patient case, and H and W are the height and width, uniformly 256×256 pixels after this preprocessing stage. The corresponding binary segmentation mask for each patient case was saved in an identical $(D \times 256 \times 256)$ format. During model training and validation, the data loading pipeline is responsible for reading these per-patient `.npy` files and yielding individual $256 \times 256 \times 1$ 2D slices for input to the segmentation models.

This processed dataset, consisting of per-patient `.npy` files containing stacks of 2D slices, is referred to as the `preprocessed` dataset throughout this thesis.

3.2.2. Data Augmentation Strategies

Various data augmentation techniques were applied per 2D slice during training:

- **Labeled Data Augmentation:** Geometric augmentations were applied identically to images and masks.

- **Unlabeled Data Augmentation (for SSL):** Differing policies (weak vs. strong) generated distinct views for Mean Teacher or pseudo-label generation in MixMatch.

The common augmentation pipeline included:

1. **Geometric Transformations**: Random horizontal/vertical flips (50% probability each) and random 90-degree rotations. For SSL distinct views, unique random seeds were used.

2. **Intensity and Noise Perturbations**:

   - *Weak Augmentation*: Random brightness (±10%) and contrast (factor [0.9, 1.1]).

- *Strong Augmentation*: More aggressive brightness (±20%), contrast (factor [0.8, 1.2]), Gaussian noise (mean 0, stddev 0.05), and Cutout (25% patch, 30% probability).

All augmented pixel values were clipped to [0, 1] [SK19; WLC+22].

3.3. Experimental Framework and Evaluation

3.3.1. Ground Truth Data

Binary segmentation masks from the MSD Pancreas-CT dataset, delineating pancreatic tissue (including tumors) as 1 and background as 0, served as ground truth.

3.3.2. Experimental Design and Data Partitioning

For all experiments focusing on the evaluation of semi-supervised learning in a limited data regime, the `preprocessed` dataset (derived from 281 unique patient CT scans as described in Section 3.1) was partitioned based on patient cases. A fixed random seed of 42 was used for all splitting procedures to ensure reproducibility. The partitioning resulted in the following sets, with each patient case contributing multiple 2D axial slices after preprocessing (Section 3.2):

- **Labeled Training Set (L):** Comprised 50 distinct patient cases. Assuming an average of approximately 102 processed 2D slices per patient case, this set provided roughly $50 \times 65 = 5100$ individual labeled 2D slices for direct supervised training signals.
- **Labeled Validation Set (V):** Consisted of 30 distinct patient cases, yielding approximately $30 \times 65 = 3060$ individual 2D slices used for model validation during training.
- **Unlabeled Training Pool (U):** Contained the remaining 201 distinct patient cases (calculated as $281 - 50 - 30 = 201$). This pool provided approximately $201 \times 65 = 20{,}502$ individual 2D slices available as unlabeled data for the semi-supervised learning frameworks.

A direct supervised baseline was established by training a model exclusively on all 2D slices derived from the 50 Labeled Training Set patient cases and validating it on all 2D slices from the 30 Validation Set patient cases. This ensures a fair comparison for assessing the benefits of leveraging the Unlabeled Training Pool via SSL techniques. The test set defined in the original MSD challenge (139 cases) was held out and not used during the model development or validation phases reported in this thesis, being reserved for potential future final model evaluation.

3.3.3. Evaluation Metrics

The performance of the segmentation models was quantitatively assessed using standard metrics widely adopted in the field of medical image segmentation. These include the Dice Similarity Coefficient, Precision, and Recall.

1. **Dice Similarity Coefficient (DSC):** The DSC is a common metric for evaluating the spatial overlap between two segmentations. It measures the similarity between the predicted segmentation mask (X) and the ground truth segmentation mask (Y) [Dic45]. The DSC is calculated as:

$$\text{DSC} = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{1}$$

   where:

   - X represents the set of pixels belonging to the predicted segmentation mask (e.g., predicted pancreatic tissue).
   - Y represents the set of pixels belonging to the ground truth segmentation mask (e.g., actual pancreatic tissue).
   - $|X \cap Y|$ denotes the number of pixels common to both the predicted and ground truth masks (i.e., the true positives for the foreground class).
   - $|X|$ denotes the total number of pixels in the predicted segmentation mask.
   - $|Y|$ denotes the total number of pixels in the ground truth segmentation mask.

   A DSC value ranges from 0 (indicating no overlap between the prediction and ground truth) to 1 (indicating perfect overlap). Higher DSC values signify better segmentation accuracy.

2. **Precision and Recall:** Precision and Recall offer complementary insights into the model's classification performance for the foreground class (pancreatic tissue), particularly its ability to correctly identify positive pixels while minimizing false classifications [MSK22]. They are defined based on the concepts of True Positives (TP), False Positives (FP), and False Negatives (FN):

   - **True Positives (TP):** The number of pixels correctly identified by the model as belonging to the pancreatic tissue (i.e., pixels that are part of the pancreas in both the ground truth and the prediction).
   - **False Positives (FP):** The number of pixels incorrectly identified by the model as pancreatic tissue, which are actually background pixels in the ground truth (also known as Type I error).
   - **False Negatives (FN):** The number of pixels belonging to the pancreatic tissue in the ground truth that the model failed to identify (i.e., pixels missed by the model, also known as Type II error).

   Using these definitions, Precision and Recall are calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

Precision measures the proportion of correctly predicted positive pixels among all pixels predicted as positive by the model (i.e., of all pixels the model called ”pancreas,” how many actually were pancreas?). High precision indicates a low false positive rate. Recall (also known as sensitivity or true positive rate) measures the proportion of actual positive pixels that were correctly identified by the model (i.e., of all actual pancreas pixels, how many did the model find?). High recall indicates a low false negative rate.

3.3.4. Common Model Architecture and Training Protocol

- **Model Architecture:** All models used the `PancreasSeg` U-Net class (Figure 7), with 4 encoder-decoder levels, filter doubling (32 to 512 at bridge), halving in decoder, and skip connections.

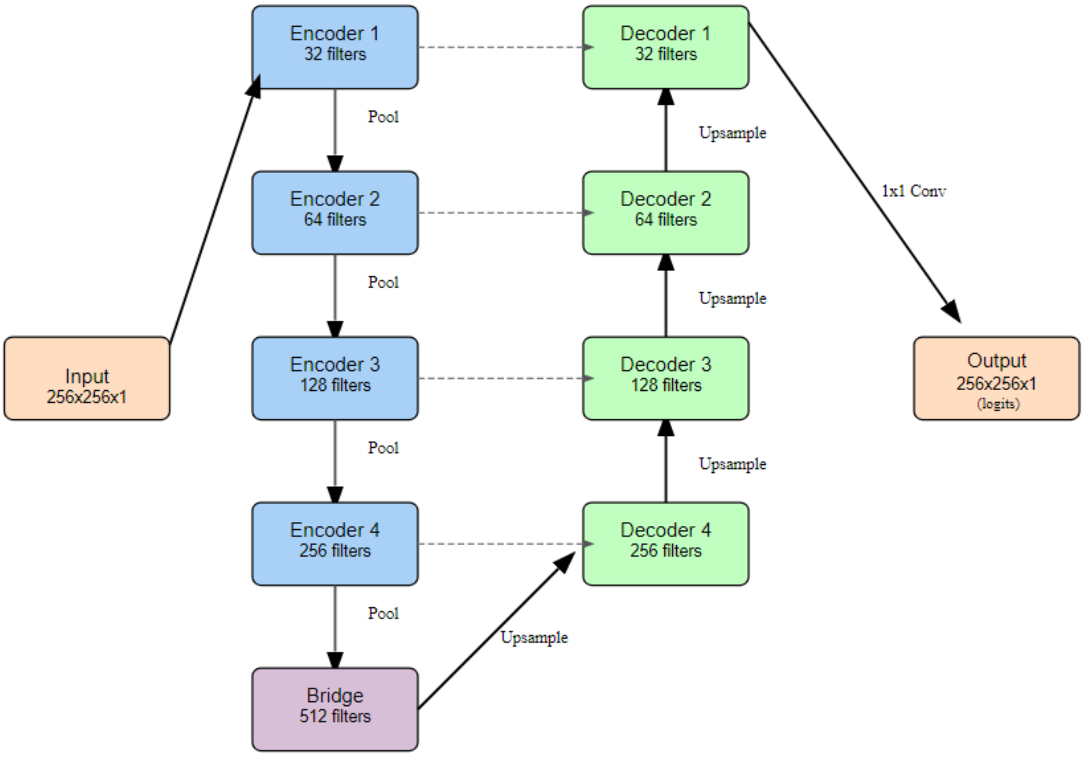

Fig. 7. Schematic of the U-Net architecture (`PancreasSeg`) employed. Input/output dimensions are $256 \times 256 \times 1$ (logits). Filter details are as described in the text.

- **Optimization:** Adam optimizer (`tf.keras.optimizers.Adam`) was used.

- **Performance Monitoring and Model Selection:** Validation DSC was monitored post-epoch. Early stopping terminated training if validation DSC did not improve for a specified patience (detailed per experiment). Best DSC epoch weights were saved.

3.4. Implementation of Supervised and Semi-Supervised Learning Algorithms

3.4.1. Supervised Learning Baseline Implementation

The primary supervised baseline, designed for direct comparison with the SSL methods, utilized all constituent 2D slices derived from the 50 designated labeled patient cases (from the `preprocessed` dataset) for training. Similarly, all 2D slices from the 30 distinct patient cases allocated to the validation set were used for performance monitoring.

The `PancreasSeg` model, incorporating a dropout rate of 0.1, was employed. Training was conducted using the Adam optimizer with an initial learning rate of 1×10^{-4} and the `DiceBCELoss` function. The model was trained for a target of 150 epochs, with early stopping triggered if the validation DSC did not improve for 25 epochs. A ReduceLROnPlateau learning rate scheduler was also active. The batch size for training was 8. Steps per epoch were determined by one full pass through all available training 2D slices (i.e., all slices from the 50 labeled patient cases).

3.5. Mean Teacher Method Implementation

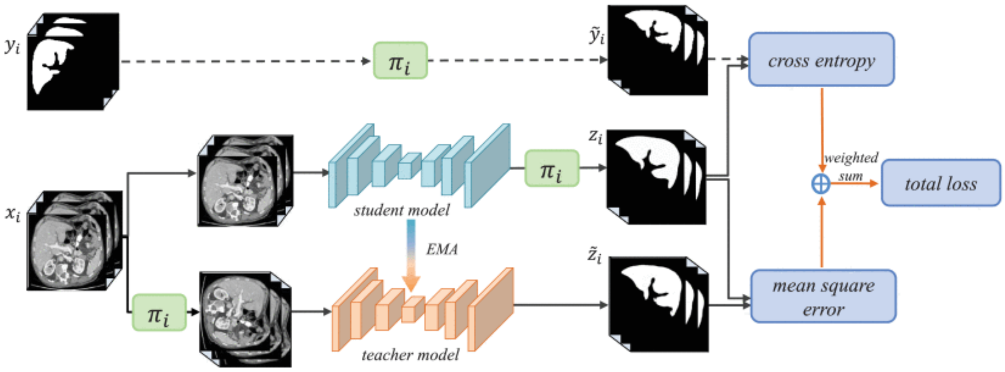

Fig. 8. Conceptual diagram of the Mean Teacher model application, adapted from [LYC+21].

