Medical Image Segmentation with Minimal Labeling Effort: How Far Can We Push the Limits?

Yizhe Zhang

School of Computer Science and Engineering
Nanjing University of Science and Technology
zhangyizhe@njust.edu.cn

Abstract

We demonstrate for the first time that a medical image segmentation model can achieve near fully supervised performance using only a single annotated image and abundant unlabeled data. We present MedSMILE, a novel framework that synergistically integrates transductive and inductive learning for this extreme one-label semi-supervised setting. Its core novelty lies in an iterative loop where a foundation model both bootstraps and refines pseudo-labels for an inductive segmentation model. This process begins with the foundation model performing transductive inference to generate an initial set of pseudo-labels for the unlabeled data pool. This bootstraps an iterative self-training process where the segmentation model is trained and used to generate progressively better labels, with an inter-round refinement step that re-leverages the foundation model to correct errors in uncertain predictions. Experiments on seven datasets across four modalities show MedSMILE recovers 90%–95% of the fully supervised Dice score while decisively outperforming existing semi-supervised techniques that require substantially more annotation. MedSMILE sets a new standard for label-efficient learning in medical image segmentation.

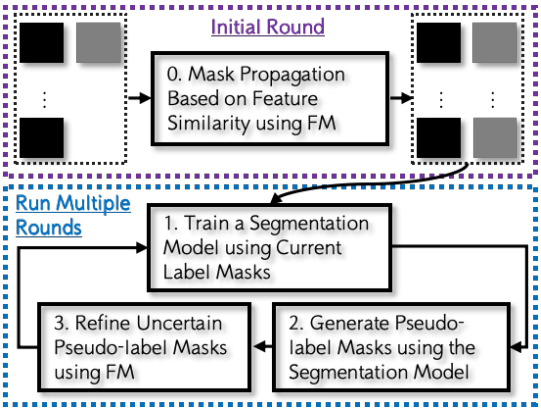

Figure 1: A high-level view of the MedSMILE method, where black blocks represent input images and gray blocks are their segmentation masks. The process uses as few as one labeled image and outputs a trained segmentation model.

Introduction

Deep learning models have demonstrated state-of-the-art performance on medical image segmentation tasks (Yao et al. 2024). However, their success typically hinges on the availability of large datasets with pixel-level annotations, which are notoriously expensive and time-consuming to acquire due to the need for expert clinical knowledge (Awudong et al. 2024; Chen et al. 2020a). This annotation bottleneck significantly hinders the development and deployment of automated segmentation tools, especially for diverse medical conditions and imaging modalities.

To mitigate this challenge, various paradigms for learning with limited annotations have emerged, including few-shot learning (FSL) (Awudong et al. 2024; Alsaleh et al. 2024), one-shot learning (OSL) (Kato and Hotta 2023; Jiang et al. 2024), and semi-supervised learning (SSL) (Qiu et al. 2024; Li et al. 2024). FSL/OSL methods often focus on generalizing to novel classes using one or few examples, typically involving complex mechanisms like prototype matching or attention between support and query sets (Awudong et al. 2024; Tang et al. 2025; Lin et al. 2023; Huang et al. 2023b). SSL leverages large amounts of unlabeled data alongside a small labeled set, commonly employing pseudo-labeling (self-training) or consistency regularization to improve performance on the seen classes (Qiu et al. 2024; He et al. 2024; Das et al. 2024). While promising, FSL/OSL often requires specific support/query structures not always present, and SSL typically assumes access to more than a single labeled sample and necessitates careful handling of pseudo-label noise (Qiu et al. 2024). Recently, GenSeg (Zhang et al. 2025) proposed leveraging generative AI to synthesize training images in scenarios with limited labeled data.

The advent of large-scale foundation models (FMs) pre-trained on vast datasets, such as Vision Transformers (ViTs) (Dosovitskiy et al. 2020), DINOv2 (Jiang et al. 2024) offers another avenue. These models learn rich, transferable features (Zhou et al. 2023; Huang et al. 2023a) that hold significant potential in low-data regimes. However, effectively adapting them to the nuances of medical segmentation, particularly with minimal manual supervision, faces challenges due to domain shifts and the demand for high pixel-level precision (Zhou et al. 2023).

In this work, we push the boundaries of annotation efficiency by tackling the extreme challenge of training a segmentation model with just a single annotated image and a pool of unlabeled data. We introduce MedSMILE (**Med**ical Image **S**egmentation with **MI**nimal **L**abeling **E**ffort), a novel framework that creates a virtuous cycle of label generation and refinement. MedSMILE uniquely synergizes transductive inference and inductive learning to bootstrap performance from this minimal starting point.

