HCMC University of Technology and Education

Faculty of Mechanical Engineering

Department of Mechatronics

MEDICAL IMAGE SEGMENTATION

Lecturer: Dr. Tran Vu Hoang

Members

Order	Name	ID
1	Huynh Thanh Phong	22134009
2	Le Quoc Thinh	22134013

ThuDuc, 26th December

Table of Contents

I. Introduction

1.1 Motivaton

Medical imaging is a vital tool in modern healthcare, enabling doctors to visualize and analyze parts of the body that are not otherwise accessible. This technology is essential for various tasks such as detecting diseases at early stages, planning effective treatments, and monitoring how well treatments are working. The roots of medical imaging can be traced back to the discovery of X-rays in 1895 by Wilhelm Röntgen, which revolutionized the field of diagnostic medicine. Since then, the development of advanced imaging modalities, such as MRI, CT, and ultrasound, has significantly expanded the scope and accuracy of medical diagnosis and treatment.

One critical application within medical image analysis is segmentation. Segmentation involves isolating and defining specific areas within an image, such as organs, tissues, or abnormalities. For instance, segmentation helps doctors accurately determine the size, shape, and boundaries of a tumor or organ, enabling precise diagnoses and treatment plans. Historically, segmentation was performed manually, relying on the expertise and precision of radiologists or medical professionals. While effective, this process posed significant challenges. Manual segmentation is inherently time-intensive and laborious, often requiring up to 4 hours per scan for detailed analysis, compared to the 30 minutes or less achievable with AI-based methods [1]. This prolonged process can delay critical diagnoses and treatments, which is especially concerning in time-sensitive cases such as cancer or acute cardiac conditions. Moreover, manual segmentation is prone to variability, as results can differ depending on the individual performing the task, potentially leading to inconsistencies in diagnoses and treatment [2].

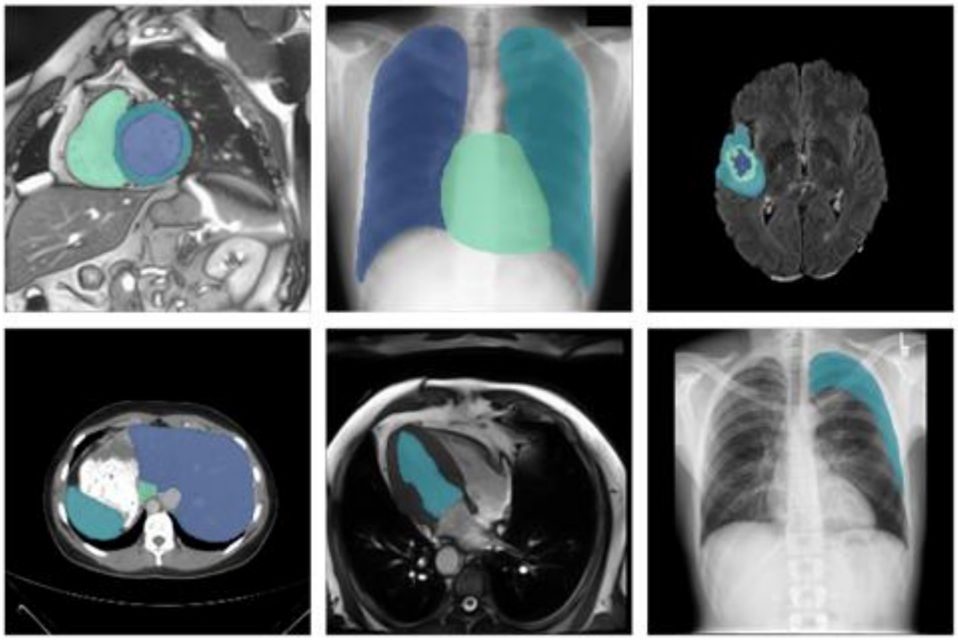


Figure 1.1 Medical Image Segmentation Application [3]

The demand for faster, more reliable, and scalable methods has paved the way for the integration of artificial intelligence (AI) into segmentation workflows. AI has emerged as a transformative solution, automating segmentation processes to improve both efficiency and accuracy. These systems significantly reduce the time required to analyze medical images, allowing healthcare providers to make quicker and more informed decisions. By eliminating inconsistencies caused by human variability, AI ensures a higher level of precision and consistency in segmentation results. This advancement directly addresses the limitations of manual segmentation, ultimately enhancing the quality of patient care and enabling faster diagnosis and treatment, especially in time-sensitive cases.

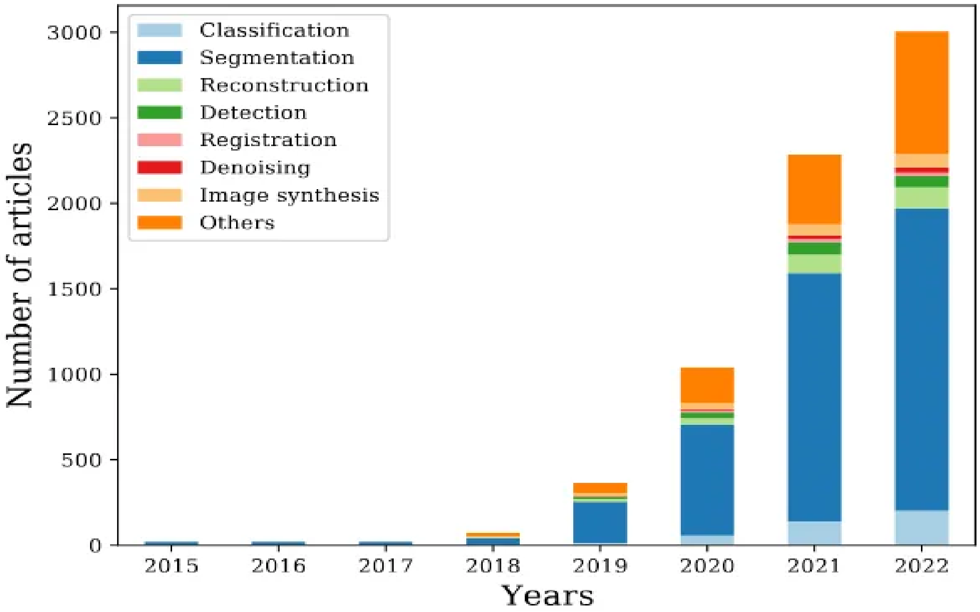


Figure 1.2: Trend in medical imaging research [4]

The increasing focus on medical image segmentation reflects its growing importance in healthcare, as shown in Figure 1.2. Over the last decade, segmentation has risen to prominence in medical imaging research because of its vital applications in diagnostics, treatment planning, and personalized medicine. This upward trend is fueled not only by theoretical advancements but also by the pressing need to overcome real-world challenges in clinical practice. Researchers and practitioners alike recognize segmentation as a critical tool for extracting precise and actionable insights from medical images, thereby bridging the gap between cutting-edge research and practical healthcare needs.

Base on these information, it is necessary for applying artificial intelligence to segment the medical image.

1.2 Objectives

Develop an AI-driven model capable of automatically segmenting medical images with high presion and reliability, assisting doctors in diagnosis and treatment.

1.3 Scope

Platform Constraints: This project is designed to operate on standard personal computers available in hospitals. Data must be transferred from MRI machines to the PC for processing, as the system cannot directly access imaging machines. However, the system can preprocess the transferred data and display segmentation results directly on the PC's screen, ensuring practical usability within the hospital environment.

2D Image Limitation: Due to time constraints and the complexity involved in handling 3D medical images, this project focuses exclusively on 2D image segmentation.

Dataset Scope: The project is specifically focused on cardiac imaging tasks, utilizing the ACDC dataset. It targets the segmentation of three key cardiac structures: the left ventricle (LV), right ventricle (RV), and myocardium (MYO). Other types of medical images, such as those of the lungs or brain, are beyond the scope of this project and cannot be addressed by the current model.

Research Purpose only: This project is developed for research purposes and is not intended for direct clinical deployment. While the system can temporarily run on a hospital's PC to demonstrate its functionality, it currently does not support a user interface or dedicated application.

1.4 Working table

Time	Huynh Thanh Phong	Le Quoc Thinh
6/11 – 14/11	Build Backbone Unet2D	Build Mean Teacher
14/11 – 29/11	Data preprocessing – Data module	Build Bi-directional Copy Paste
29/11 – 15/12	Fine tuning: Mask + Mask size	Fine-tuning: U_weight
15/12 – 18/12	Collect results	Clean code - Comment
18/ 12 – 26/12	Write report + Powerpoint	Write report

Table 1.1 Time and task distribution

II. Related works

2.1 Challenges in Medical Images Segmentation

Annotation Cost

The annotation process in medical segmentation follows a rigorous and systematic protocol to ensure accuracy and consistency. Typically, two medical experts independently annotate the images, with their annotations reviewed by a third expert to identify and resolve discrepancies. Any inconsistencies require the original annotators to revisit and refine their work through multiple iterations until a consensus is reached. This meticulous process guarantees high-quality annotations, which are crucial for training reliable segmentation models.

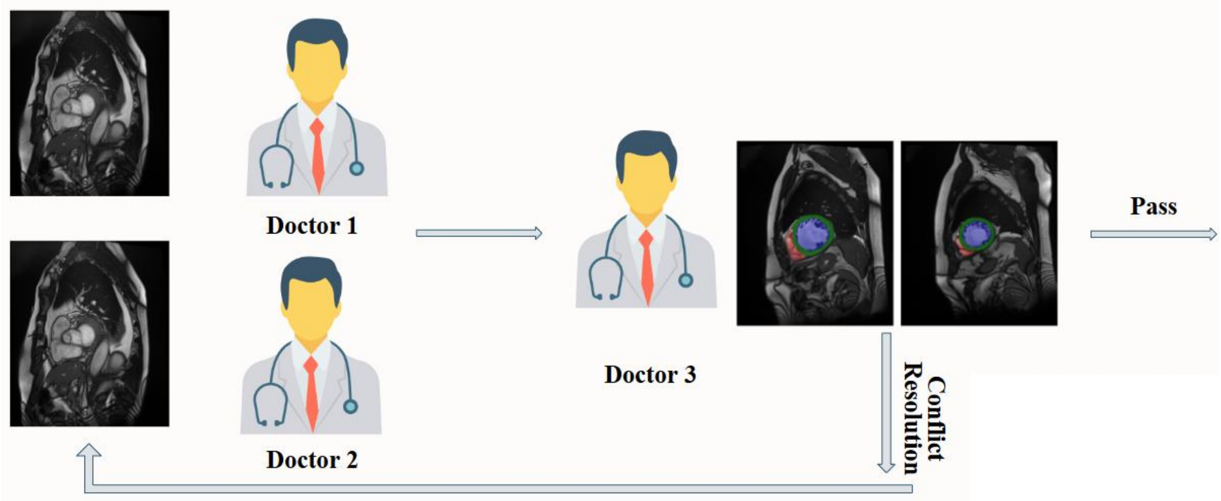


Figure 2.1 An example of annotation protocol in medical image

However, this protocol is both time-intensive and expensive. The need for highly specialized domain experts, such as radiologists or pathologists, significantly increases the financial cost of annotation. Their expertise comes at a premium, and the complexity of tasks like pixel-level labeling for multi-class segmentation adds further to the time and effort required. These challenges make the annotation process a major bottleneck, particularly for large-scale projects with limited resources.