The Mean Teacher SSL algorithm [TV17] was implemented as follows:

1. **Student-Teacher Architecture:** Identical `PancreasSeg` U-Nets. Student trained with supervised and consistency loss; teacher weights are EMA of student's. Dropout $p = 0.1$ was applied to student's encoder/bridge during training.

2. **Teacher Model Update Mechanism:**

   - **Initial Direct Copy Phase:** For $N_{copy} = 10$ epochs, teacher weights explicitly synchronized with student's at epoch start.
   - **EMA Phase:** Subsequently, and per-batch within copy phase, teacher weights updated with EMA decay $\alpha = 0.999$. Batch Normalization statistics copied directly.

3. **Loss Functions:** $L_{total} = L_S + \lambda_C(t) \cdot L_C$.

- **Supervised Loss (L_S):** DiceBCELoss ($w_{dice} = 0.5, w_{bce} = 0.5$).
- **Consistency Loss (L_C):** MSE between student's predictions on strongly augmented unlabeled views (x_{uS}) and sharpened teacher's predictions on weakly augmented views (x_{uT}). Sharpening used $T_{sharp} = 0.5$, logits clipped to [-15, 15].

4. **Consistency Weight Schedule ($\lambda_C(t)$):** Linearly ramped from 0 to $\lambda_{C,max} = 10.0$ over $E_{ramp} = 50$ epochs, starting after N_{copy} epochs.

5. **Training Configuration and Hyperparameters (Mean Teacher Experimental Run):** The student model was initialized from a 20-epoch supervised pre-training phase, which used all 2D slices from the 50 designated labeled patient cases (achieving a validation DSC of 0.7223 on slices from the 30 validation cases). The subsequent Mean Teacher SSL phase also utilized all 2D slices derived from the 50 Labeled (L), 30 Validation (V), and 201 Unlabeled (U) patient cases. This phase targeted 50 epochs. Key hyperparameters, consistent with the experimental run detailed in Section 4.2.2 included: Adam optimizer with a learning rate of 2×10^{-5}; batch processing involving 4 labeled and 4 unlabeled 2D slices per conceptual step; student dropout rate of 0.2; random seed 42. Early stopping monitored the validation DSC with a patience of 30 epochs. The number of steps per epoch was calculated as $\lceil$(Total 2D slices from 50L cases)$/4\rceil$. Other SSL-specific parameters such as `TEACHER_DIRECT_COPY_EPOCHS` (10), `BASE_EMA_DECAY` (0.999), `SHARPENING_TEMPERATURE` (0.5), `CONSISTENCY_MAX` (10.0), `CONSISTENCY_RAMPUP` (50 epochs), and the `DELAY_CONSISTENCY_FLAG` were set as defined in the corresponding execution script.

3.6. MixMatch Method Implementation

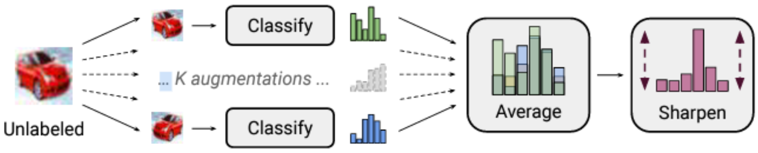

Fig. 9. Conceptual diagram of the label guessing process in MixMatch, adapted from [BCG+19].

3.6.1. MixMatch Framework.

The MixMatch SSL algorithm [BCG+19] was implemented as follows:

1. **Pseudo-Label Generation for Unlabeled Data:** For each x_u: $K = 2$ weakly augmented views averaged, then sharpened with $T = 0.5$.

2. **Data Augmentation and MixUp:** Labeled (X_l) and unlabeled (X_u) data strongly augmented. MixUp $\alpha_{mixup} = 0.75$ applied separately: X_l with P_l (ground truth), X_u with Q_u (sharpened pseudo-labels).

3. **Loss Functions:** $L = L_S + \lambda_u(t) \cdot L_U$.

   - **Supervised Loss (L_S):** BCE with logits on mixed labeled data.

   - **Unsupervised Consistency Loss (L_U):** MSE on mixed unlabeled data.

4. **Consistency Weight ($\lambda_u(t)$):** Linearly ramped from 0 to $\lambda_{u,max} = 25.0$ over $S_{ramp} \approx 525$ steps.

5. **Training Configuration (MixMatch Experimental Run):** The student model was initialized from the same 20-epoch supervised pre-training on all 2D slices from the 50 labeled patient cases (validation DSC: 0.7223). The MixMatch SSL phase utilized all 2D slices from the 50L/30V/201U patient case split and targeted 150 epochs. Key hyperparameters, consistent with the experimental run detailed in Section 4.3.1, included: Adam optimizer with an initial learning rate of 5×10^{-4} and a cosine decay schedule. The batch size for both the labeled and unlabeled data iterators was 8. Assuming approximately [Total L Slices, e.g., 2000] labeled 2D slices from the 50 patient cases, this resulted in `steps_per_epoch` of $\lceil$[Total L Slices]$/8\rceil$ $=$ [Calculated Steps, e.g., 250]. Student dropout rate was 0.1. Early stopping (patience 30 epochs was active. The random seed was 42. Other MixMatch-specific parameters such as `MIXMATCH_K` (2), `MIXMATCH_T` (0.5), `MIXMATCH_ALPHA` (0.75), `MIXMATCH_CONSISTENCY_MAX` (25.0), and `MIXMATCH_RAMPUP_STEPS` ($\sim$525 steps, calculated based on 75 effective epochs) were set as defined in the corresponding execution script.

This implementation strategy aims to harness the core strengths of MixMatch while adapting it for 2D medical image segmentation.

4. Experimental Results and Analysis

This section presents the empirical evaluation of the supervised learning baseline and two semi-supervised learning (SSL) algorithms—Mean Teacher and MixMatch—for the task of 2D pancreatic segmentation from CT slices. All experiments were conducted using the consistently `preprocessed` dataset, where each patient file corresponds to a single $256 \times 256 \times 1$ 2D axial slice, as detailed in Section 3.2.1. Model performance was primarily assessed using the Dice Similarity Coefficient (DSC) on the designated validation set of 30 slices. The underlying model architecture for all experiments was a U-Net [RFB15] (the `PancreasSeg` model, detailed in Section 3.3.4) incorporating a dropout rate of 0.1. The Adam optimizer was utilized for all training procedures.

4.1. Supervised Learning Baseline Performance with Limited Labeled Data

To establish a robust performance benchmark directly comparable to the semi-supervised learning approaches utilizing limited labeled data, a supervised learning experiment was conducted using 50 labeled 2D slices from the `preprocessed` dataset for training and 30 slices for validation.

4.1.1. Experimental Configuration for Supervised Baseline (50 Labeled Slices)

The `PancreasSeg` model was trained using 50 patient files (each a $256 \times 256 \times 1$ 2D slice) for the training set and 30 distinct patient files for the validation set. The model architecture incorporated a dropout rate of 0.1. Training employed the Adam optimizer with an initial learning rate of 1×10^{-4} and the `DiceBCELoss` function. Standard data augmentation techniques (Section 3.2.2) were applied. The model was trained for a target of 150 epochs, with early stopping (patience of 25 epochs on validation DSC) and a ReduceLROnPlateau learning rate scheduler.

4.1.2. Performance and Training Dynamics (50 Labeled Slices)

The training dynamics for this supervised baseline are illustrated in Figure 10.

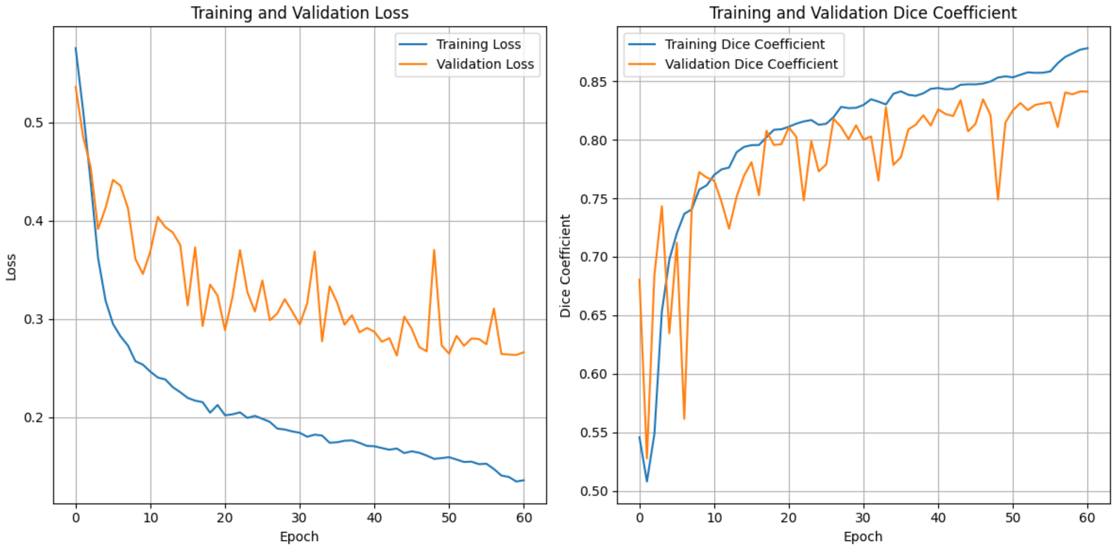

Fig. 10. Supervised learning training metrics using 50 labeled slices from the `preprocessed` dataset. Left: Training and Validation Loss. Right: Training and Validation Dice Coefficient. Training proceeded for approximately 61 epochs before early stopping (confirm exact epoch from logs).

The training loss (Figure 10, left panel, blue line) demonstrated a consistent decrease from approximately 0.58 to about 0.15 by epoch 61. The validation loss (orange line) initially decreased from 0.55 but plateaued around 0.27–0.30 after epoch 20–25, exhibiting increased volatility and indicating the onset of overfitting. The training Dice score (right panel, blue line) steadily rose from ∼0.52 to ∼0.87. The validation Dice score (orange line) showed rapid initial improvement from below 0.55 to over 0.80 by epoch 15–20, eventually reaching a peak validation Dice score of **0.8453 at epoch 58** (value and epoch from experimental logs). Training was terminated by early stopping after 61 epochs. This DSC of 0.8453 serves as the primary supervised benchmark for evaluating SSL methods using the same 50 labeled slices.

A separate, shorter 20-epoch supervised run on these same 50 labeled slices (serving as pre-training for SSL models) achieved a peak validation DSC of 0.7223. This highlights that extended training, even with limited data, can significantly improve supervised performance up to a point.