The process, illustrated in Fig. 1, begins with a transductive step: leveraging powerful features from a pre-trained foundation model, the single annotation is propagated across the entire unlabeled dataset to generate initial pseudo-labels. This sparks an iterative self-training loop. In each round, a segmentation model is trained inductively on the current pseudo-labels and then generates potentially improved labels for the next round. Critically, an inter-round transductive refinement step uses the foundation model's features again to correct predictions for uncertain samples by propagating information from high-confidence ones. This iterative process allows MedSMILE to progressively distill high-quality supervision from the unlabeled data, starting from just one labeled example. Our main contributions are:

- We introduce MedSMILE, a novel framework that synergistically integrates transductive and inductive learning for the extreme one-label semi-supervised setting. Its core novelty lies in the iterative loop where a foundation model both bootstraps and refines pseudo-labels for an inductive segmentation model.
- Our approach significantly outperforms state-of-the-art semi-supervised methods while using orders of magnitude less labeled data, establishing a new benchmark for annotation efficiency in medical image segmentation.
- We demonstrate, for the first time to our knowledge, that competitive segmentation performance is achievable from a single annotated sample. Across diverse medical datasets, MedSMILE recovers 90%–95% of the fully supervised performance (in terms of Dice score).

Our work shows that the annotation bottleneck in medical imaging can be dramatically alleviated. By structuring the interplay between foundation models and self-training, MedSMILE charts a promising path toward developing powerful segmentation tools with minimal human effort.

Related Work

Our proposed method, MedSMILE, is a semi-supervised learning approach situated at the intersection of one-shot segmentation, semi-supervised learning, and foundation model application. We position our exploration within the context of these three paradigms.

Few-Shot and One-Shot Segmentation (FSS/OSS)

FSS/OSS aims to segment novel object classes given only one or a few annotated support examples (Awudong et al. 2024; Kato and Hotta 2023). A dominant paradigm involves prototypical networks, where class prototypes are derived from support features (e.g., via masked average pooling) and query pixels are classified based on feature similarity (often cosine) to these prototypes (Awudong et al. 2024; Huang et al. 2023b). Recognizing the limitations of single prototypes for complex structures, recent works explore multiple prototypes (Tang et al. 2025), adaptive local prototypes (Huang et al. 2023b), part-aware prototypes (Tang et al. 2025), or vector quantization (Huang et al. 2023b). Attention mechanisms and Transformers are also widely used to model interactions and align features between support and query images (Lin et al. 2023; Ding et al. 2023).

Specific one-shot medical segmentation methods often employ tailored strategies. ProtoSAM (Jiang et al. 2024) uses DINOv2 features within a prototype network to generate prompts for the Segment Anything Model (SAM) (Zhou et al. 2023), leveraging SAM's zero-shot capabilities without fine-tuning it. OneSeg (He et al. 2023) tackles 3D volumes by learning inter-slice correspondence via self-supervised reconstruction and propagating a single 2D slice annotation. Kato and Hotta (2023) use an attention mechanism guided by small visual prompt pairs derived from a single annotated image.

MedSMILE differs from typical FSS/OSS frameworks in its objective and structure. Firstly, while initialized by one example, its goal is to segment the entire dataset (assumed to contain instances of the same class(es) as the reference), rather than generalizing to novel classes based on support/query pairs. Secondly, it employs an iterative refinement process over multiple rounds, progressively improving labels for the whole dataset, which contrasts with the typical single-pass segmentation in FSS/OSS after learning the support-query relationship. Thirdly, our initial pseudo-label generation relies on direct similarity to reference features, while subsequent rounds use the trained segmentation model's predictions, refined using FM feature similarity between dataset samples guided by uncertainty, rather than explicitly constructed prototypes refined through meta-learning or complex interaction modules common in FSS/OSS.

Semi-Supervised Segmentation (SSS)

SSS leverages abundant unlabeled data alongside limited labeled data to improve segmentation performance, typically for classes present in the labeled set (Qiu et al. 2024). Pseudo-labeling (or self-training) is a cornerstone technique: a model trained on labeled data generates labels for unlabeled data, which are then used to retrain the model, often iteratively (Qiu et al. 2024; Zheng et al. 2020). Consistency regularization is another key approach, enforcing prediction consistency under input perturbations or model variations (Qiu et al. 2024; Das et al. 2024). Many SSS methods combine both strategies (He et al. 2024) and incorporate sophisticated techniques to handle pseudo-label noise, such as uncertainty estimation and weighting (Qiu et al. 2024), teacher-student frameworks (Li et al. 2024), adversarial learning (Awudong et al. 2024), contrastive learning (Wang et al. 2024), CRF-based refinement (Qiu et al. 2024), or dynamic sample selection (Shao et al. 2024). FS-RENet (Long et al. 2023) uses feature similarity constraints, while SGRS-Net (Wang et al. 2024) uses regional supervi-

sion based on pseudo-label synergy.