Limit dataset

A limited dataset in medical image segmentation poses significant challenges due to the high cost and effort required for labeling. Annotation relies on medical professionals, such as radiologists or pathologists, for precise, pixel-level annotations, making the process expensive and time-intensive. As a result, only a small fraction of medical imaging data is labeled, leaving most data unused for model training.

This limitation impacts model performance, as small labeled datasets can lead to overfitting and biased models that struggle with rare cases or diverse scenarios. Addressing this issue requires cost-effective strategies, such as focusing annotations on critical samples and simplifying the annotation process. However, the reliance on expert knowledge continues to hinder the creation of large, high-quality datasets, slowing progress in medical image segmentation.

Privacy and Ethics

Privacy and ethics are critical challenges in medical image segmentation and AI in healthcare. Regulations like HIPAA and GDPR ensure data protection but limit access to diverse, large-scale datasets needed for robust models. While anonymization helps, advanced re-identification techniques still pose risks.

Ethically, using medical data requires informed consent and careful handling to avoid violations and maintain trust. Biases in data or models can lead to inequitable treatment outcomes, raising fairness concerns. Balancing compliance with innovation necessitates privacy-preserving techniques, like federated learning, and embedding ethical considerations into workflows.

2.2 General method

2.2.1 Deep Learning

Deep learning has revolutionized medical image segmentation by automating the process, eliminating the need for handcrafted features. Models like convolutional neural networks (CNNs) learn directly from raw imaging data, enabling accurate, efficient, and automatic segmentation across various modalities such as MRI, CT, and ultrasound.

Fully Convolutional Networks (FCNs)

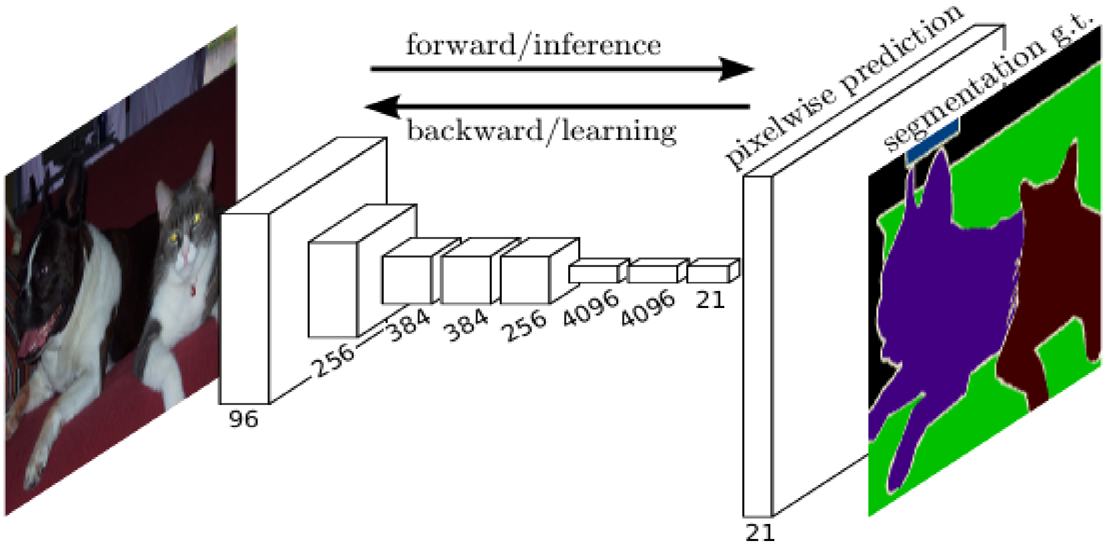


Figure 2.2 Fully Convolutional Networks [5]

Fully Convolutional Networks (FCNs) [cited], introduced by Long et al. in 2015, enable pixel-level predictions for semantic segmentation by replacing fully connected layers with convolutional layers. FCNs consist of an encoder for feature extraction and a decoder for upsampling, with skip connections combining low- and high-level features to recover spatial details. While FCNs balance semantic understanding and spatial precision, they face challenges like the loss of fine details and dependency on input size, highlighting the need for further advancements in segmentation architectures.

U-Net

U-Net, introduced by Ronneberger et al. in 2015, is a landmark architecture for semantic segmentation, particularly in medical imaging. It enhances the Fully Convolutional Network (FCN) framework with a symmetric encoder-decoder structure and skip connections, making it highly effective for small datasets common in medical applications.

U-Net's architecture includes an encoder for capturing high-level features and a decoder for restoring spatial resolution. Skip connections combine low-level spatial details with high-level features, preserving fine-grained information for accurate segmentation.

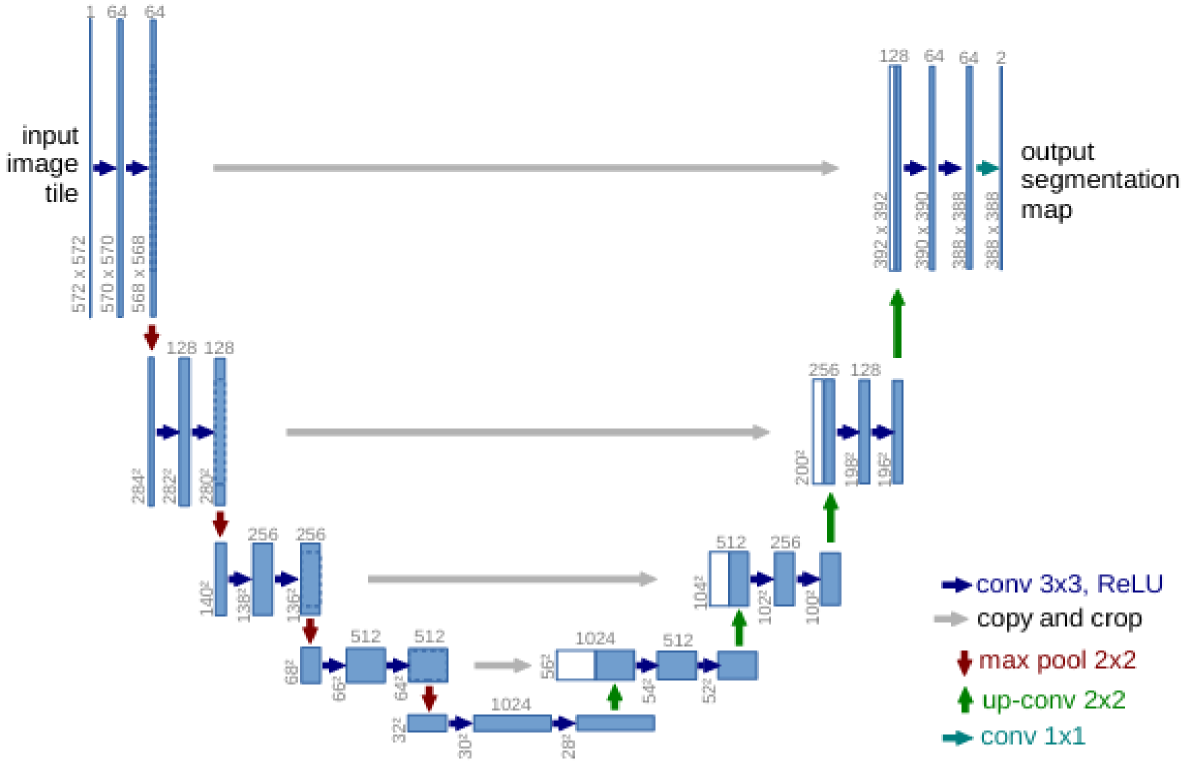


Figure 2.3 U-Net architecture [6]

From Figure 2.3, The U-Net architecture consists of a **contracting path** (left side) and an **expansive path** (right side). The contracting path follows a typical convolutional network structure, with repeated applications of two unpadded convolutions, each foll 3×3 owed by a ReLU activation and a 2×2 max pooling operation with stride 2 for downsampling. At each downsampling step, the number of feature channels is doubled. The expansive path involves upsampling the feature maps, applying a 2×2 up-convolution to halve the number of feature channels, concatenating the upsampled feature map with the cropped feature map from the contracting path, and applying two 3×3 convolutions with ReLU activations. The cropping accounts for the border pixel loss during convolutions. Finally, a 1×1 convolution is applied at the output layer to map each feature vector to the desired number of classes. The network has a total of 23 convolutional layers.

U-Net is an ideal network for medical image segmentation because of its specialized architecture, which combines an encoder-decoder design with skip connections. The encoder captures high-level semantic features, while the decoder restores spatial resolution. Skip connections ensure the preservation of fine details by merging low-level spatial information with high-level semantic features. This design is particularly effective in segmenting complex and small anatomical structures with high precision. Additionally, U-Net performs well with limited annotated data, a common challenge in medical imaging, making it highly suitable for this domain.

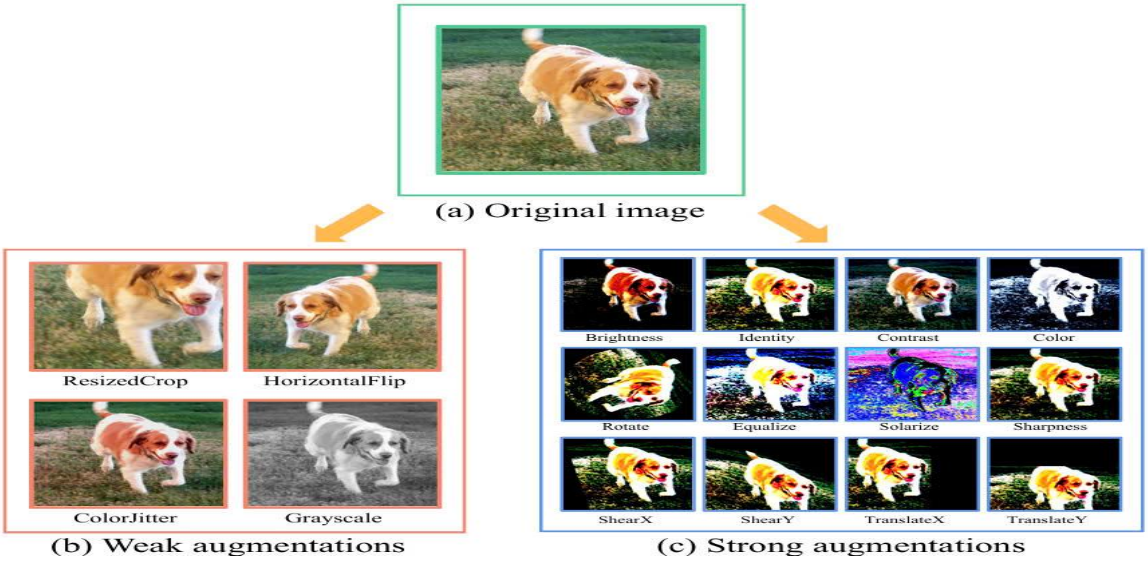
2.2.2 Semi-supervised Learning

Although deep learning has revolutionized medical image segmentation by automating the process and achieving remarkable accuracy, it faces significant challenges related to annotation cost and limited datasets. Medical imaging requires expert annotations, which are time-consuming and expensive, often making large-scale labeled datasets impractical. These challenges restrict the applicability of fully supervised learning, driving the need for alternative approaches. Semi-supervised learning (SSL) addresses these limitations by effectively utilizing both labeled and unlabeled data to train robust segmentation models.