4.2. Semi-Supervised Learning: Mean Teacher

The Mean Teacher algorithm [TV17] was investigated for its potential to improve pancreatic segmentation by leveraging unlabeled data. The core methodological details, including the U-Net architecture, phased teacher updates, and consistency mechanisms, are described in Section 3.5. This section details an iterative experimental process undertaken to evaluate and understand the behavior of the Mean Teacher framework under conditions of limited labeled data (50 patient cases for labeled training, 30 for validation, and 201 for unlabeled data, unless otherwise specified for a particular experiment). All student models for the Mean Teacher phase were initialized from a 20-epoch supervised pre-training on the 50 labeled patient cases, which achieved a peak validation DSC of 0.7223.

4.2.1. Initial Exploration and Emergence of Teacher Model Instability

Initial experiments with the Mean Teacher framework, employing standard configurations (e.g., moderate sharpening temperature, significant consistency weight ramp-up, and a 15-epoch direct teacher-student copy phase followed by EMA updates with $\alpha = 0.999$), revealed a critical challenge early on: the instability of the teacher model.

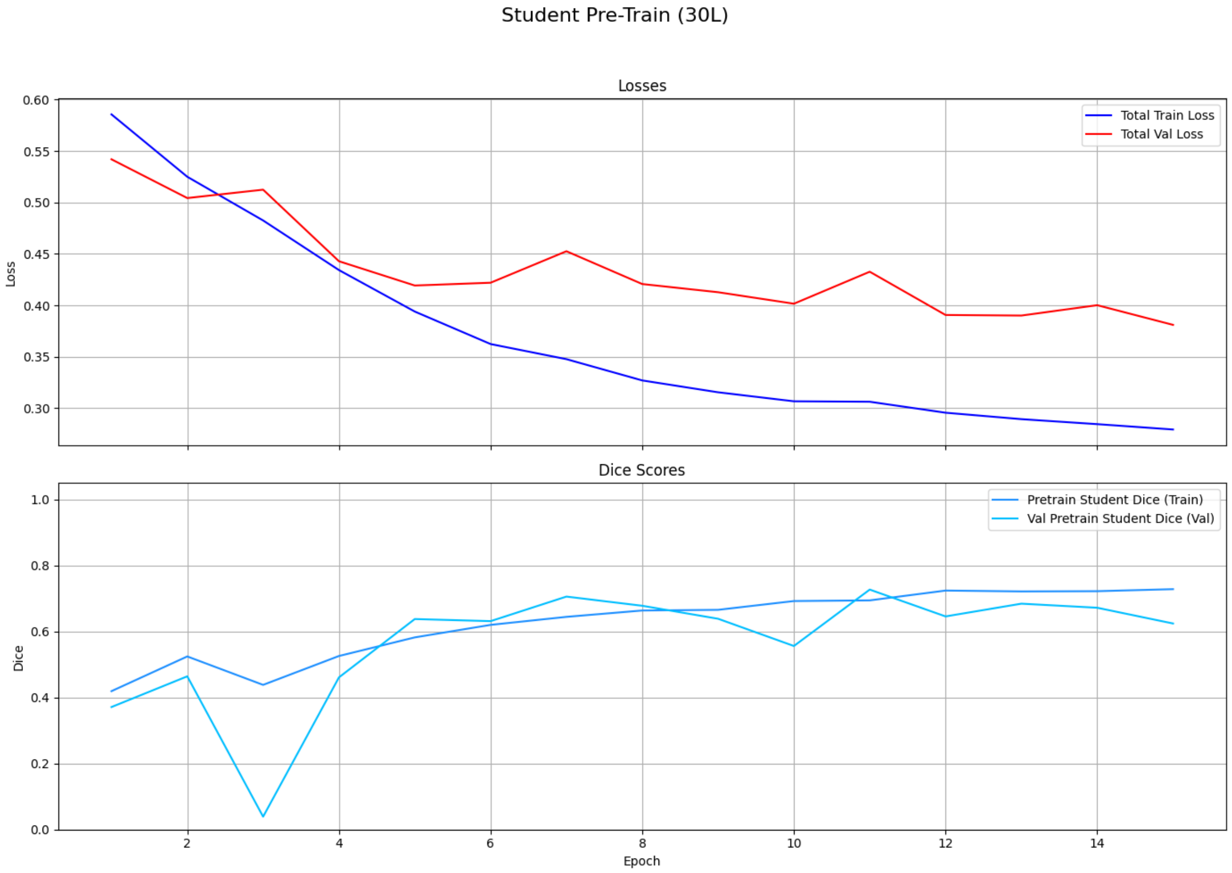

Fig. 11. Early Mean Teacher experiment, Student pre-train with 15 epoch and 30 labeled dataset

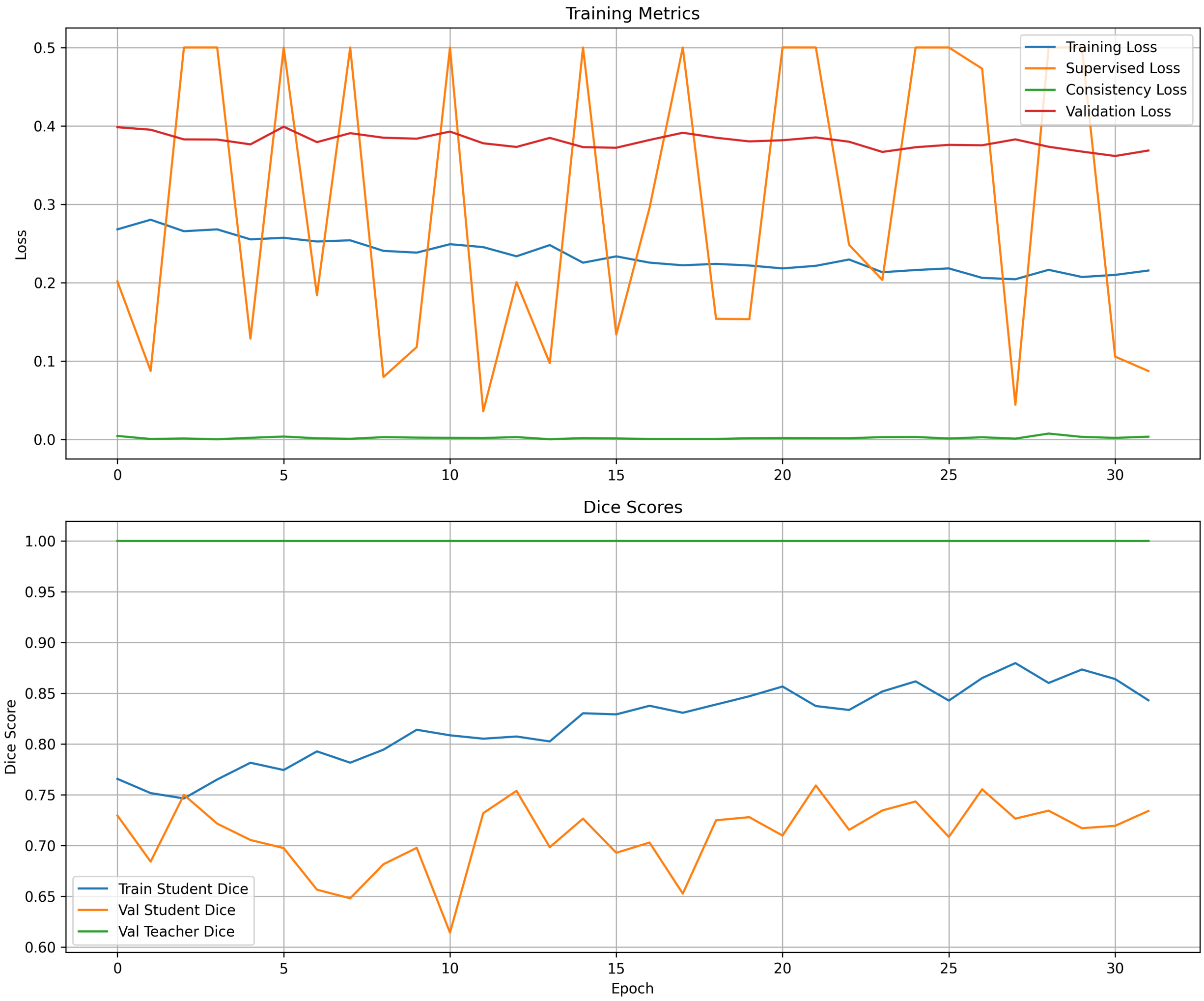

Fig. 12. Early Mean Teacher experiment showing rapid teacher model validation DSC artifact.

Across several initial configurations, a common pattern emerged where the student model, benefiting from pre-training, would exhibit reasonable initial validation performance. However, the teacher model's validation Dice score would often rapidly converge to an artifactual 1.0, particularly after the direct copy phase concluded and EMA updates began. Concurrently, the calculated consistency loss between student and teacher predictions on unlabeled data would remain very low, suggesting that either both models were converging to similar (potentially trivial, e.g., background-dominant) predictions, or the teacher was not providing a useful learning signal. This behavior indicated that the teacher model was failing to maintain robust foreground segmentation capabilities and was likely collapsing to a state of predicting predominantly background, with the 1.0 DSC arising from evaluation on background-only validation slices or slices with minimal, incorrectly predicted foreground. This observation necessitated a more detailed investigation into the interplay of hyperparameters and the teacher update mechanism.

4.2.2. Investigating Mean Teacher with Phased Updates and Tuned SSL Parameters (50 Labeled Cases)

Building on initial observations, a key experiment was conducted using 50 labeled patient cases, with specific attention to the teacher update strategy and SSL hyperparameter settings.

4.2.3. Experimental Configuration

The student model was initialized from the 20-epoch supervised pre-training (validation DSC: 0.7223). The Mean Teacher phase utilized slices from 50L/30V/201U patient cases and was targeted for 50 epochs. Key hyperparameters included: a 10-epoch direct teacher copy phase, EMA decay $\alpha = 0.999$, sharpening temperature $T_{sharp} = 0.5$, and a consistency loss weight $\lambda_C(t)$ ramping to a maximum of 10.0 over 50 epochs , with the ramp-up delayed until after the copy phase . The Adam optimizer with a learning rate of 2×10^{-5} and student dropout of 0.2 were employed. Early stopping (patience 30 epochs) was active. The random seed was 42.

4.2.4. Performance and Training Dynamics

The learning curves for this experiment are presented in Figure 13.