MedSMILE operates under the semi-supervised learning setup. It contrasts with traditional SSS methods primarily in its starting point (only one label) and its specific refinement mechanism. Unlike standard iterative self-training where the same model architecture is continually fine-tuned, MedSMILE re-initializes the Mask2Former model from pretrained weights each round. This strategy encourages the model to learn primarily from the current, potentially improved, set of pseudo-labels, reducing reliance on potentially biased knowledge accumulated from poorer labels in early rounds. While MedSMILE in its core form lacks explicit consistency regularization or complex co-training structures found in some advanced SSS methods, the iterative refinement loop itself, particularly with the uncertainty-guided KNN propagation step, serves as an implicit mechanism for improving pseudo-label quality and robustness over rounds. The refinement step shares conceptual similarities with uncertainty-aware SSS methods but employs a distinct mechanism based on FM feature propagation.

Foundation Models in Medical Imaging

Foundation models, pre-trained on large-scale datasets, are increasingly explored in medical imaging (Zhou et al. 2023; Huang et al. 2024, 2023a). Vision Transformers like DINOv2 (Jiang et al. 2024) provide powerful off-the-shelf feature extractors. CLIP (Guo et al. 2023), pre-trained on image-text pairs, enables zero-shot capabilities via text prompts, with medical adaptations emerging (Huang et al. 2023a; Hua et al. 2024). SAM (Kirillov et al. 2023) offers promptable segmentation but often requires adaptation for optimal performance on medical images (Zhou et al. 2023; Konwer et al. 2025). Mask2Former (Cheng et al. 2022) is a state-of-the-art Transformer-based segmentation architecture. Using FMs in low-data medical settings often involves strategies beyond simple fine-tuning, including parameter-efficient adaptation (Liu, Luo, and Zhu 2024), prompt engineering (Zheng et al. 2024), or integration into specialized frameworks (Jiang et al. 2024). A key challenge remains the domain gap between FM pre-training data and specific medical image characteristics (Zhou et al. 2023). The optional self-supervised fine-tuning of the DINOv2 encoder in MedSMILE is one strategy to address this domain gap.

Methodology

Our method, MedSMILE, is a multi-round pipeline for medical image segmentation model training. Given a single annotated reference image (I_{ref}, M_{ref})—where $M_{ref} \in \{0, \dots, C-1\}^{H_{ref} \times W_{ref}}$ is a C-class label map—and a dataset of N unlabeled images $\mathcal{D}_U = \{I_u\}_{u=1}^{N}$, the pipeline comprises an optional encoder fine-tuning step, an initialization phase (Round 0) utilizing DINOv2 features (LVD-142M pretrained), followed by R iterative refinement rounds ($r = 1, \dots, R$). Subsequent rounds employ segmentation model training, pseudo-label generation, and a DINO-based pseudo-label refinement. The final output of the pipeline is a trained segmentation model.

Self-Supervised Encoder Fine-tuning

To enhance the quality of features extracted by the DINOv2 model, especially if a significant domain shift exists between its pre-training data and the target unlabeled medical images $\mathcal{D}_U$, an optional self-supervised fine-tuning step can be performed on the DINOv2 encoder before Round 0. This step aims to adapt the encoder to the specific characteristics of $\mathcal{D}_U$. The fine-tuning employs a contrastive learning framework, specifically a Momentum Contrast (MoCo)-style approach (He et al. 2020; Chen et al. 2020b).

- **Architecture**: A student and a teacher network are utilized. Both networks are initialized with the same DINOv2 encoder (Φ_{DINO}), which comes pre-trained on the LVD-142M dataset. Following the encoder, each network has a projection head (e.g., a shallow MLP) that maps the output features to a lower-dimensional space where the contrastive loss is computed. The student network's parameters are updated via backpropagation, while the teacher network's parameters are an exponential moving average (EMA) of the student's and are not updated by gradients.
- **Data Augmentation**: A multi-crop augmentation strategy is applied to each image in $\mathcal{D}_U$. Typically, this generates two global views (larger crops) and several local views (smaller crops) of the same image.
- **Training Objective**: The student encoder processes one set of augmented views (e.g., one crop), while the teacher encoder processes another set of augmented views (e.g., the other crop) from the same original images. The InfoNCE loss (Oord, Li, and Vinyals 2018) is then used to train the student network. The loss encourages the projection of an augmented view from the student to be similar to the projection of a different augmented view of the same image (from the teacher, serving as a positive key) and dissimilar to projections from different images (negative keys) within the batch.
- **Output**: After fine-tuning for a set number of epochs, the updated student DINOv2 encoder, denoted as Φ'_{DINO}, is saved. If this step is performed, Φ'_{DINO} is used as the feature extractor in the subsequent pseudo-label generation process. If this step is skipped, the original pre-trained Φ_{DINO} is used.