Semi-supervised learning (SSL) aims to utilize a large amount of unlabeled data in conjunction with labeled data to train higher-performing segmentation models. In SSL settings, a dataset is typically divided into a labeled subset, denoted as $D_L = \{(x_i^l, y_i)\}_{i=1}^{M}$ and the unlabeled subset, denoted as $D_u = \{\, x_i^u\}_{i=1}^{N}$, where M << N. The goal is to build a data-efficient deep learning model that combines D_L and D_U to achieve comparable performance to a fully supervised model trained on a fully annotated dataset. By leveraging consistency regularization and pseudo-labeling, SSL ensures improved generalization and reduced dependency on costly labeled data, making it a practical solution for medical image segmentation [7].

Consistency Regularization

Consistency regularization is a key technique in semi-supervised learning that ensures the model produces stable and consistent predictions for the same input under different perturbations or augmentations. This technique aligns predictions across varied transformations of an image, enhancing the model's robustness to input variability and its ability to generalize effectively. Perturbations can include operations such as random cropping, flipping, noise addition, or intensity shifts, which simulate variations in real-world imaging conditions. By enforcing consistency, the model learns to make reliable predictions regardless of these perturbations, an essential characteristic for medical imaging applications where variability in imaging modalities and patient conditions is common.


(a) Original image

(b) Weak augmentations

(c) Strong augmentations

Figure 2.4 Data augmentation used in consistency regularization

Mean-Teacher Framework

The Mean-Teacher framework is a prominent implementation of consistency regularization in SSL. It employs a dual-model architecture consisting of a “student” model and a “teacher” model. The student model learns directly from both labeled and unlabeled data, while the teacher model generates pseudo-labels for the unlabeled data. The teacher’s parameters are updated as the exponential moving average (EMA) of the student’s weights, providing a stable reference for the student during training.

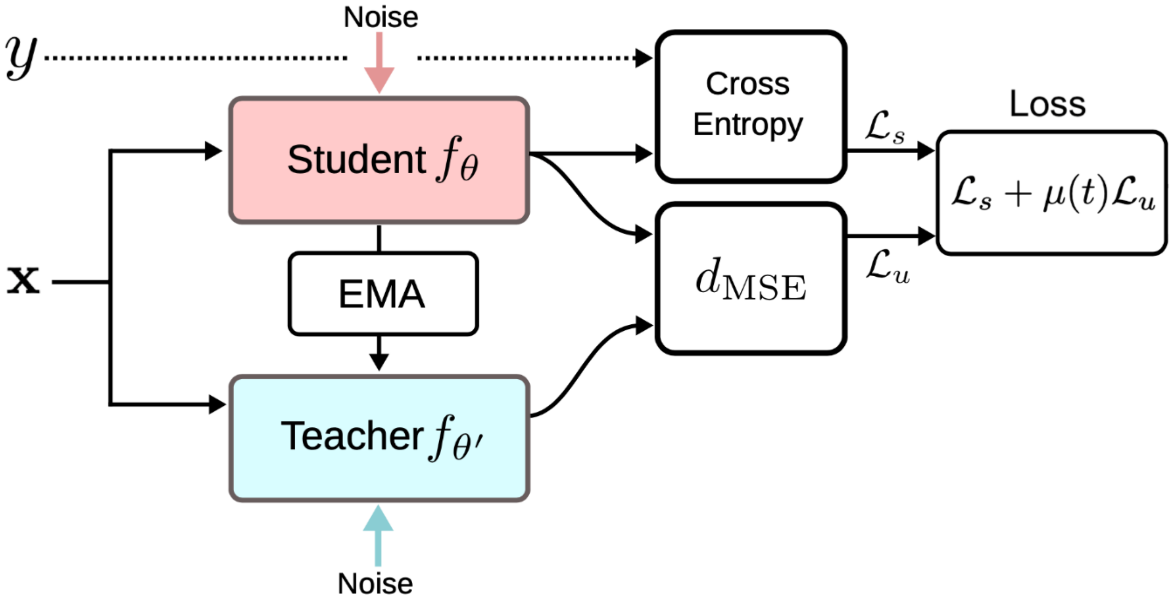


Figure 2.6 Overview of the Mean Teacher framework [8]

This framework enforces consistency by minimizing the difference between the student's predictions and the teacher's pseudo-labels for unlabeled data, ensuring that the student model generates similar predictions for the same input under different perturbations, such as augmentations or noise injections. By aligning predictions across these variations, the model becomes more robust to input variability and generalizes better across diverse imaging conditions. The Mean-Teacher approach has demonstrated superior performance in semi-supervised medical image segmentation tasks by effectively leveraging both labeled and unlabeled data. [9]

2.2.3 Copy-paste method

The Mean-Teacher framework is a leading method in semi-supervised learning for medical image segmentation but faces challenges with confirmation bias [10] caused by reliance on noisy pseudo-labels. While consistency regularization enforces stable predictions through perturbations, it can amplify errors in pseudo-labels, limiting the model's generalization.

Copy-Paste methods [11] address these issues by enhancing data diversity through simple yet effective augmentations, such as copying regions of interest and pasting them onto other images. This approach generates realistic augmented samples without introducing artificial noise, reducing reliance on pseudo-labels and mitigating confirmation bias. In medical image segmentation, Copy-Paste is particularly beneficial for tasks like cardiac segmentation, where labeled data is scarce and preserving anatomical realism is crucial.

By improving data diversity and addressing pseudo-label challenges, Copy-Paste complements semi-supervised frameworks, enhancing segmentation performance in medical imaging.

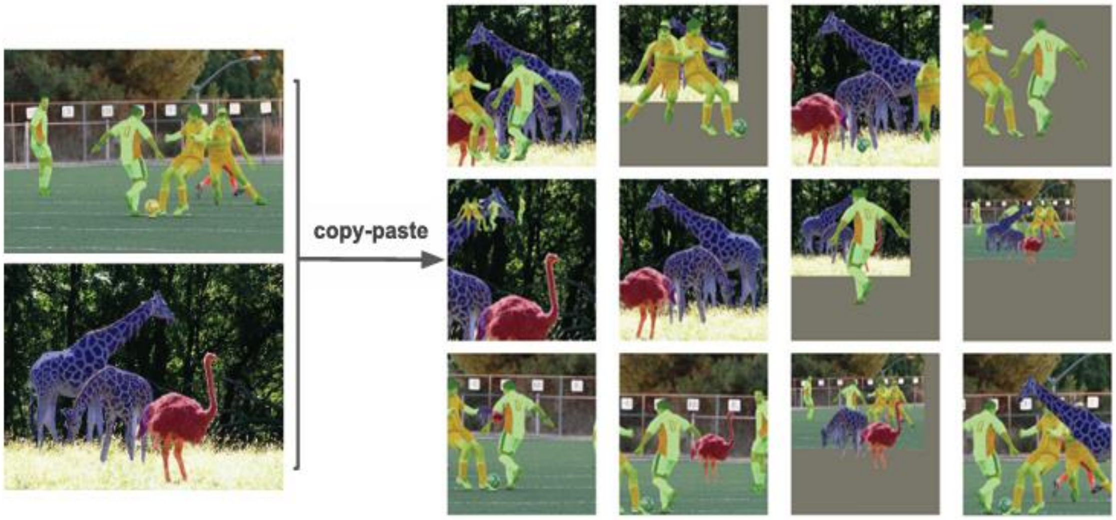


Figure 2.7: Illustration of the Copy-Paste Method for Data Augmentation in Segmentation Tasks

2.2.4 Bi-directional Copy-Paste

Traditional Copy-Paste methods are widely used in medical image segmentation to enhance data diversity by augmenting labeled datasets. However, these methods face significant challenges in semi-supervised learning settings. They often treat labeled and unlabeled data separately during augmentation, leading to an empirical distribution mismatch [12]. This mismatch hinders effective knowledge transfer from labeled to unlabeled data, reducing overall performance.

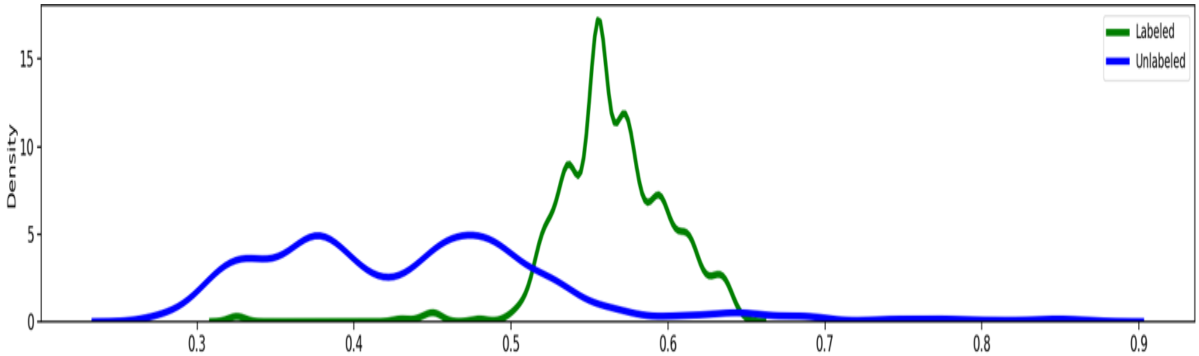


Figure 2.8: Visualization of empirical distribution mismatch (KDE)

To address these limitations, Bidirectional Copy-Paste (BCP) introduces a novel approach that bidirectionally augments both labeled and unlabeled datasets. By pasting labeled images onto unlabeled ones and vice versa, BCP aligns the distributions of both datasets, effectively leveraging ground-truth labels and pseudo-labels for consistent learning. This strategy reduces the reliance on noisy pseudo-labels while promoting better generalization, resulting in significant performance gains. The KDE visualization demonstrates how BCP minimizes distribution mismatch, ensuring a more balanced and robust learning process