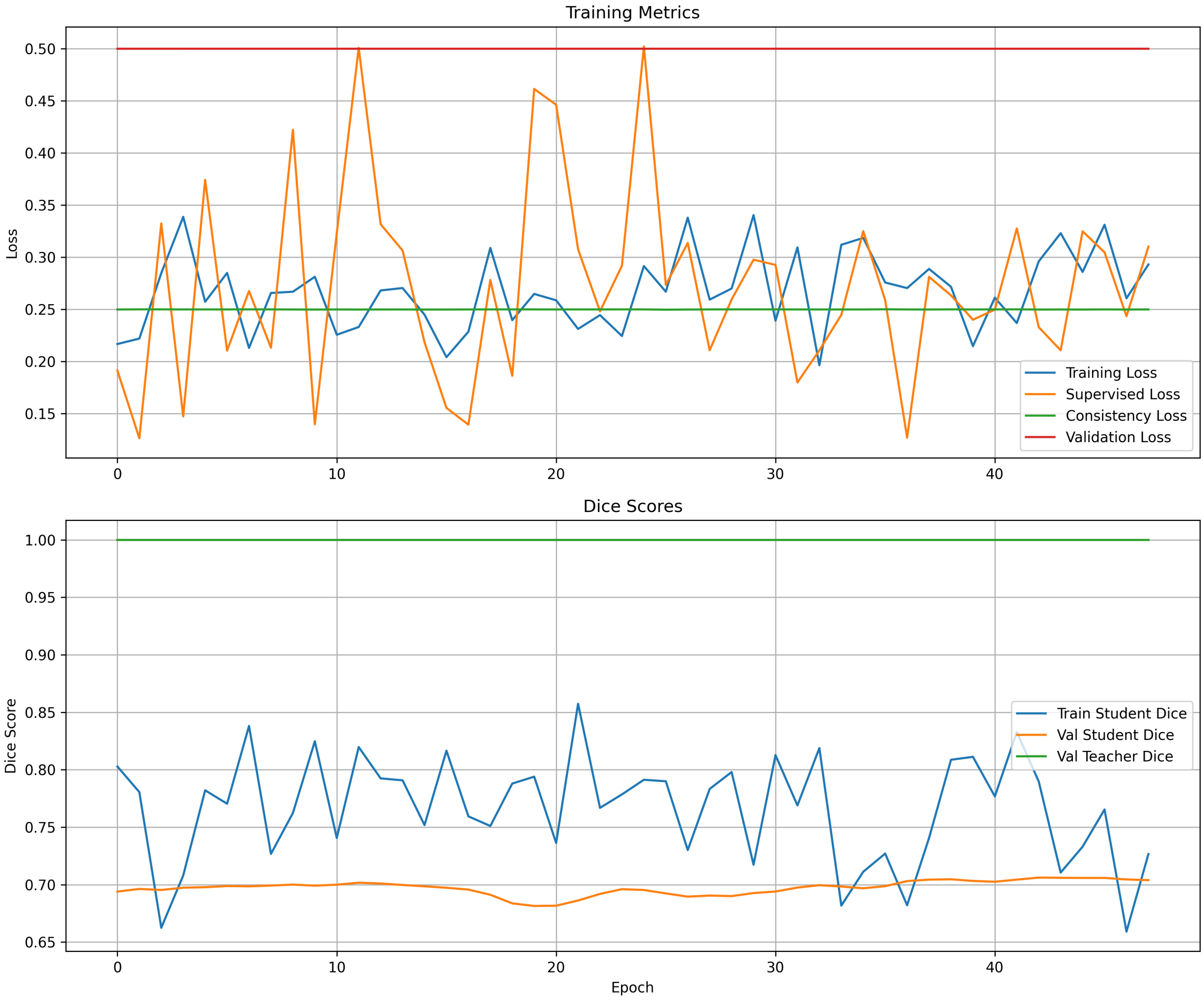

Fig. 13. Mean Teacher training metrics (50 Labeled patient cases, 45 epochs). Top: Losses (Total Train - blue, Supervised - orange, Consistency - green, Val - red). Bottom: Dice Scores (Train Student - blue, Val Student - orange, Val Teacher - green). Teacher Val DSC shows an artifactual 1.0.

The student model's validation Dice score (Figure 13, bottom panel, orange line) started at approximately 0.69–0.70. It exhibited considerable volatility throughout the 45 epochs of training, fluctuating primarily between 0.66 and 0.70. While early peaks reached towards 0.75 (e.g., approximately 0.75 at epoch 9, confirm from logs), no sustained improvement beyond its pre-training level (0.7223) was achieved. The training terminated at epoch 45 due to early stopping, with a final validation DSC of approximately 0.69. This performance did not surpass the more extensively trained 50-slice supervised baseline (0.8453, Section 4.1.2).

The teacher model's validation Dice score (green line) mirrored the student's performance during the initial direct copy phase but rapidly converged to and persistently remained at an artifactual 1.0 thereafter. The (weighted) consistency loss (top panel, green line) was very low throughout training after an initial value, indicating a weak SSL signal. The validation loss (red line) remained high and flat (∼0.50).

4.2.5. Interpretation of Teacher Model Behavior (Artifactual DSC)

The persistent validation Dice score of 1.0 for the teacher model, after its initial synchronization with the student, is a critical observation. This artifact does not indicate perfect segmentation of pancreatic tissue. Instead, it likely arises from a combination of factors:

1. **Teacher Model Collapse:** The teacher model, updated via a slow EMA and influenced by potentially noisy gradients from the student on unlabeled data (even with sharpening), appears to collapse towards a trivial solution, predominantly predicting background.

2. **Dice Coefficient Properties:** When evaluating on validation slices that are entirely background, if the teacher correctly predicts all background (TP=0, FP=0, FN=0 for foreground), the Dice score becomes 0/0, often handled as 1.0 or 0.0 depending on implementation. More likely, if many validation slices are truly background-only, and the teacher predicts them perfectly as background, these contribute a DSC of 1.0 (for the background class, if it were multi-class) or are handled such that empty prediction on empty ground truth yields 1.0 in some Dice implementations for binary cases when focusing on foreground. Given your 'DiceCoefficient' class calculates Dice for the foreground, an empty prediction on an empty foreground ground truth (all background slice) would lead to (2*0+smooth) / (0+0+smooth) = 1.0.

3. **Averaging Effect:** If a significant portion of the validation set consists of slices with no pancreatic tissue (all background), and the teacher correctly identifies these, these high DSC values (1.0) can dominate the average validation Dice score, masking poor performance on slices that do contain pancreatic tissue.

This artifactual performance means the teacher is not providing reliable pseudo-labels for foreground structures on unlabeled data, undermining the core mechanism of Mean Teacher.

4.2.6. Exploratory Run with Reduced Labeled Data (30 Labeled Cases)

To observe the algorithm's behavior under even more stringent data limitations, an exploratory run was conducted using only 30 labeled patient cases for pre-training and for the supervised component of Mean Teacher training (Figure 14). While specific hyperparameters for this run were not identical to the 50L experiment (e.g., different random seed, potentially different SSL parameter tuning during debugging), the general trends observed were informative.

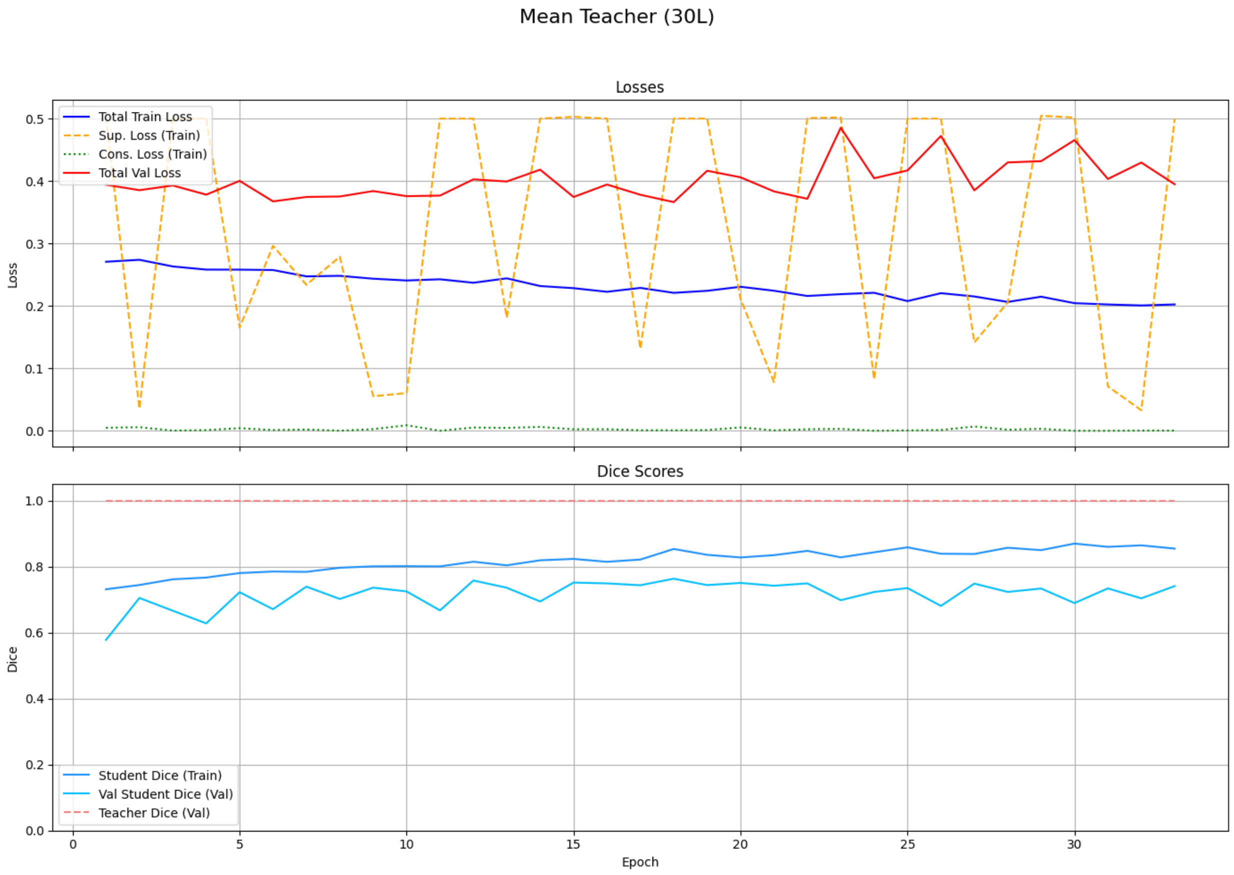

Fig. 14. Mean Teacher training metrics with 30 Labeled patient cases (33 epochs). Top: Losses. Bottom: Dice Scores. Teacher Val DSC (red dashed line) also exhibits an artifactual 1.0.

In this 30L scenario, the student model's validation Dice score (Figure 14, bottom panel, cyan line) again showed volatility, generally performing in the 0.60–0.75 range after an initial pre-training phase (details of pre-training for this specific 30L run are not shown but would be lower than the 50L pre-training). The teacher model's validation Dice (red dashed line) also rapidly achieved and maintained an artifactual 1.0. The consistency loss component (top panel, green dotted line) was again observed to be very low. This experiment, despite the different labeled data count, reinforced the observation that the teacher model's instability and the subsequent ineffectiveness of the consistency signal were persistent challenges, potentially exacerbated by further reducing the labeled set size.

4.2.7. Analysis of Mean Teacher Performance and Challenges (Overall Summary)

Across various configurations and data conditions (50L and exploratory 30L) for Mean Teacher, including student pre-training and phased teacher updates, a consistent set of challenges emerged, preventing the SSL framework from significantly enhancing segmentation performance over a well-established supervised baseline or even its own pre-training.