2.2.5 Comparision table

Based on the information above, the methods are compared against key features relevant to the design requirements. The table is summarized below:

Feature	FCN [5]	UNet [6]	MT [9]	MT-CP	MT-BCP [12]
Data Requirement	★★☆☆☆	★★★☆☆	★★★★☆	★★★★☆	★★★★★
Robustness to Data Variability	★★★☆☆	★★★★★	★★★★☆	★★★★☆	★★★★★
Integrate Unlabeled Data	★☆☆☆☆	★★☆☆☆	★★★★★	★★★★★	★★★★★
Augmentation Quality	★☆☆☆☆	★★☆☆☆	★★★☆☆	★★★★☆	★★★★★
Computational Cost	★★★★☆	★★★★☆	★★★★☆	★★★☆☆	★★★☆☆
Dice Score	★★☆☆☆	★★★★☆	★★★★☆	★★★★☆	★★★★★
HD (Hausdorff Distance)	★★★☆☆	★★★☆☆	★★★★☆	★★★★☆	★★★★★

Table 2.1: Comparison of segmentation methods

III. Proposed method

3.1 Design requirements

The proposed method is designed with the following requirements in mind:

- **Utilization of Limited Labeled Data:** Perform effectively with only 10% of the dataset labeled (less than 150 images generally), minimizing the need for extensive manual annotation.
- **Integration of Unlabeled Data:** Enhance segmentation performance by leveraging unlabeled data through pseudo-labeling and data augmentation, while mitigating risks like confirmation bias from excessive augmentation.
- **Generalization and Consistent Learning:** Ensure consistent strategies for labeled and unlabeled data to bridge distribution gaps and generate diverse, realistic training examples, improving robustness and generalization.
- **Achieve High Segmentation Accuracy:** Ensure the delivery of accurate and reliable predictions, as reliability is critical in the medical environment. To meet these high standards, the Dice score should surpass 0.85 and the Hausdorff Distance (HD) should remain below 3.0, ensuring the results align with established medical segmentation standards.

In conclusion, these design requirements establish clear and measurable standards for selecting an effective method. Any chosen approach must demonstrate its ability to work effectively with limited labeled data, leverage unlabeled data for improved performance, generate diverse training examples, and achieve high segmentation accuracy as defined by rigorous evaluation metrics.

3.2 Select proposed method

First, considering the utilization of limited labeled data, FCN and UNet are unsuitable due to their reliance on large annotated datasets, which makes them ineffective for this project with only 10% labeled data (approximately 136 images). In contrast, the Mean-Teacher framework (MT) and its extensions, MT-CopyPaste (MT-CP) and MT-BCP, address this challenge by leveraging pseudo-labeling and consistency regularization to effectively incorporate unlabeled data. Therefore, FCN and UNet are excluded from further consideration.

Feature	FCN	Unet	**MT**	**MT-CP**	**MT-BCP**
Data Requirement	★★☆☆☆	★★★☆☆	★★★★☆	★★★★☆	★★★★★
Integrate Unlabeled Data	★☆☆☆☆	★★☆☆☆	★★★★★	★★★★★	★★★★★

Table 3.1 Outperformance of Semi-Supervised Over Supervised Methods on Integrated Data

Secondly, evaluating the integration of unlabeled data, the Mean-Teacher framework struggles with confirmation bias caused by excessive augmentation, leading to degraded pseudo-label quality. MT-CP improves upon this by blending labeled and pseudo-labeled data, but it lacks a balanced approach. MT-BCP resolves these issues through bidirectional augmentation, ensuring realistic and anatomically consistent training samples. This approach enhances segmentation performance, making MT-BCP the superior choice over MT and MT-CP in this aspect.

Feature	MT	**MT-CP**	**MT-BCP**
Robustness to Data Variability	★★★★☆	★★★★☆	★★★★★
Augmentation Quality	★★★☆☆	★★★★☆	★★★★★

Table 3.2 Evaluation of MT, MT-CP, and MT-BCP on Robustness and Augmentation

Next, assessing generalization and consistent learning, MT-CP offers some improvements through simple augmentation strategies. However, MT-BCP excels by employing balanced bidirectional augmentation, which ensures that training data remain realistic and representative of the true distribution. This significantly enhances robustness and reliability, further solidifying MT-BCP as the most suitable method.

After thoroughly assessing the requirements for limited labeled data, effective integration of unlabeled data, and robust generalization, **MT-BCP** stands out as the only method that fully meets all criteria.. It comprehensively addresses these challenges and is identified as the optimal choice for this project.

3.3 Mean Teacher – Bidirectional Copy Paste

3.3.1 Architecture

The proposed architecture, MT-Bidirectional Copy-Paste (MT-BCP), is an extension of the Mean-Teacher framework, tailored for semi-supervised medical image segmentation. The architecture consists of a teacher model and a student model, both built on a shared backbone network, combined with the innovative Bidirectional Copy-Paste (BCP) augmentation module. This design enhances the traditional Mean-Teacher framework by addressing the empirical distribution mismatch between labeled and unlabeled data while leveraging advanced augmentation strategies to improve training diversity and robustness.

Backbone

The architecture employs U-Net as its backbone network, a choice discussed and analyzed in the related work section. The U-Net structure, with its encoder-decoder design and skip connections, is well-suited for capturing multi-scale features and preserving spatial resolution, essential for precise segmentation tasks.

Teacher model

The teacher model generates pseudo-labels for unlabeled data, serving as a supervisory guide for the student model. It is not directly updated through backpropagation; instead, its parameters are updated via an Exponential Moving Average (EMA) of the student model's weights. This EMA update ensures that the teacher model remains stable and produces reliable pseudo-labels, which are crucial for effective semi-supervised learning.

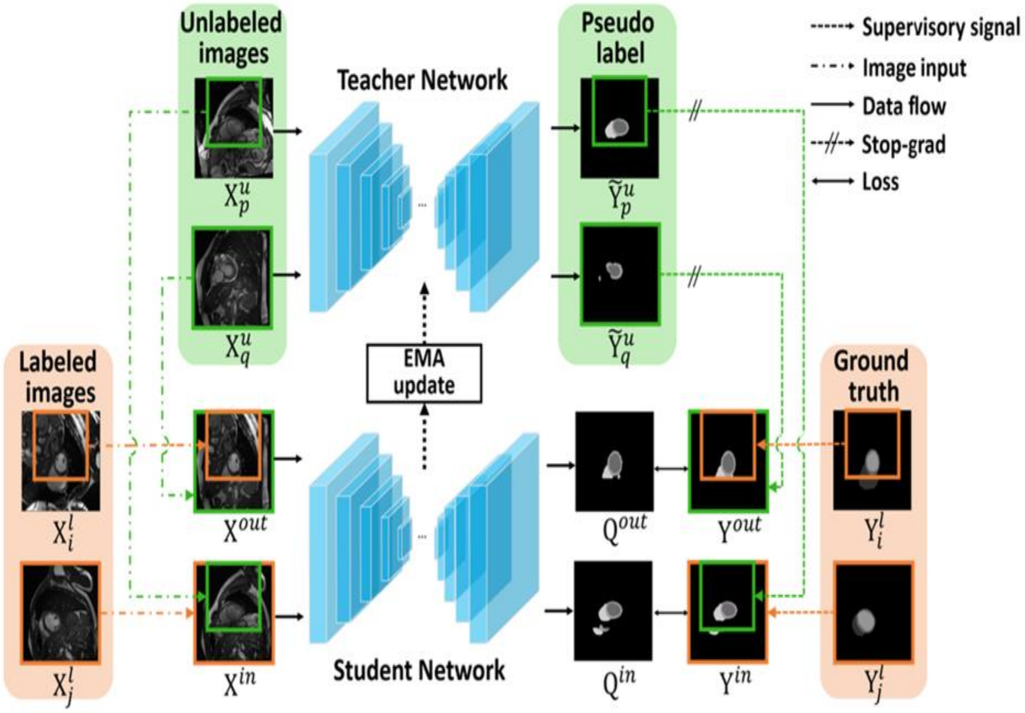


Figure 3.1 Bidirectional copy-paste framework in Mean Teacher architecure [12]

Student model

The student model serves as the trainable component of the architecture, learning from both labeled data and pseudo-labeled data provided by the teacher. The integration of the Bidirectional Copy-Paste module further enhances the student model's training by introducing diverse and realistic augmented samples, allowing it to generalize better across various medical imaging conditions.

3.3.2 Components

Bidirectional Copy-Paste

The Bidirectional Copy-Paste (BCP) module is the core of this method, acting as a data augmentation technique that integrates labeled and pseudo-labeled datasets. It aligns distributions, enhances data diversity, and improves model generalization while reducing dependency on limited labeled data.

First, we generate a zero-centered mask $M \in \{0, 1\}^{1xWxH}$, designed for 2D images. In this mask, the value 0 represents the foreground, and 1 represents the background. The size of the zero-value region is $\beta H\ x\ \beta W$ where $\beta$is a hyper-parameter and $\beta \in \{\ 0, 1\}$. This mask determines the regions to be copied during the augmentation process.

Bidirectional Augmentation: The core idea of BCP is to exchange regions between labeled and pseudo-labeled datasets bidirectionally, ensuring that both images and their corresponding labels are augmented consistently. Labeled regions are pasted into pseudo-labeled images to enrich the pseudo-labeled dataset with high-quality labeled information. Conversely, pseudo-labeled regions are pasted into labeled images to introduce variability into the labeled dataset. This operation mitigates the empirical distribution mismatch between the two datasets, creating balanced and realistic training samples.

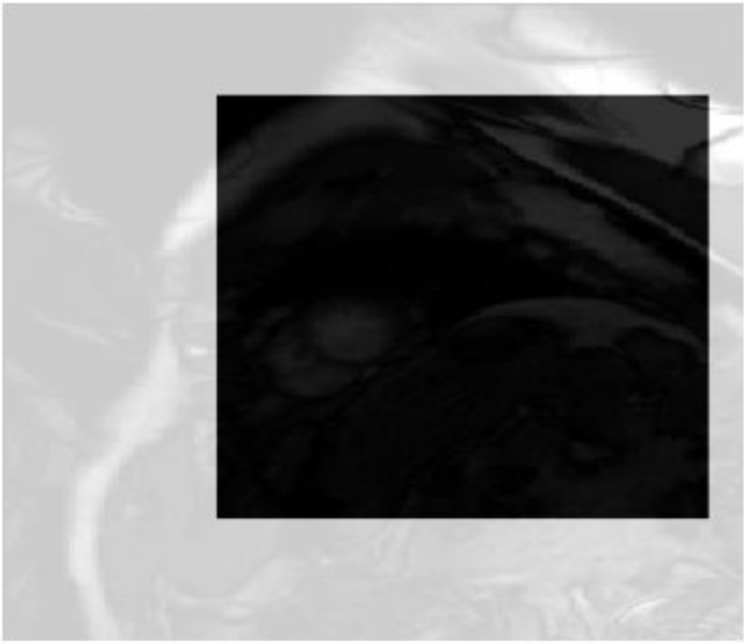


Figure 3.2 Zero-centered mask visualization

The bidirectional copy-paste process can be mathematically represented as follows:

$$X^{in} = X_j^l \odot M + X_p^u \odot (1 - M), \tag{1}$$

$$X^{out} = X_q^u \odot M + X_i^l \odot (1 - M), \tag{2}$$

$$Y^{in} = Y_j^l \odot M + \hat{Y}_p^u \odot (1 - M), \tag{3}$$

$$Y^{out} = \hat{Y}_q^u \odot M + Y_i^l \odot (1 - M). \tag{4}$$

Where $X_i^l, X_j^l \in D^l, i \neq j$, $X_p^u, X_q^u \in D^u, p \neq q$, Where $\hat{Y}_q^u \; and \; \hat{Y}_p^u$ are the pseudo label generate by teacher model,Y_i^l and Y_j^l are the labeled ground truth and $\odot$ mean element-wise multiplication.

Two distinct labeled and unlabeled images are used to ensure input diversity, improving model robustness. By enriching the pseudo-labeled dataset with reliable labeled information and adding variability to the labeled dataset, the BCP process addresses empirical distribution mismatch and ensures anatomical realism. This bidirectional mixing of labeled and pseudo-labeled regions enhances training data diversity and significantly improves model generalization and performance.

Exponential Moving Average (EMA)

The Exponential Moving Average (EMA) [9] is used in the MT-BCP framework to update the teacher model's parameters based on the student model. EMA ensures stability and consistency by applying a weighted average of the student's parameters over time, defined as:

$$\theta_{teacher} \leftarrow \alpha . \theta_{teacher} + (1 - \alpha). \theta_{student} \tag{5}$$

where α is the smoothing factor. EMA reduces noise, improves pseudo-label stability, and enhances generalization by ensuring the teacher model evolves gradually. This mechanism is crucial for generating reliable pseudo-labels, which are combined with labeled data through Bidirectional Copy-Paste (BCP) to improve training efficiency in semi-supervised learning.

Loss function

- **Dice Loss:**

  The Dice Loss is a region-based loss function designed to measure the overlap between the predicted segmentation mask and the ground truth. It ensures that the model captures the correct regions of interest, making it particularly effective in medical image segmentation due to its focus on maximizing overlap.

  The Dice Loss is defined as:

$$L_{Dice} = 1 - \frac{2\sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum g_i + \epsilon}.$$

Where

  - p_i: predicted value for pixel i
  - g_i; ground truth value for pixel i
  - ϵ: a small constant to prevent division by zeros

– **Cross-Entropy Loss (CE Loss)**

The Cross-Entropy Loss is a pixel-wise classification loss that evaluates the accuracy of predictions for each pixel in the image. It penalizes incorrect predictions more heavily, ensuring the model learns to classify each pixel correctly.

The Cross-Entropy Loss is defined as:

$$L_{CE} = \frac{-1}{N}\sum[g_i \log(p_i) + (1 - g_i)\log(1 - p_i)].$$

Where:

  - N: Total number of pixels
  - p_i: Predicted probability for pixel i
  - g_i: Ground truth label for pixel i

- **Mixed Loss:**

  The Mixed Loss combines Dice Loss and CE Loss to leverage the strengths of both. Dice Loss focuses on capturing the global structure of the segmentation, while CE Loss ensures pixel-level classification accuracy. This combination is particularly useful in medical image segmentation, where both local details and global consistency are critical.

  The Mixed Loss is defined as:

$$L_{seg} = L_{Dice} + L_{CE},$$

$$L^{in} = L_{seg}\left(Q^{in}, Y^{in}\right) \odot M + \alpha L_{seg}\left(Q^{in}, Y^{in}\right) \odot (1 - M), \quad (6)$$

$$L^{out} = L_{seg}(Q^{out}, Y^{out}) \odot (1 - M) + \alpha L_{seg}(Q^{out}, Y^{out}) \odot M. \quad (7)$$

Where:

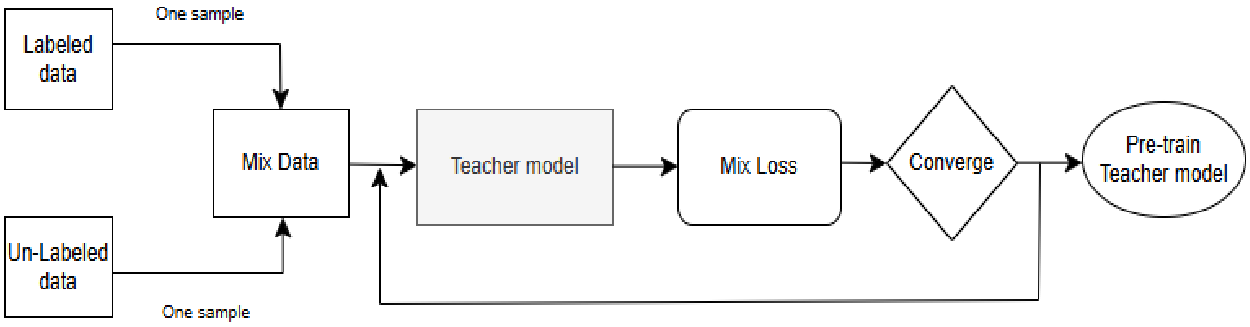
  - Q^{in}, Q^{out}: predictions of model for the mix inputs

- Y^{in} , Y^{out}:grouund truth lables for the mix data
- M: the zero-centered binary mask
- α: the weight that determines the contribution of labeled and pseudo-labeled

3.3.3 Workflow

Pre-train phase:

In the pre-train phase, the teacher model is pre-trained on the labeled data. During this phase, the copy-paste augmentation is applied, where labeled data is copied and pasted into unlabeled data to create a mixed dataset. This augmentation helps the model learn from both labeled and unlabeled data in a supervised manner, ensuring an effective initialization of the teacher model.



Figiure 3.3: Visualize about the pre-train phase workflow

Self-train phase

In the self-train phase, the student model is trained while the teacher model updates its parameters based on Exponential Moving Average (EMA). In this phase, the Bidirectional Copy-Paste (BCP) augmentation is utilized, where labeled data is pasted onto unlabeled data and vice versa to generate diverse training samples. The teacher model generates pseudo-labels for the unlabeled data, and the BCP method is applied to further enhance the training process. This iterative training helps the student model learn effectively from both labeled and pseudo-labeled data.

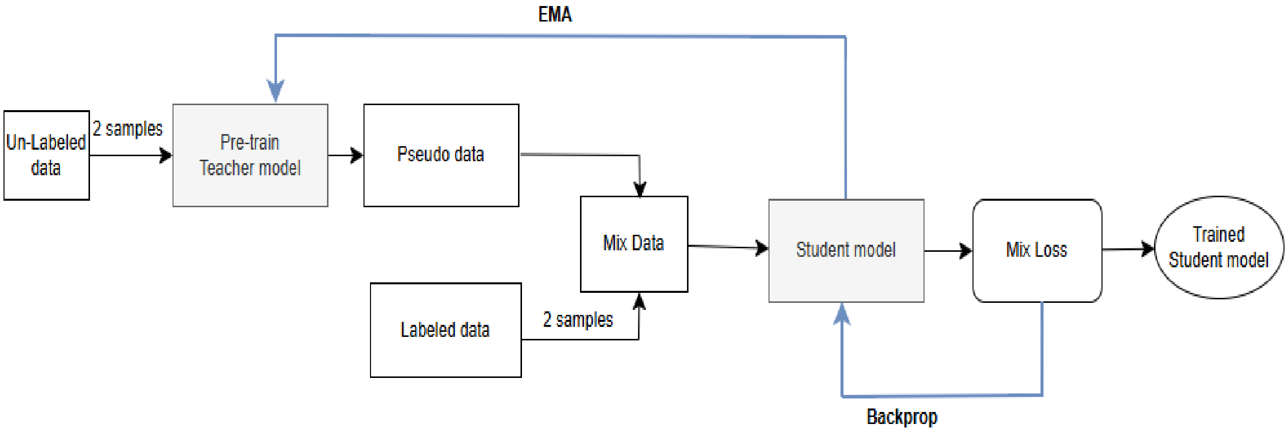


Figure 3.4: Visualize about the self-train phase workflow

IV. Experimental Results

4.1 Dataset

The Automated Cardiac Diagnosis Challenge (ACDC) dataset [13] is a well-established benchmark in medical image segmentation, specifically for cardiac MRI analysis. This dataset, introduced as part of the ACDC challenge, promotes advancements in automated heart disease diagnosis using deep learning. It includes short-axis cine-MRI scans from 100 patients across diverse clinical conditions. The dataset provides annotations for key anatomical structures: the left ventricle (LV), right ventricle (RV), myocardium (MYO), and background.