The pivotal issue was the ”teacher model's instability and collapse”. While initialized effectively through pre-training and direct copy mechanisms, the teacher model consistently failed to maintain robust foreground segmentation capabilities once its weights were updated via EMA. It rapidly converged to a state of predominantly predicting background, as evidenced by its artifactual validation Dice score of 1.0. This rendered the ”consistency regularization mechanism largely ineffective”. The student model, when regularized towards the teacher’s predictions on unlabeled data, was essentially being guided by pseudo-labels that were either uninformative (all background) or incorrect for foreground regions. Consequently, the very low observed consistency loss did not translate into improved student generalization for the primary task.

Several factors likely contribute to this outcome:

- **Pseudo-Label Quality from Unstable Teacher:** An unstable or collapsed teacher cannot generate reliable pseudo-labels, breaking the SSL learning loop.
- **Sensitivity in Low-Data Regimes:** With only 30-50 patient cases for labeled data, the initial models may not be robust enough to guide the SSL process effectively on diverse unlabeled data. The teacher might be overly influenced by ambiguous unlabeled samples or by noise in the student’s gradients.
- **2D Slice Limitations:** Isolated 2D slices may lack sufficient context for the teacher to make confident predictions on unlabeled data, especially for an anatomically complex and variable organ like the pancreas. This ambiguity could make it easier for the teacher to settle into a low-complexity solution (e.g., predict background).
- **Hyperparameter Effects:** While parameters like sharpening temperature ($T_{sharp} = 0.5$) and consistency weight ($\lambda_{C,max} = 10.0$) were chosen based on common practices and initial tuning, their interaction with an already struggling teacher model might not have been optimal. Aggressive sharpening of poor pseudo-labels can be detrimental.

These findings underscore the significant difficulty in successfully applying the standard Mean Teacher algorithm to this specific 2D pancreatic CT segmentation task under severe data scarcity. The primary bottleneck appears to be the challenge of ensuring the teacher model generates and maintains high-quality, reliable supervisory signals from unlabeled data throughout the training process. Without a dependable teacher, the consistency mechanism fails to provide a beneficial learning gradient.

4.3. Semi-Supervised Learning: MixMatch

The MixMatch algorithm [BCG+19], known for its holistic approach to SSL by combining data augmentation, pseudo-label sharpening, and MixUp consistency, was evaluated. Methodological details are provided in Section 3.6.1. This section presents findings from several experimental configurations aimed at understanding MixMatch’s efficacy for 2D pancreatic CT segmentation with limited labeled data.

4.3.1. Initial MixMatch Configuration and Performance (50 Labeled Cases)

4.3.2. Experimental Configuration.

The primary MixMatch experiment utilized 50 labeled patient cases (all constituent 2D slices) for the supervised component, 30 validation cases, and 201 unlabeled cases from the `preprocessed` dataset. The student model was initialized from a 20-epoch supervised pre-training on the 50 labeled cases (achieving a validation DSC of 0.7223). Key MixMatch hyperparameters for this run included: $K = 2$ weak augmentations for pseudo-label guessing, sharpening temperature $T = 0.5$, MixUp $\alpha_{mixup} = 0.75$, a maximum consistency weight $\lambda_{u,max} = 25.0$ (ramped over ∼525 training steps), and an Adam optimizer with an initial learning rate of 5×10^{-4} (cosine decay). A dropout rate of 0.1 was used. Early stopping (patience 30 epochs) was active, and steps per epoch were 13 (based on slices from 50L cases and a labeled stream batch size of 4, `BATCH_SIZE_MM=8` in the script implied 4L+4U streams).

4.3.3. Performance and Training Dynamics

The learning curves for this 50L MixMatch experiment are presented in Figure 15.

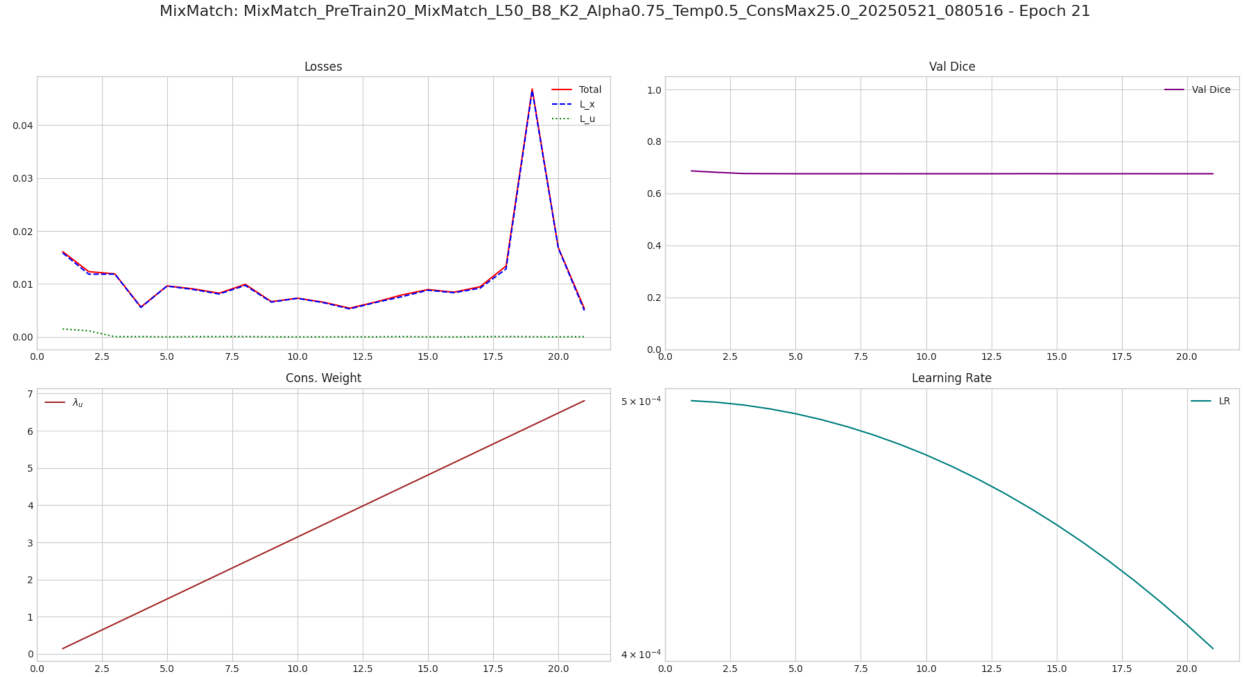

Fig. 15. MixMatch training metrics (50 Labeled patient cases, 21 epochs post-pre-training). Top-left: Losses (Total, L_x, L_u). Top-right: Validation Dice. Bottom-left: $\lambda_u(t)$. Bottom-right: Learning Rate. Validation Dice rapidly degrades.

Training terminated after 21 epochs due to early stopping. The supervised loss (L_x) on mixed labeled data decreased (e.g., to ∼0.16 by epoch 21). However, the unsupervised consistency loss (L_u) remained effectively zero throughout, despite $\lambda_u(t)$ ramping up. Most critically, the validation Dice score (Figure 15, top-right) catastrophically declined from its pre-training level (starting ∼0.68 at the beginning of the MixMatch phase) to near-zero values (best during SSL phase ≈ 0.0114)

within the first few epochs.

4.3.4. Initial Analysis

This primary configuration demonstrated a significant failure of the MixMatch SSL component. The near-zero L_u coupled with the validation Dice collapse strongly suggests that the pseudo-labels generated for unlabeled data were of extremely poor quality (likely predicting background or other trivial solutions). The model then effectively learned to replicate these incorrect targets, making the SSL component detrimental.

4.3.5. Exploring MixMatch with Varied Labeled Data and Pre-training Durations

To further investigate the behavior of MixMatch and its sensitivity to the amount of initial supervised signal, additional exploratory experiments were conducted with different numbers of labeled cases and corresponding pre-training durations, while keeping other core MixMatch hyperparameters ($K = 2, T = 0.5, \alpha_{mixup} = 0.75, \lambda_{u,max} = 25.0$, LR 5×10^{-4}) largely consistent.

4.3.6. Experiment with 30 Labeled Cases (10 Epochs Pre-training).

In this setup, 30 patient cases were used for labeled data, and the student model was pre-trained for 10 epochs. The subsequent MixMatch phase (Figure 16) ran for 21 epochs before early stopping.

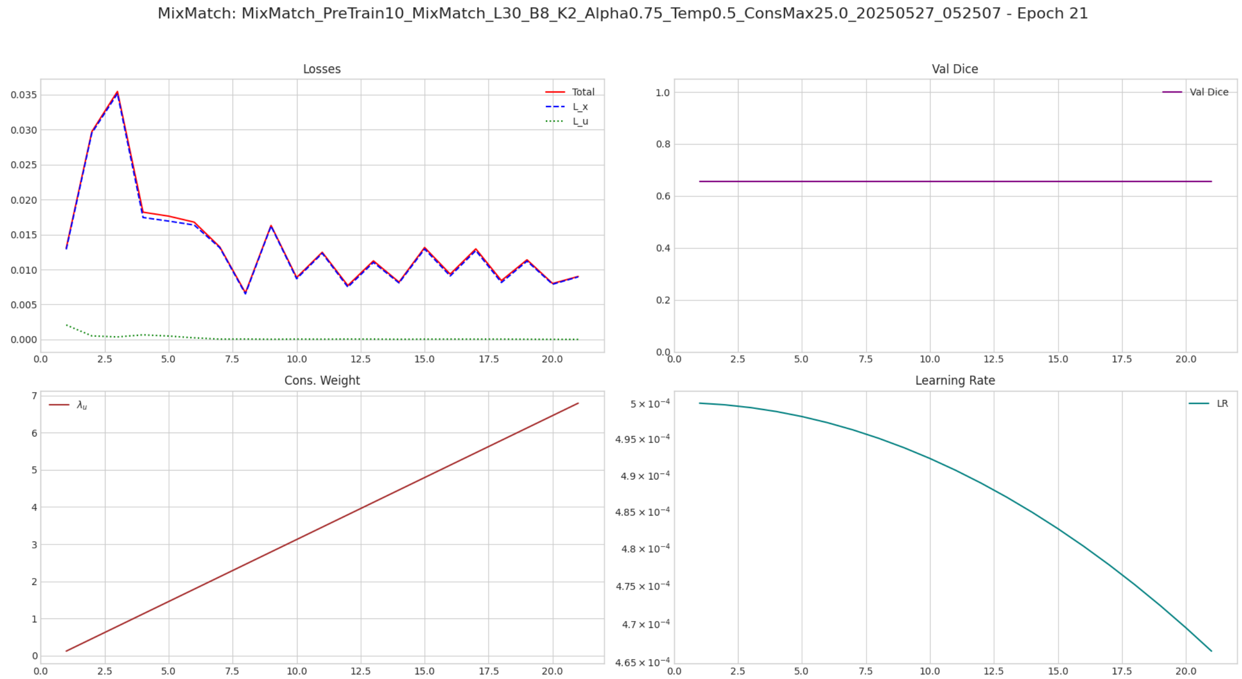

Fig. 16. MixMatch training metrics (30 Labeled patient cases, 10 epochs pre-training, 21 MixMatch epochs). Panels show Losses, Validation Dice, Consistency Weight, and Learning Rate. Validation Dice remains at pre-training level.