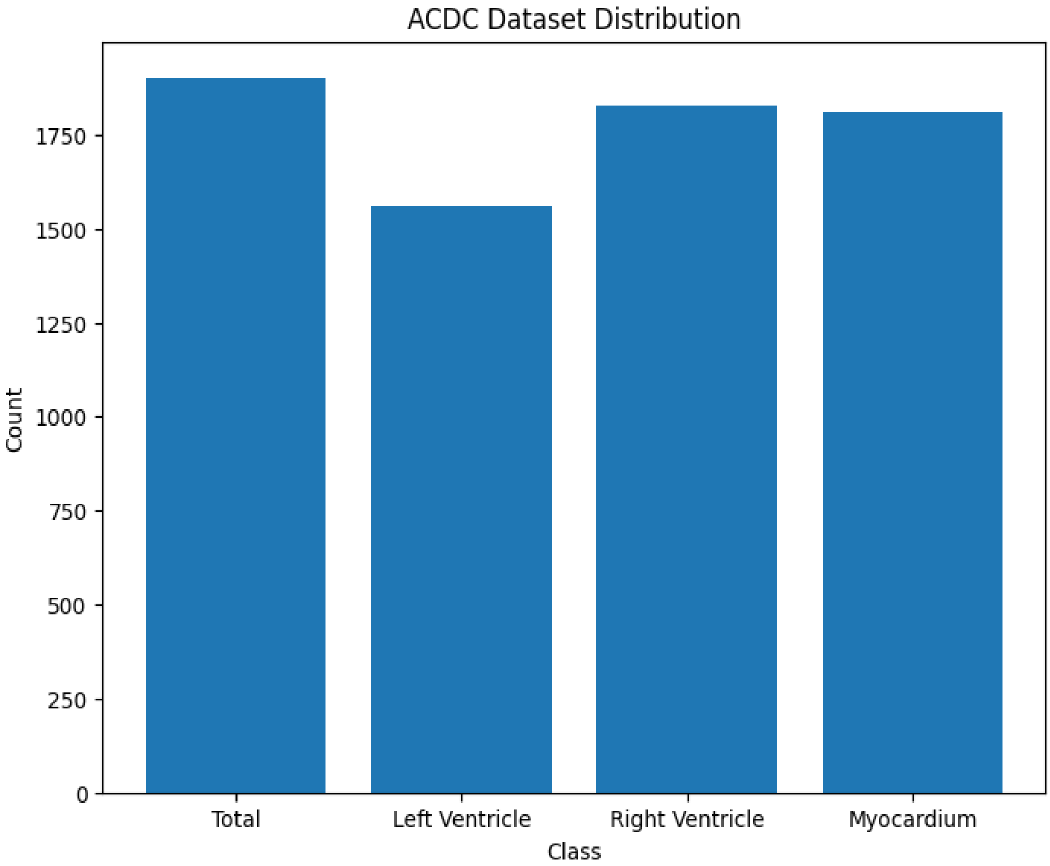


Figure 4.1 ACDC Dataset Distribution

Preprocessing

Before training, the ACDC dataset underwent preprocessing to prepare it for model input. Each MRI scan was sliced into 2D images, and the data was converted into the HDF5 format for efficient storage and access during training. This step simplifies data handling and ensures compatibility with the training pipeline.

Dataset splitting

The dataset will be split into three subsets:

- **Training Set:** 70% of the data (70 patients' scans) for model training.

- **Validation Set:** 10% of the data (10 patients' scans) for hyperparameter tuning.
- **Test Set:** 20% of the data (20 patients' scans) for final model evaluation.

This distribution ensures a balanced evaluation of the model's performance.

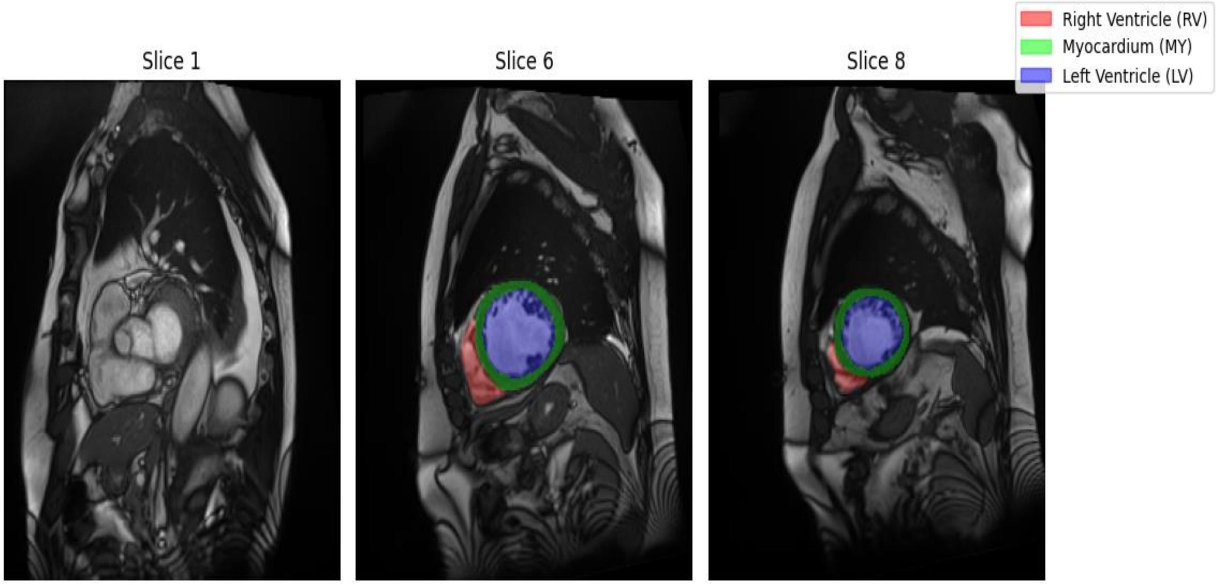


Figure 4.2 Visualization of Cardiac MRI Slices with Annotated Segmentation

Data augmentation

To enhance model robustness and generalization while minimizing the risk of confirmation bias, simple augmentation techniques were applied during training.

- **Random Flip:** Horizontal and vertical flips were applied randomly to increase image diversity.
- **Random Rotation:** Images were rotated randomly within a range of -20° to 20°, simulating variations in cardiac orientations while preserving anatomical structures.

4.2 Evaluation metrics

In this project, a combination of evaluation metrics is used to comprehensively assess the performance of the segmentation model. Each metric serves a specific purpose, ensuring a robust evaluation of both overlap accuracy and boundary precision in the segmented results:

Labeled-to-Total Ratio

Definition: a metric that quantifies the proportion of labeled samples in the entire dataset

Formula:

$$R = \frac{Labeled}{Labeled + Unlabeled}$$

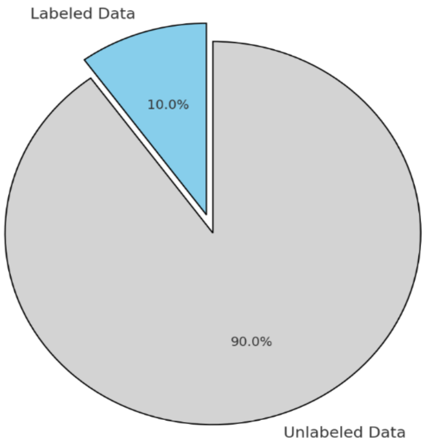


Figure 4.3 Proportion of Labeled and Unlabeled Data in the Dataset

Meaning: This metric is crucial for meeting the design requirements of this project, ensures labeled data is smaller than unlabeled data, aligning with the semi-supervised framework. It supports efficient training by leveraging abundant unlabeled data while minimizing reliance on labeled samples.

Dice Score

Definition: The Dice Score measures the overlap between the predicted segmentation and the ground truth segmentation. It evaluates the similarity between the two sets and ranges from 0 (no overlap) to 1 (perfect overlap).

Formula:

$$\boldsymbol{Dice} = \frac{\mathbf{2} \times |\boldsymbol{A} \cap \boldsymbol{B}|}{|\boldsymbol{A}| + |\boldsymbol{B}|}$$

where A is the predicted segmentation and B is the ground truth segmentation.

Meaning**:** The Dice Score provides a robust evaluation of the overlap between predicted and ground truth segmentation. For the ACDC dataset, it is particularly useful for monitoring the model's performance in accurately segmenting small or irregular cardiac structures.

Intersection over Union (IoU)

Definition: Also known as the Jaccard Index, it measures the overlap between the predicted and ground truth segmentation areas, similar to the Dice coefficient, but with a different calculation approach.

Formula:

$$\textbf{IoU} = \frac{|\boldsymbol{A} \cap \boldsymbol{B}|}{|\boldsymbol{A} \cup \boldsymbol{B}|}$$

Where:

- A: the predicted segmentation
- B: the ground truth segmentation

Meaning**:** IoU offers a stricter overlap evaluation than the Dice Score, making it valuable for validating the precise segmentation of complex boundaries.

Hausdorff Distance (HD)

Definition: The Hausdorff Distance measures the maximum distance from a point on the boundary of the predicted segmentation to the nearest point on the boundary of the ground truth segmentation

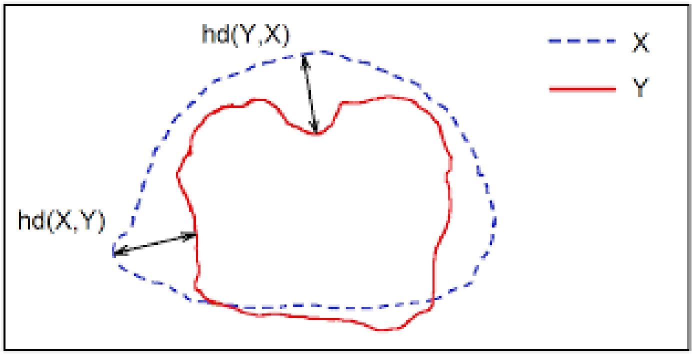


Figure 4.4 Visualization of Hausdorff Distance [14]

Formula:

$$\boldsymbol{HD(A,B)} = \boldsymbol{max(sup}\ (\boldsymbol{inf}\ ||\boldsymbol{a} - \boldsymbol{b}||), \boldsymbol{sup}\ (\boldsymbol{inf}\ ||\boldsymbol{a} - \boldsymbol{b}||))$$

Where:

- $a \in A\ b \in B\ \ and\ b \in B\ a \in A$

Meaning**:** This is critical for assessing how well the model handles edge cases and small structural differences in the cardiac segmentation task. This metric highlights areas where the BCP method can improve boundary precision

Average Surface Distance (ASD)

Definition: ASD provides an average measure of the distance between boundaries of the predicted and ground truth segments

Formula:

$$\boldsymbol{ASD(A,B) = (\sum_{a\in A} \min_{b\in B}|a-b| + \sum_{b\in B} \min_{a\in A}|a-b|)\frac{1}{|A|+|B|}}.$$

Meaning: ASD provides a balanced measure of boundary accuracy, offering insights into the average deviation between predicted and ground truth boundaries.

<u>Kernel Density Estimation</u>

Kernel Density Estimate (KDE) is used to analyze the distribution of segmentation errors, providing a non-parametric estimation of the dataset's probability density to understand deviations from the ground truth.

Definition**:** Kernel Density Estimate (KDE) evaluates the distribution of data points by smoothing individual data contributions across a defined bandwidth.

Formul

$$\boldsymbol{f(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{x-x_i}{h}}.$$

Where:

- $f\,(x)$: The estimated density function at point x .
- n : The number of data points.
- h : The bandwidth, controlling the smoothness of the estimate.
- K : The kernel function (commonly Gaussian).

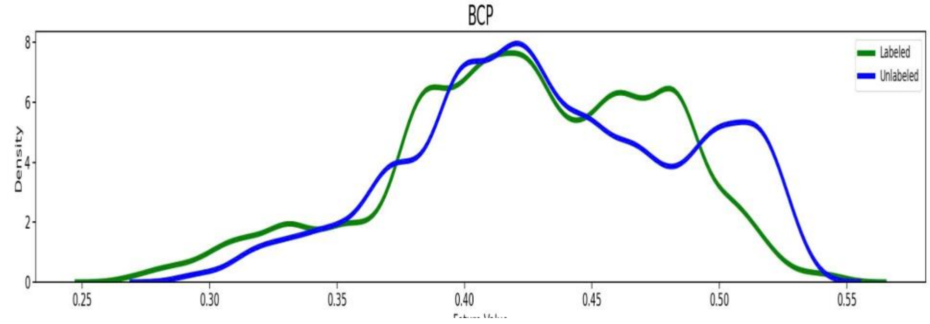


Figure 4.5 Illustration of Kernel Density Estimation Plot

In this project, we address the **empirical distribution mismatch** between the large amount of unlabeled data and the small amount of labeled data by utilizing Kernel Density Estimation (KDE). KDE helps to align the distributions by analyzing and smoothing the differences, ensuring more consistent and effective model training.

4.3 Implement

Implementation Environment

The project was implemented using **Python 3.12 and Pytorch** for their compatibility with the latest tools. Due to the resource-intensive nature of the training process, which exceeded the capabilities of a local laptop, the training was conducted on **Kaggle**. Using the **NVIDIA Tesla P100 GPU** available on Kaggle, the training process took approximately 4 hours and 30 minutes.

Hyper-parameters

The table below summarizes the hyperparameters used for the best-performing model. These settings were carefully tuned to optimize the model's performance:

Hyper-parameters	Value	Description
Batch size	24	Number of samples processed in one iteration.
Labeled batch size	12	Number of labeled samples in each batch.
Learning rate	0.01	Initial learning rate for model optimization.
U_weight (α)	0.5	Weight for the loss of unlabeled data.
Mask size (β)	0.5	A scaling factor multiplied by the image dimensions to calculate the mask size
Number of labeled data	136	Number of labeled data
Pre-train iterations	15000	Number of iterations for pre-train
Self-train iterations	30000	Number of iterations for self-train
Mask strategy	Center mask	Strategy for generating masks during training.

Table 4.1: Hyper-parameters for best model

4.4 Results

4.4.1 Quantitative

Before presenting the best results and comparing them with other semi-supervised methods, it is essential to first evaluate the impact of fine-tuning two critical hyperparameters, β (mask size) and α (U_weight). Understanding the influence of these hyperparameters on model performance provides valuable insights for optimizing the Bidirectional Copy-Paste (BCP) method and ensuring robust segmentation outcomes.

Beta (β) Comparison

Exploration focused on adjusting the rectangular mask size (β) to optimize model performance and refine the masking strategy. The results are summarized in Table 4.2, showing the Mean Dice Coefficients for different values of β on the validation and test sets.

The results indicate that values of β between 0.3 and 0.5 provide the best balance between model performance and effective masking. Specifically, $\beta = 0.5$ achieved the highest Mean Dice Coefficients of **0.887** on the validation set and **0.880** on the test set, demonstrating its suitability for generating optimal mask sizes.

Mean Dice Coefficient of model with different β

Beta (β)	Valid Set	Test Set
2/3	0.880	0.867
1/2	**0.887**	**0.880**
1/3	0.895	0.879
5/6	0.618	0.847

Table 4.2 Mean Dice coefficient of model with different β

From Table 4.2, we can analyze that the value of β determines the size of the mask, which directly affects the effectiveness of the Copy-Paste strategy. Higher β values produce larger masks, reducing the augmentation's meaningfulness as the copied region becomes too large, resembling the original image. This diminishes the diversity introduced by the Copy-Paste strategy, rendering it less effective for improving model performance.

U-weight Comparison

The U-weight (α) controls the influence of unlabeled data on the training process, making it a critical hyperparameter in semi-supervised learning. From the results in Table 4.3, it is evident that different U-weight values significantly impact model performance.

U-Weight	Valid Set	Test Set
3/2	0.882	0.879
1/2	**0.888**	**0.883**
1/3	0.886	0.874

Table 4.3 Mean Dice Coefficient of model with different U-Weight

A lower U-weight ($\alpha = 1/3$) limits the contribution of unlabeled data, leading to suboptimal performance. Conversely, a higher U-weight ($\alpha = 3/2$) overemphasizes the pseudo-labeled data, which may contain noise, thereby reducing segmentation accuracy. The best performance was achieved with $\alpha = 1/2$, yielding the highest Mean Dice Coefficients of **0.888** on the validation set and **0.883** on the test set. This indicates that a balanced U-weight ensures effective utilization of unlabeled data while avoiding the propagation of noisy pseudo-labels, thereby enhancing the model's generalization ability.

Result of the best models

After fine-tuning the hyperparameters (β and α), the performance of the proposed Bidirectional Copy-Paste (BCP) method was compared to state-of-the-art (SOTA) approaches in medical image segmentation.

As we can see in the Table 4.4 The proposed Bidirectional Copy-Paste (BCP) method successfully addresses the design requirement of achieving high segmentation accuracy. It achieved a **Dice coefficient of 88.29**, the highest among models trained with 10% labeled data, along with an **IoU of 79.41**. Additionally, BCP demonstrated excellent precision in boundary delineation, with the lowest **Hausdorff Distance (HD) of 2.47** and a minimal **Average Surface Distance (ASD) of**

0.79. These metrics highlight BCP's capability to produce highly accurate and reliable segmentation results, meeting the project's primary objective.

Method	Labeled data	Dice	IoU	HD	ASD
UNet [12]	10%	79.41	68.11	9.35	2.70
UNet [12]	100%	91.44	84.59	4.30	0.99
UA-MT [12]	10%	81.65	70.64	6.88	7.75
SASSNet [12]	10%	84.50	74.34	5.42	6.06
BCP	10%	**88.29**	**79.41**	**2.47**	**0.79**

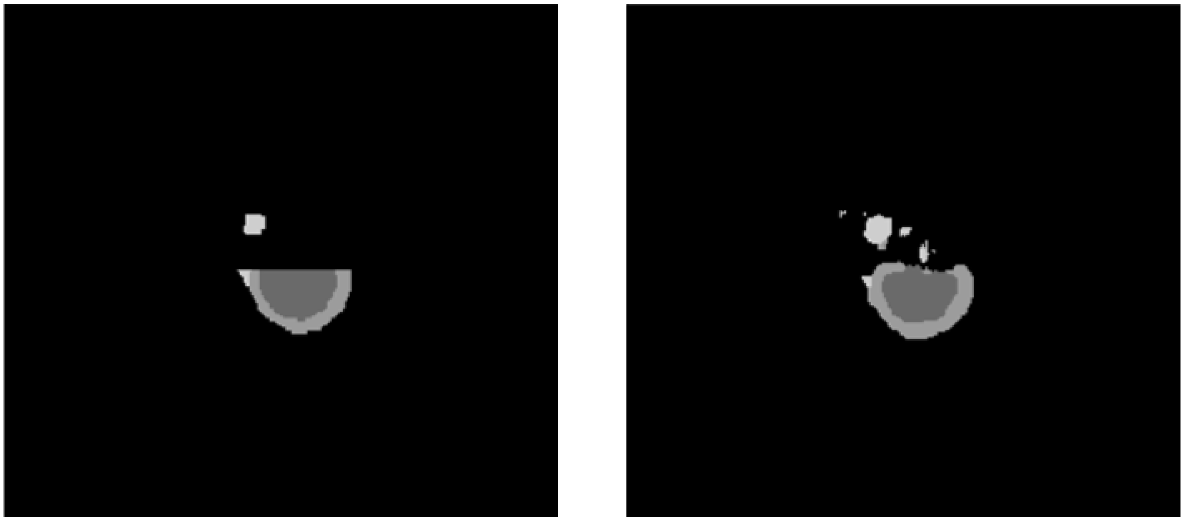
Table 4.4: Comparision with SOTA method on ACDC dataset

Compared to other semi-supervised approaches, BCP outperformed UA-MT and SASSNet in all key metrics, demonstrating superior performance in leveraging limited labeled data. Even when compared to the fully supervised UNet trained on 100% labeled data, BCP significantly narrowed the performance gap, achieving comparable results with only 10% of labeled data.

The results confirm that the BCP method meets the design requirement of achieving high segmentation accuracy while utilizing limited labeled data. By outperforming other semi-supervised methods and nearing fully supervised performance, BCP proves to be a robust and scalable solution for medical image segmentation, addressing challenges in annotation-constrained scenarios.

4.4.2 Qualitative

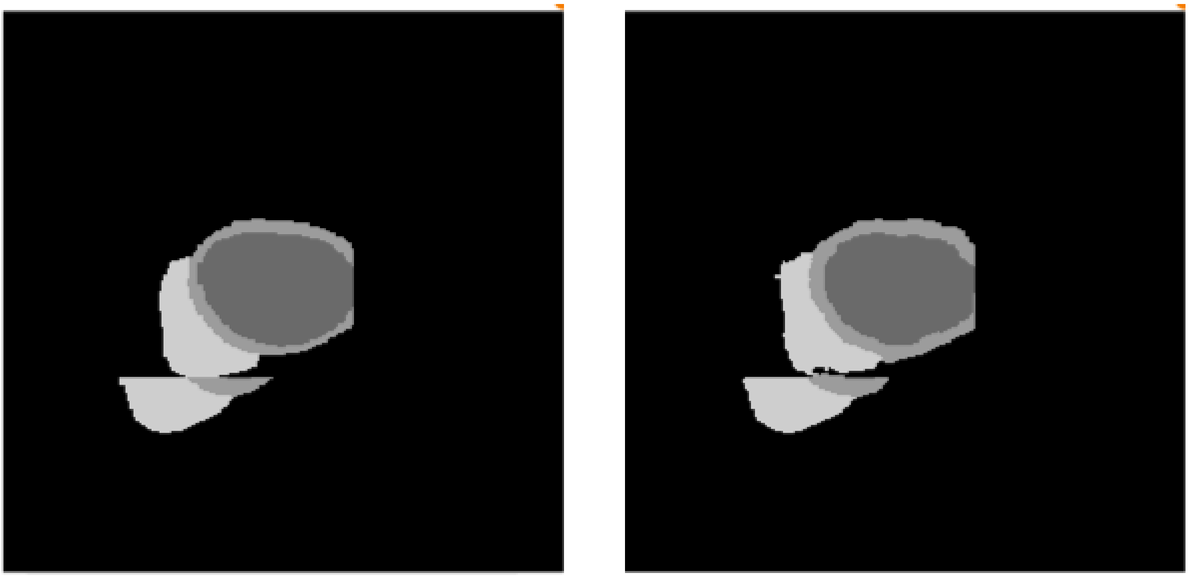
The pseudo label generate by teacher model can be visualized:



(a) Ground Truth (b) Pseudo label

Figure 4.6 Comparison of Ground Truth and Pseudo Labels

The prediction of student model on the Mix data:

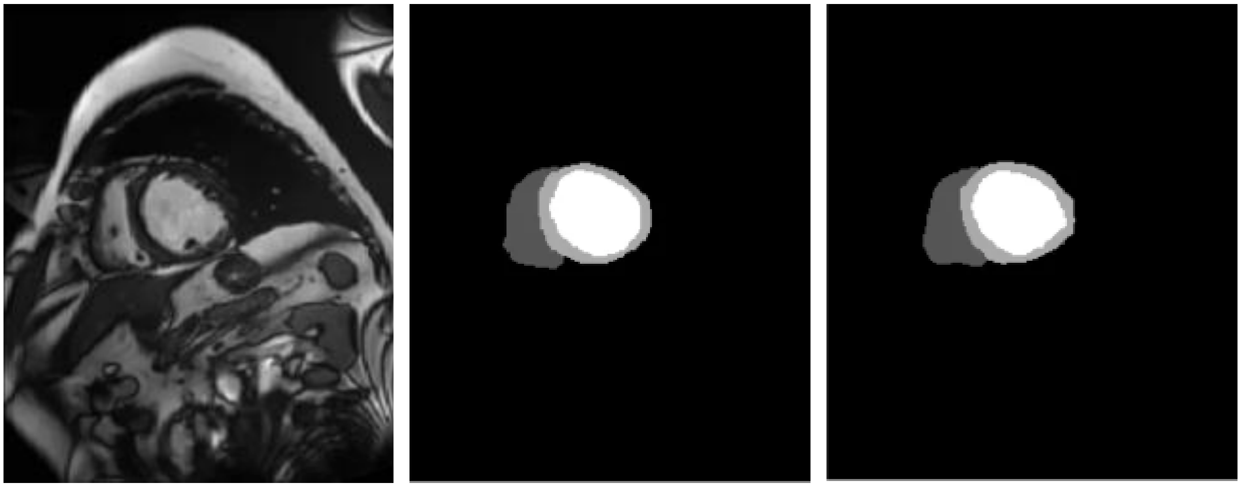


(a) Ground truth (b) Prediction

Figure 4.7 Comparison of Ground Truth and Student Model Predictions on Mixed Data

The predictions from the student model are shown in Figure 4.7. Compared to the pseudo-labels, the student model produces more accurate and complete segmentations that closely align with the ground truth.

Comparison of the model's predictions with the ground truth for test set observations is visualized below:



(a) Input image (b) Ground Truth (c) Model prediction

Figure 4.8 Visualization of Model Predictions Compared to Ground Truth on the Test Set

From the model predictions, we can conclude that the proposed approach successfully meets the design requirement of achieving high segmentation accuracy. The improvements in segmentation quality validate the robustness of the BCP method in handling noisy pseudo-labels and leveraging limited labeled data effectively.

<u>KDE analysis</u>

One of the purposes of the proposed method in this project is to address the empirical distribution mismatch caused by the large amount of unlabeled data:

From Figure 2.8, we can indicate that:

- The Kernel Density distribution shows a noticeable mismatch between the labeled (green) and unlabeled (blue) data distributions. This highlights the **empirical distribution mismatch** problem, where the features of labeled and unlabeled data differ significantly.
- Such a gap can hinder the model's ability to generalize effectively, as the model struggles to align these distributions during training.

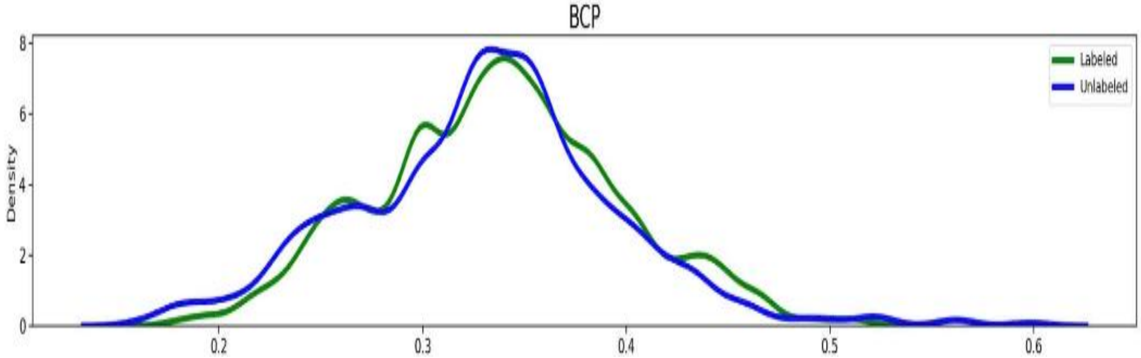


Figure 4.9 The Kernel Density Estimation of BCP MT (our)

From Figure 4.9, we can take some observations:

- The kernel density distribution of our BCP method demonstrates a much closer alignment between labeled and unlabeled data distributions. The overlap between the green (labeled) and blue (unlabeled) curves indicates that the proposed BCP approach successfully mitigates the empirical distribution mismatch.
- This alignment suggests improved consistency in feature representation between labeled and unlabeled data, contributing to better semi-supervised learning performance.

The results demonstrate that our BCP method significantly reduces the distribution mismatch between labeled and unlabeled data. This improved alignment suggests that our approach effectively integrates unlabeled data into the training process, enhancing overall model performance and reliability.

4.4.3. Limitations

The qualitative results from Figure 4.6, 4.7 highlight a key limitation of the pretrained model used for generating pseudo labels. While the pseudo labels aim to provide supervision for unlabeled data, they are often **imprecise and contain noise**, as seen in the predictions. The generated pseudo labels exhibit inaccuracies, including incomplete segmentation and false positive regions, particularly in areas where the model struggles to correctly delineate boundaries or smaller structures. These inconsistencies can negatively impact the self-training process, as the model relies on these pseudo labels for learning. The presence of noise in the pseudo labels may lead to the propagation of errors during training, reducing the overall segmentation accuracy and robustness of the model.

One more barrier, while the ACDC dataset provides valuable 2D cardiac MRI slices for segmentation, relying solely on 2D data introduces inherent limitations. One significant drawback is the lack of spatial continuity between slices, which can lead to fragmented and inconsistent segmentation results, particularly for complex or elongated structures. In 3D medical imaging, context across adjacent slices is crucial for capturing volumetric relationships and ensuring accurate boundary delineation. The absence of this information in 2D data restricts the model's ability to fully leverage the anatomical context, resulting in potential segmentation errors. Additionally, the limited spatial resolution in 2D data can make it challenging to resolve finer details or small anatomical structures. Addressing these limitations requires transitioning to 3D data to capture the full spatial context and improve segmentation accuracy.

V. Conclusion

5.1. Conclusion

This project provided a comprehensive overview of the **Bidirectional Copy-Paste (BCP)** method, a novel approach for improving medical image segmentation. BCP addresses multiple challenges in semi-supervised learning, including the empirical distribution mismatch between labeled and unlabeled data, while also enhancing model robustness and segmentation accuracy. By utilizing advanced masking strategies, pseudo-labeling techniques, and self-training frameworks, the method leverages the potential of unlabeled data to improve performance. Applied to the ACDC dataset, the BCP method demonstrated its effectiveness in overcoming key limitations in medical image segmentation, paving the way for more robust and accurate solutions in clinical applications.***Advantages:***

- **Improved Distribution Alignment**: The BCP method demonstrated effectiveness in narrowing the distribution gap between labeled and unlabeled data, as evidenced by better kernel density alignment.
- **Addressing Annotation Cost and Dataset Limitation**: By leveraging pseudo-labeling and self-training, the method reduced the dependency on manually labeled data, effectively mitigating the high annotation cost and dataset limitation often associated with medical image segmentation tasks.
- **Enhanced Model Performance**: Experiments with optimal β mask sizes and U-weight adjustments resulted in significant improvements in segmentation metrics like Dice Coefficient and IoU.

Disadvantages:

- **Noise in Pseudo Labels**: The pretrained model generated pseudo-labels with noise and inaccuracies, leading to suboptimal training in some scenarios.
- **Limited 2D Data Context**: The current implementation on 2D data limits the model's ability to fully leverage spatial continuity present in medical imaging datasets.

5.2. Future Work

To address the noise and inaccuracies in pseudo labels generated by the pretrained model, we propose the following solutions:

<u>*Increase Pretrain Iterations:*</u>

- By increasing the number of iterations during the pretraining phase, the model can learn more robust and accurate representations. This improvement in the pretrained model's segmentation quality will lead to higher-quality pseudo labels, reducing noise and errors during the self-training phase.

<u>*Implement Multi-Teacher or Multi-Student Framework:*</u>

- Leveraging methods like multiple students or teachers can be introduced to enhance diversity in pseudo labels. This approach helps mitigate biases and errors inherent in a single-teacher or single-student framework by promoting diverse pseudo-label generation and improving consistency in predictions.
- For example, using structurally different students or teachers ensures varied perspectives, enabling the correction of biases through cross-consistency learning and discrepancy-informed correction.

Extend the method to 3D data:

- Training on 3D data allows the model to produce smoother and more consistent pseudo labels across slices, reducing noise and inaccuracies. This is especially beneficial for self-training, as better pseudo labels lead to improved supervision and segmentation performance.

References

[1] R. A. Alvarez et al., "Evaluation of a New Multimodal Method for Early Detection of Alzheimer's Disease: Combining Structural MRI, FDG-PET, and Cognitive Tests," *Molecular Imaging and Biology.*

[2] F. Yepes-Calderon and J. G. McComb, "Manual Segmentation Errors in Medical Imaging. Proposing a Reliable Gold Standard," *Communications in computer and information science*, pp. 230–241, Jan.

[3] S. M. Hooper, *Label-Efficient Machine Learning for Medical Image Analysis*, Ph.D. dissertation, Dept. of Electrical Engineering, Stanford Univ., Stanford, CA, USA, June 2023.

[4] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical Image Segmentation Review: The Success of U-Net

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. doi: 10.48550/arXiv.1411.4038.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," presented at the *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. doi: 10.48550/arXiv.1505.04597.

[7] R. Jiao, Y. Zhang, L. Ding, B. Xue, J. Zhang, R. Cai, and C. Jin, "Learning with Limited Annotations: A Survey on Deep Semi-supervised Learning for Medical Image Segmentation," *arXiv preprint arXiv:2207.14191*, 2022.

[8] L. Weng, "Semi-Supervised Learning: An Overview," *Lil'Log*, Dec. 5, 2021.

[9] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *arXiv preprint arXiv:1703.01780*, 2018.

[10] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning," *arXiv preprint arXiv:1908.02983*, 2020.

[11] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, "Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation," *arXiv preprint arXiv:2012.07177*, 2021.

[12] Y. Bai, D. Chen, Q. Li, W. Shen, and Y. Wang, "Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[13] "Automated Cardiac Diagnosis Challenge (ACDC) Dataset," *CREATIS Laboratory*, [Online].

[14] S. Jadon, "A survey of loss functions for semantic segmentation," *arXiv preprint arXiv:2006.14822*, Jun. 2020.