The validation Dice score (Figure 16, top-right, purple line) started at approximately 0.65–0.68 (reflecting its 10-epoch pre-training on 30L cases) and remained essentially flat at this level throughout the 21 epochs of MixMatch training. The unsupervised loss (L_u, top-left, green dotted line) again started very low and stayed negligible. This indicates that even with fewer initial labels and shorter pre-training, the MixMatch SSL component did not provide any positive learning signal to improve upon the pre-trained baseline.

4.3.7. Experiment with 40 Labeled Cases (20 Epochs Pre-training).

The number of labeled cases was increased to 40, and pre-training was extended to 20 epochs. The MixMatch phase (Figure 17) ran for 15 epochs.

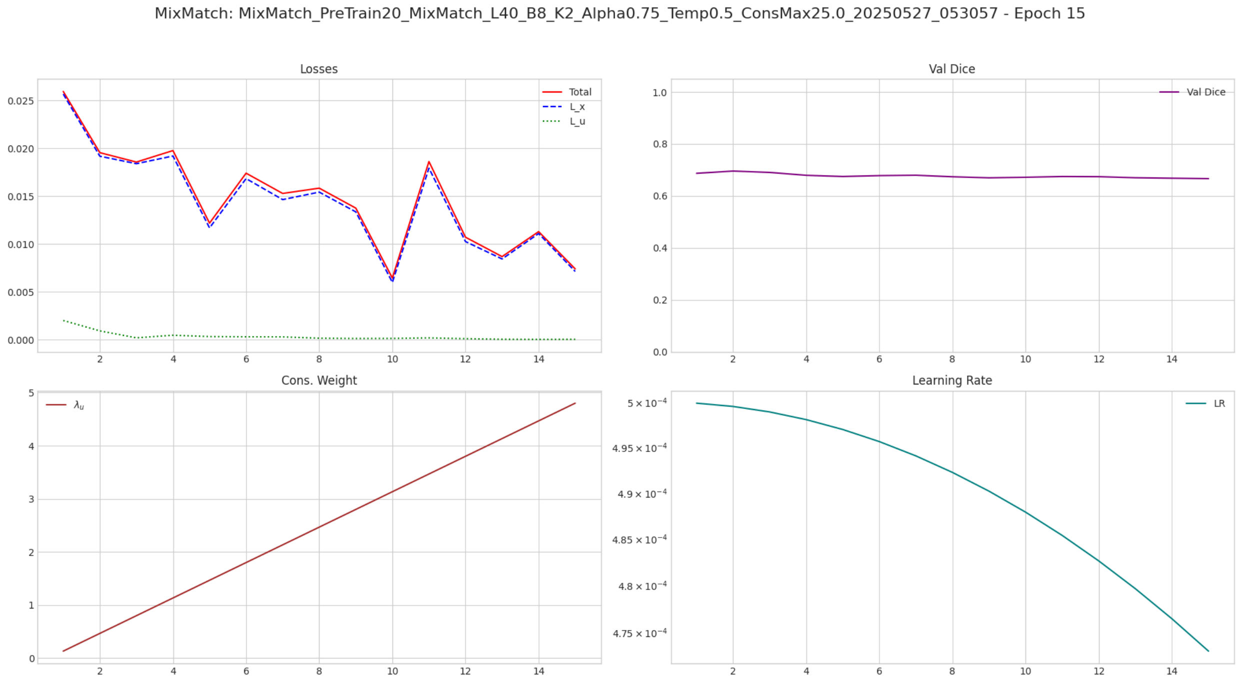

Fig. 17. MixMatch training metrics (40 Labeled patient cases, 20 epochs pre-training, 15 MixMatch epochs). Panels show Losses, Validation Dice, Consistency Weight, and Learning Rate. Validation Dice remains at pre-training level.

Similar to the 30L experiment, the validation Dice score (Figure 17, top-right, purple line) started around 0.68–0.70 (reflecting its 20-epoch pre-training on 40L cases) and showed no improvement during the MixMatch phase, plateauing at this level. The unsupervised loss (L_u) also remained consistently near-zero.

4.3.8. Analysis of Varied Setups.

These exploratory runs with 30 and 40 labeled cases, despite different pre-training intensities, consistently showed that the MixMatch SSL phase did not enhance performance beyond the level achieved by supervised pre-training alone. The persistent near-zero L_u across these experiments reinforces the hypothesis that the pseudo-label generation process within MixMatch, under these conditions, failed to produce useful targets for the unlabeled data. The model achieved trivial consistency with likely background-dominant pseudo-labels. Increasing the initial labeled set from 30 to 40 (and the pre-training duration) resulted in a slightly better starting validation DSC, but the SSL component was still unable to capitalize on this.

4.3.9. Overall Analysis of MixMatch Performance and Challenges

Across all tested configurations for MixMatch, including the primary run with 50 labeled cases and exploratory runs with 30 and 40 labeled cases, the algorithm failed to leverage unlabeled data to improve segmentation performance. In the 50L case, it was markedly detrimental, while in the 30L and 40L cases, it led to stagnation at the pre-training performance level.

The consistent observation of a near-zero unsupervised consistency loss (L_u), irrespective of the number of labeled samples (30, 40, or 50) or the strength of pre-training, is the most critical finding. This strongly suggests a fundamental failure in the pseudo-labeling pipeline of MixMatch for this specific 2D pancreatic CT segmentation task with limited data. It is hypothesized that:

- **Poor Quality of Initial Pseudo-Labels:** The student model, even after pre-training, likely generates pseudo-labels for unlabeled data that are heavily biased towards the background class or are otherwise uninformative for segmenting the relatively small and complex pancreatic structures.

- **Ineffectiveness of Averaging and Sharpening:** The process of averaging predictions from $K = 2$ weak augmentations and then applying temperature sharpening ($T = 0.5$) was insufficient to correct or refine these poor initial pseudo-labels into useful supervisory signals. Instead, sharpening might have amplified confidence in incorrect (e.g., background) predictions.

- **Trivial Consistency:** The student model could easily achieve very low L_u by predicting outputs consistent with these flawed pseudo-labels on the mixed unlabeled data, without actually learning meaningful foreground features from the unlabeled set.

- **Impact of 2D Slice Input:** The lack of 3D context in isolated 2D slices likely makes it harder for the model to generate reliable pseudo-labels, especially for an organ with complex inter-slice continuity and variable appearance.

The MixMatch framework, therefore, in its current implementation and under these challenging low-data 2D conditions, provided a strong but ultimately misdirected learning signal from the unlabeled data, either leading to performance collapse or stagnation. This underscores the critical dependence of MixMatch on the generation of at least moderately reliable pseudo-labels for the SSL mechanism to be beneficial.

5. Discussion

The empirical evaluations detailed in Chapter 4 offer significant insights into the application of supervised and semi-supervised learning (SSL) methodologies for the challenging task of 2D pancreatic CT image segmentation, particularly under the constraints of severely limited labeled data. This chapter synthesizes these experimental findings, provides a comparative analysis of the evaluated learning approaches, delves into the broader implications and inherent difficulties encountered, candidly outlines the limitations of the current study, and proposes actionable directions for future research in this domain.

5.1. Comparative Analysis of Learning Approaches with Limited Labeled Data

All semi-supervised learning experiments presented in this thesis were conducted utilizing 50 labeled patient cases (comprising all their constituent 2D slices after preprocessing) from the `preprocessed` dataset. The performance of these SSL methods was critically benchmarked against a direct supervised U-Net model, which was also trained on the identical set of 50 labeled patient cases and subjected to extensive training to establish its optimal performance. A consolidated summary of the peak validation Dice Similarity Coefficient (DSC) achieved and key observational takeaways for each learning approach is presented in Table 2.

Table 2. Performance Summary: Supervised vs. Semi-Supervised Learning with 50 Labeled Patient Cases from the `preprocessed` Dataset.

Training Approach	**Peak Validation DSC**	**Key Observations Summary**
Supervised U-Net	**0.8453**	Robust and stable learning; established the performance benchmark for limited labeled data.
(50 Labeled Cases)	(at epoch 58)[a]	
Mean Teacher	$\sim$0.75 (initial peak)[b]	Teacher model rapidly collapsed to an artifactual DSC (≈ 1.0); SSL phase did not yield sustained improvement over pre-training performance.
(50 Labeled + 201 Unlabeled Cases)	$\sim$0.69 (at early stop)[c]	
MixMatch	0.7223 (from pre-training)[d]	Validation DSC catastrophically declined to $\approx$0.0114 during SSL phase; SSL component proved highly detrimental to overall performance.
(50 Labeled + 201 Unlabeled Cases)	$\approx$ **0.0114** (during SSL)	

[a] Peak validation DSC and corresponding epoch for the 50-labeled-case supervised baseline run (Figure 10).

[b] Approximate peak validation DSC observed during the initial phase (reflecting pre-training and direct teacher copy) of Mean Teacher training.

[c] Approximate validation DSC at termination of Mean Teacher training due to early stopping (Figure 13).

[d] Validation DSC after 20-epoch supervised pre-training, serving as initialization for MixMatch SSL phase.

The direct supervised baseline, when trained comprehensively using all 2D slices derived from the 50 labeled patient cases (as detailed in Section 4.1.2), achieved a strong peak validation Dice score of 0.8453 at epoch 58. This result effectively established a high-performance benchmark, clearly indicating the segmentation accuracy attainable by the `PancreasSeg` U-Net architecture with this specific limited dataset under optimized, fully supervised conditions.

The Mean Teacher algorithm, despite being initialized with a student model pre-trained to a validation DSC of 0.7223 (from 20 epochs on the same 50 labeled cases), did not yield sustained improvements through the semi-supervised learning phase. While the student's validation DSC exhibited an early peak (e.g., approximately 0.75, largely influenced by its pre-training and the initial direct teacher-copy phase), it subsequently failed to consistently surpass this level, ultimately settling around 0.69 at the point of early stopping (Figure 13). The primary impediment, as analyzed in Section 3.5, was identified as the teacher model's rapid convergence to an artifactual state (validation DSC ≈ 1.0), which rendered the consistency loss mechanism ineffective for guiding meaningful foreground segmentation.

Similarly, the MixMatch algorithm, also initialized from the identical pre-trained student model (validation DSC 0.7223), demonstrated a markedly counterproductive effect during its SSL training phase. The validation Dice score experienced a catastrophic decline from its pre-training level, plummeting to near-zero values (best recorded DSC during the SSL phase was ≈ 0.0114). The consistently negligible unsupervised loss ($L_u \approx 0$), as detailed in Section 3.6, suggested that while the model achieved superficial consistency with its generated pseudo-labels, these pseudo-labels were of extremely poor quality, likely reflecting trivial solutions such as predominantly background predictions.

In essence, under the specific configurations and data conditions tested, neither the Mean Teacher nor the MixMatch algorithm successfully leveraged the 201 unlabeled patient cases (and their constituent slices) to improve upon the performance achieved by a direct supervised model trained with only the 50 labeled patient cases from the `preprocessed` dataset.

5.2. Computational Efficiency Analysis

An analysis of the computational overhead associated with each training approach revealed differences in per-epoch processing times. The supervised baseline training on data from 50 labeled patient cases required approximately 230 seconds per epoch on an Tesla A100 GPU. The Mean Teacher algorithm, which involves operations for both student and teacher models alongside consistency loss computation, exhibited a per-epoch training time of approximately 340 seconds. The MixMatch algorithm, with its K-augmentations for pseudo-label guessing and subsequent MixUp operations, required a comparable time of approximately 346 seconds per epoch.

While the SSL methods inherently introduce additional computational steps per epoch compared to direct supervised training, the observed per-epoch time differences are not excessively prohibitive for research or offline training scenarios. The primary factor influencing total training duration in these experiments was the number of epochs executed before convergence criteria were met or early stopping was triggered. Given the performance outcomes, where SSL methods did not demonstrate superior efficacy, the choice between these approaches, in their current implementations for this task, would be predominantly driven by segmentation performance rather than marginal differences in per-epoch computational cost.

5.3. Challenges in Semi-Supervised Learning for 2D Pancreatic Slice Segmentation

The experimental outcomes from this study highlight several significant challenges inherent in the application of established SSL methods to 2D slice-based pancreatic CT segmentation, particularly when operating with a severely limited pool of labeled data. A recurrent primary issue observed across both Mean Teacher and MixMatch was the instability and suboptimal quality of the teacher model's predictions or the generated pseudo-labels. With only 50 labeled patient cases (providing a limited set of 2D slices), which may inadequately represent the full spectrum of pancreatic and tumor appearances, as well as the considerable surrounding anatomical variability, the models struggled to develop a sufficiently robust and generalizable internal representation. This difficulty manifested as a functional collapse of the teacher model in the Mean Teacher framework and likely contributed to the generation of poor-quality, uninformative pseudo-labels in MixMatch, thereby preventing the SSL mechanisms from imparting beneficial learning signals for accurate foreground segmentation.

Furthermore, SSL methods often introduce a complex array of new hyperparameters, including consistency weights, sharpening temperatures, and parameters for auxiliary mechanisms like MixUp. The optimization of this expanded hyperparameter space can be particularly arduous in low-data regimes. Validation signals derived from a small validation set may themselves be noisy, and the learning signals from unlabeled data can be misleading if not properly regularized or guided. For instance, the aggressive sharpening temperatures and high consistency pressures explored in some Mean Teacher configurations might have inadvertently exacerbated teacher model instability rather than promoting robust learning.

The inherent nature of 2D slice-based segmentation, which discards valuable 3D contextual information, further complicates the learning task. Pancreatic boundaries are frequently ambiguous on isolated 2D CT slices, making it more challenging for models to learn robust distinguishing features from a few examples. This ambiguity may also render it easier for SSL models to converge to trivial solutions, such as predicting all background, when processing unlabeled data where no ground truth constraint is present to penalize such behavior. Lastly, the pancreas is a relatively small organ, and tumor characteristics can be subtle and highly variable. Without a sufficient number of diverse labeled examples covering this variability, models may struggle to differentiate foreground pancreatic tissue from complex surrounding background textures—a problem that SSL, if reliant on flawed pseudo-signals derived from these same challenging inputs, may not readily overcome.

5.4. Limitations of the Current Study

While this research provides valuable insights into the application of SSL for pancreatic cancer segmentation, several limitations inherent to the study's scope and execution must be acknowledged. Firstly, the exploration of the SSL hyperparameter space, while guided by common practices and initial experiments, was not exhaustive. A comprehensive grid search or more advanced automated hyperparameter optimization for all respective parameters of Mean Teacher and MixMatch (e.g., various ramp-up schedules, EMA decay rates, MixUp α values, number of K-augmentations for pseudo-labeling) was beyond the practical computational scope of this Master's thesis. It remains plausible that alternative, meticulously tuned configurations might yield improved SSL performance.

Secondly, the evaluation was focused on two prominent SSL algorithms. Other SSL paradigms, such as various forms of self-training employing more sophisticated pseudo-label selection criteria, or advanced uncertainty-guided consistency methods beyond those implicitly part of the Mean Teacher framework, were not investigated. The choice was made to deeply evaluate two distinct, influential methods rather than superficially survey a larger number.

A third significant limitation is the inherent 2D slice-based approach adopted for model architecture and data processing. This methodology, while common for initial explorations and computationally more tractable, fundamentally limits the models' ability to leverage 3D spatial context, which is widely recognized as beneficial, if not essential, for robust volumetric medical image segmentation tasks. Fourthly, all experiments were conducted on a single, albeit publicly available and relevant, dataset (MSD Pancreas-CT), processed through a specific, standardized pipeline. The generalizability of the findings to other pancreatic CT datasets, or to datasets subjected to different preprocessing and data representation strategies, was not assessed.

Finally, the 'enhancement' aspect of the thesis objectives—beyond the systematic evaluation—was constrained by available computational time and resources. This limited the breadth and depth of iterative algorithmic modifications and re-evaluations that could be performed for Mean Teacher and MixMatch specifically to optimize their performance under the identified challenges within the timeframe of this project.

5.5. Future Directions

The findings and limitations of this study illuminate several promising and critical directions for future research aimed at improving the efficacy of semi-supervised learning for medical image segmentation, particularly in challenging low-data scenarios such as pancreatic cancer. A primary and overarching focus should be on enhancing the robustness of teacher models in EMA-based frameworks like Mean Teacher, and more generally, improving the quality and reliability of pseudo-label generation processes in methods like MixMatch and self-training. For Mean Teacher, this could involve investigating techniques to stabilize the teacher, such as adaptive EMA decay rates that respond to training dynamics, exploring alternative student-teacher architectural relationships (e.g., different capacities or regularization), or dynamically incorporating uncertainty estimates to modulate the teacher's influence on the consistency loss, thereby down-weighting less reliable pseudo-signals. For MixMatch and related self-training paradigms, future work should explore more sophisticated pseudo-label generation and selection strategies, potentially integrating model confidence scores, predictive uncertainty metrics, or ensemble-based predictions to filter or refine the pseudo-labels that are used for training on unlabeled data.

Exploring advanced consistency regularization strategies beyond simple Mean Squared Error loss on output probabilities is another important avenue. This could include enforcing consistency at intermediate feature levels within the network, employing adversarial training techniques to align the distributions of predictions on labeled and unlabeled data, or developing regularization terms based on higher-order statistical properties or invariances under various image transforms (e.g., in the Fourier domain). A particularly significant step forward for this specific application would likely be the transition from 2D to 3D segmentation models, such as 3D U-Nets or V-Nets, for both supervised and SSL approaches. Utilizing full volumetric information may inherently improve model robustness and reduce slice-level ambiguity, potentially creating a more favorable environment for SSL techniques to generate reliable signals from unlabeled data.

Furthermore, the development of hybrid SSL approaches warrants thorough investigation. Such methods could aim to combine the strengths of different SSL techniques (e.g., integrating robust pseudo-labeling with strong consistency regularization) or integrate SSL with other learning paradigms. For instance, leveraging transfer learning from models pre-trained on larger, related medical imaging datasets could provide a stronger initialization for SSL fine-tuning. Incorporating contrastive self-supervised pre-training on the available unlabeled data itself, prior to commencing the SSL phase, could also help learn more discriminative initial features. Investigating methods for adaptive hyperparameter tuning for SSL, where critical parameters like consistency weights or sharpening temperatures are dynamically adjusted during training based on the evolving model state or data characteristics, could also prove beneficial over fixed schedules, especially in non-stationary learning environments. Finally, to robustly assess generalizability and clinical relevance, any developed or enhanced SSL methods must be rigorously evaluated on larger and more diverse datasets, potentially including multi-modal imaging data if pertinent to the clinical application and data availability.

The insights gained from this work, despite the challenges encountered with the direct application of standard SSL implementations, contribute to the broader understanding of applying these techniques to complex medical image segmentation tasks where annotated data is scarce. The identified failure modes provide a clear and empirically grounded basis for targeted research into more robust and effective semi-supervised learning strategies for the medical domain.

6. Conclusion

This thesis undertook an evaluation of semi-supervised learning (SSL) algorithms, specifically Mean Teacher and MixMatch, for the task of 2D pancreatic tumor segmentation from CT images, with a primary focus on scenarios constrained by limited labeled data. The research aimed to assess the efficacy of these SSL methods in leveraging unlabeled data to achieve segmentation accuracy comparable to or exceeding that of fully supervised models, thereby addressing the challenge of costly and time-consuming data annotation in medical imaging. Through systematic experimentation using the `preprocessed` dataset derived from the MSD Pancreas-CT challenge, this work has provided insights into both the theoretical promise and the significant practical challenges of applying these SSL techniques.

The principal conclusions drawn from this research are as follows:

- **SSL Performance in Low-Data Regime:** Under the tested configurations and with only 50 labeled patient cases (and their constituent 2D slices) for training, neither Mean Teacher nor MixMatch demonstrated an ability to outperform a robustly trained supervised U-Net baseline that utilized the same limited labeled dataset. The supervised baseline, after extensive training, achieved a peak validation Dice Similarity Coefficient (DSC) of approximately 0.8453.

- **Mean Teacher Challenges:** The Mean Teacher algorithm, while benefiting from student pre-training, was significantly hampered by teacher model instability. The teacher model rapidly converged to an artifactual state (yielding a DSC near 1.0), rendering the consistency regularization mechanism ineffective for improving foreground segmentation. This highlights a critical vulnerability in the student-teacher knowledge transfer process under severe data scarcity.

- **MixMatch Inefficacy and Detrimental Effects:** The MixMatch algorithm proved detrimental to performance, causing a catastrophic decline in validation DSC from its pre-trained initialization to near-zero values. This failure was attributed to the generation of extremely poor-quality pseudo-labels, with which the model achieved trivial consistency, leading to a strong but misdirected learning signal.

- **Criticality of Pseudo-Signal Quality:** A core finding across both SSL methods is the paramount importance of the quality and stability of the supervisory signals derived from unlabeled data (i.e., teacher predictions or pseudo-labels). In low-data regimes, the generation of reliable pseudo-signals is exceptionally challenging, and flawed signals can nullify or reverse any potential benefits of SSL.

- **Impact of Pre-training:** Supervised pre-training on the limited labeled set provided a strong initial performance (e.g., 0.7223 DSC after 20 epochs). However, the SSL mechanisms, as implemented, were not robust enough to build upon or even consistently maintain this initial knowledge, particularly in the case of MixMatch.

- **Computational Considerations:** While SSL methods introduced a moderate increase in per-epoch computational time compared to supervised training, the overall training duration was more significantly influenced by convergence behavior and early stopping. Performance and stability, therefore, remain more critical factors than marginal per-epoch costs in selecting an appropriate method for this task.

The contributions of this research lie in the rigorous empirical evaluation of these standard SSL techniques within a challenging clinical application, the identification of specific failure modes related to teacher stability and pseudo-label quality in data-scarce environments, and the establishment of relevant performance benchmarks for 2D pancreatic segmentation on the `preprocessed` dataset configuration.

The insights gained, particularly concerning the difficulties in generating reliable supervisory signals from unlabeled data with limited initial supervision, underscore the need for more robust SSL mechanisms. Future research, as detailed in Section 5.5, should focus on strategies to enhance model stability, improve pseudo-label generation and refinement, explore advanced consistency techniques, and investigate the potential of 3D architectures to better leverage contextual information.

In summary, while this thesis highlighted significant practical challenges in applying Mean Teacher and MixMatch to 2D pancreatic CT segmentation with very limited labeled data, the detailed analysis of their performance provides valuable lessons and a clear impetus for future research aimed at developing more effective and reliable semi-supervised learning solutions for medical imaging.

References

[AAR+24] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, et al. Medical image segmentation review: the success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[ARB+22] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[BCG+19] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: a holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

[Ben12] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade: Second edition*, pp. 437–478. Springer, 2012.

[CDY+20] Yanwen Chong, Yun Ding, Qing Yan, and Shaoming Pan. Graph-based semi-supervised learning: a review. *Neurocomputing*, 408:216–230, 2020.

[CEK+23] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. *Medical Image Analysis*, 87:102792, 2023. DOI: `10.1016/j.media.2023.102792`. URL: `https://www.sciencedirect.com/science/article/abs/pii/S0010482523013057`.

[CTQ+21] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35 of number 8, pp. 6912–6920. AAAI Press, 2021. DOI: `10.1609/aaai.v35i8.16852`. URL: `https://doi.org/10.1609/aaai.v35i8.16852`.

[Dic45] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[DL24] Weizhen Ding and Zhen Li. Curriculum consistency learning and multi-scale contrastive constraint in semi-supervised medical image segmentation. *Bioengineering*, 11(1), 2024. ISSN: 2306-5354. DOI: `10.3390/bioengineering11010010`. URL: `https://www.mdpi.com/2306-5354/11/1/10`.

[Jia+23] R. Jiao et al. Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, 152:105305, 2023. DOI: `10.1016/j.compbiomed.2023.105305`. URL: `https://www.sciencedirect.com/science/article/abs/pii/S0010482523013057`.

[JSD24] Suchi Jain, Geeta Sikka, and Renu Dhir. A systematic literature review on pancreas segmentation from traditional to non-supervised techniques in abdominal medical images. *Artificial Intelligence Review*, 57(12):317, 2024.

[KBD+19] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In M. Jorge Cardoso, Aasa Feragen, Ben Glocker, Ender Konukoglu, Ipek Oguz, Gozde Unal, and Tom Vercauteren, editors, *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, vol. 102 of *Proceedings of Machine Learning Research*, pp. 285–296. PMLR, 2019-08–10 Jul. URL: `https://proceedings.mlr.press/v102/kervadec19a.html`.

[KCS+22] Hee E Kim, Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E Maros, and Thomas Ganslandt. Transfer learning for medical image classification: a literature review. *BMC medical imaging*, 22(1):69, 2022.

[KMŠ+23] Olga Kurasova, Viktor Medvedev, Aušra Šubonienė, Gintautas Dzemyda, Aistė Gulla, Artūras Samuilis, Džiugas Jagminas, and Kęstutis Strupas. Semi-supervised learning with pseudo-labeling for pancreatic cancer detection on ct scans. In *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, pp. 1–6, 2023. DOI: `10.23919/CISTI58278.2023.10211356`. URL: `https://ieeexplore.ieee.org/abstract/document/10211356`.

[Lee13] D.-H. Lee. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, 2013. URL: `https://www.kaggle.com/blobs/download/forum-message-attachment-files/746/pseudo_label_final.pdf`.

[LLH+22] Shangqing Liu, Shujun Liang, Xia Huang, Xinrui Yuan, Tao Zhong, and Yu Zhang. Graph-enhanced u-net for semi-supervised segmentation of pancreas from abdomen ct scan. *Physics in Medicine Biology*, 67(15):155017, 2022-07. DOI: `10.1088/1361-6560/ac80e4`. URL: `https://dx.doi.org/10.1088/1361-6560/ac80e4`.

[LYC+21] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2021. DOI: `10.1109/TNNLS.2020.2995319`.

[LYF+23] Liyun Lu, Mengxiao Yin, Liyao Fu, and Feng Yang. Uncertainty-aware pseudo-label and consistency for semi-supervised medical image segmentation. *Biomedical Signal Processing and Control*, 79:104203, 2023. ISSN: 1746-8094. DOI: `https://doi.org/10.1016/j.bspc.2022.104203`. URL: `https://www.sciencedirect.com/science/article/pii/S1746809422006577`.

[LZY+23] Shanfu Lu, Zijian Zhang, Ziye Yan, Yiran Wang, Tingting Cheng, Rongrong Zhou, and Guang Yang. Mutually aided uncertainty incorporated dual consistency regularization with pseudo label for semi-supervised medical image segmentation. *Neurocomputing*, 548:126411, 2023. DOI: `10.1016/j.neucom.2023.126411`. URL: `https://www.sciencedirect.com/science/article/abs/pii/S0925231223005349`.

[MCK+23] Aliasghar Mortazi, Vedat Cicek, Elif Keles, and Ulas Bagci. Selecting the best optimizers for deep learning–based medical image segmentation. *Frontiers in Radiology*, 3, 2023. ISSN: 2673-8740. DOI: 10.3389/fradi.2023.1175473. URL: https://www.frontiersin.org/journals/radiology/articles/10.3389/fradi.2023.1175473.

[MKT+18] Andrew McGuigan, Paul Kelly, Richard C Turkington, Claire Jones, Helen G Coleman, and Stephen RS McCain. Pancreatic cancer: a review of clinical diagnosis, epidemiology, treatment and outcomes. *World Journal of Gastroenterology*, 24(43):4846–4861, 2018. DOI: 10.3748/wjg.v24.i43.4846. URL: https://www.wjgnet.com/1007-9327/full/v24/i43/4846.htm.

[MSK22] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, 15(1):210, 2022.

[Myr19] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pp. 311–320, Cham. Springer International Publishing, 2019. DOI: 10.1007/978-3-030-11726-9_28.

[PBB+19] Christian S Perone, Pedro Ballester, Rodrigo C Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019.

[Pei+18] M. Peikari et al. A cluster-then-label semi-supervised learning approach for pathology image classification. *Scientific Reports*, 8(1):24876, 2018. DOI: 10.1038/s41598-018-24876-0. URL: https://www.nature.com/articles/s41598-018-24876-0.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

[RFL+15] Holger R Roth, Amal Farag, Le Lu, Evrim B Turkbey, and Ronald M Summers. Deep convolutional networks for pancreas segmentation in ct imaging. In *Medical Imaging 2015: Image Processing*, vol. 9413, pp. 378–385. SPIE, 2015.

[RFT+16] H. Roth, A. Farag, E. B. Turkbey, L. Lu, J. Liu, and R. M. Summers. Data from pancreas-ct (version 2) [data set], 2016. DOI: 10.7937/K9/TCIA.2016.tNB1kqBU. URL: https://doi.org/10.7937/K9/TCIA.2016.tNB1kqBU.

[RVR18] YCAP Reddy, P Viswanath, and B Eswara Reddy. Semi-supervised learning: a brief review. *Int. J. Eng. Technol*, 7(1.8):81, 2018. DOI: 10.14419/ijet.v7i1.8.9977.

[SCH+24] Min Shao, Chao Cheng, Chengyuan Hu, Jian Zheng, Bo Zhang, Tao Wang, Gang Jin, Zhaobang Liu, and Changjing Zuo. Semisupervised 3d segmentation of pancreatic tumors in positron emission tomography/computed tomography images using a mutual information minimization and cross-fusion strategy. *Quantitative Imaging in Medicine and Surgery*, 14(2):1747, 2024.

[SEG17] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3D fully convolutional deep networks. *arXiv e-prints*:arXiv:1706.05721, arXiv:1706.05721, 2017-06. DOI: `10.48550/arXiv.1706.05721`. arXiv: `1706.05721 [cs.CV]`.

[SJT16] M. Sajjadi, M. Javanmardi, and T. Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, vol. 29, pp. 1163–1171, 2016. URL: `https://proceedings.neurips.cc/paper/2016/hash/7c7e8a4a6624a13108d00e0db225b7a5-Abstract.html`.

[SK19] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[SK22] Junya Sato and Shoji Kido. Large batch and patch size training for medical image segmentation, 2022. arXiv: `2210.13364 [eess.IV]`. URL: `https://arxiv.org/abs/2210.13364`.

[SRK+22] Constantin Marc Seibold, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. Reference-guided pseudo-label generation for medical semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 36 of number 2, pp. 2171–2179, 2022. DOI: `10.1609/aaai.v36i2.20114`. URL: `https://ojs.aaai.org/index.php/AAAI/article/view/20114`.

[TV17] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[WLC+22] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. Medical image segmentation using deep learning: a survey. *IET image processing*, 16(5):1243–1267, 2022.

[Wu+22] Y. Wu et al. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 74:102134, 2022. DOI: `10.1016/j.media.2022.102134`. URL: `https://www.sciencedirect.com/science/article/abs/pii/S1361841522001773`.

[YSP21] Rahul Yedida, Snehanshu Saha, and Tejas Prashanth. Lipschitzlr: using theoretically computed adaptive learning rates for fast convergence. *Applied Intelligence*, 51:1460–1478, 2021. DOI: `10.1007/s10489-020-01892-0`.

[YTW+23] Xiaosu Yang, Jiya Tian, Yaping Wan, Mingzhi Chen, Lingna Chen, and Junxi Chen. Semi-supervised medical image segmentation via cross-guidance and feature-level consistency dual regularization schemes. *Medical Physics*, 50(7):4269–4281, 2023. DOI: 10.1002/mp.16217. URL: https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16217.

[YYZ+21] Hang Yu, Laurence T Yang, Qingchen Zhang, David Armstrong, and M Jamal Deen. Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. *Neurocomputing*, 444:92–110, 2021.

[ZG23] Jiasen Zhang and Weihong Guo. A new regularization for deep learning-based segmentation of images with fine structures and low contrast. *Sensors*, 23(4), 2023. ISSN: 1424-8220. DOI: 10.3390/s23041887. URL: https://www.mdpi.com/1424-8220/23/4/1887.

[ZZH+23] Wenqiao Zhang, Lei Zhu, James Hallinan, Andrew Makmur, Shengyu Zhang, Qingpeng Cai, and Beng Chin Ooi. Boostmis: boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. *arXiv preprint arXiv:2203.02533*, 2023. DOI: 10.48550/arXiv.2203.02533. URL: https://arxiv.org/abs/2203.02533.