For clarity, the DINOv2 encoder used in the subsequent steps will be referred to as Φ_{DINO}, implicitly meaning Φ'_{DINO} if the fine-tuning was performed.

Round 0: Initial Pseudo-Label Generation

Round 0 generates initial soft pseudo-labels $\mathcal{L}^{(0)} = \{PL_u^{(0)}\}_{u=1}^{N}$ for all $I_u \in \mathcal{D}_U$ using (I_{ref}, M_{ref}). This involves three steps:

1. **Feature Extraction:** The DINOv2 Vision Transformer, Φ_{DINO} (potentially fine-tuned as described in Section) (Jiang et al. 2024), extracts patch-level features. I_{ref} and each $I_u \in \mathcal{D}_U$ are resized to DINOv2's input dimensions (e.g., 518×518). L2-normalized feature maps, $F_{ref} = \Phi_{DINO}(I_{ref}) \in \mathbb{R}^{D \times H_{feat} \times W_{feat}}$ and $F_u =$

$\Phi_{DINO}(I_u) \in \mathbb{R}^{D \times H_{feat} \times W_{feat}}$, are extracted from a designated layer (e.g., `x_norm_patchtokens`). D is the feature dimension, and (H_{feat}, W_{feat}) are the feature map's spatial dimensions.

2. **Class Prototype Calculation:** For each class c in M_{ref}, a class prototype $\boldsymbol{p}_c \in \mathbb{R}^D$ is computed. M_{ref} is resized to (H_{feat}, W_{feat}), yielding M'_{ref}. Each $\boldsymbol{p}_c$ is the L2-normalized mean of feature vectors from F_{ref} at locations corresponding to class c in M'_{ref}:

$$\boldsymbol{p}'_c = \frac{\sum_{h=1}^{H_{feat}} \sum_{w=1}^{W_{feat}} \mathbb{I}(M'_{ref}[h,w] = c) \cdot F_{ref}[:,h,w]}{\sum_{h=1}^{H_{feat}} \sum_{w=1}^{W_{feat}} \mathbb{I}(M'_{ref}[h,w] = c) + \epsilon} \tag{1}$$

$$\boldsymbol{p}_c = \frac{\boldsymbol{p}'_c}{\|\boldsymbol{p}'_c\|_2 + \epsilon} \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function and ϵ ensures numerical stability. We assume each class c in M'_{ref} has at least one corresponding feature vector.

3. **Similarity Computation and Soft Pseudo-Label Generation:** For each I_u and class c, cosine similarity is computed between its feature vectors $F_u[:,h,w]$ and $\boldsymbol{p}_c$. This yields low-resolution class similarity maps $\{S^{low}_{u,c}\}_{c=0}^{C-1}$, where $S^{low}_{u,c} \in [-1,1]^{H_{feat} \times W_{feat}}$:

$$S^{low}_{u,c}[h,w] = F_u[:,h,w] \cdot \boldsymbol{p}_c \tag{3}$$

Each $S^{low}_{u,c}$ is upsampled (e.g., via bilinear interpolation) to the target resolution (H,W), yielding $S^{high}_{u,c}$. Initial soft pseudo-labels $PL^{(0)}_u \in [0,1]^{C \times H \times W}$ are formed by applying a pixel-wise softmax across classes:

$$PL^{(0)}_u[k,h,w] = \frac{\exp(S^{high}_{u,k}[h,w])}{\sum_{j=0}^{C-1} \exp(S^{high}_{u,j}[h,w])} \tag{4}$$

, for each class k.

Rounds $r = 1...R$: Iterative Model Training and Pseudo-Label Refinement

For rounds $r = 1, \ldots, R$, the pipeline iteratively refines pseudo-labels and trains the segmentation model. The input to round r is $\mathcal{L}^{(r-1)}$ from the preceding round.

1. **Dataset Preparation:** A training dataset $\mathcal{D}^{(r)}_{train}$ is assembled. For $r = 1$, soft pseudo-labels $PL^{(0)}_u \in \mathcal{L}^{(0)}$ are converted to hard labels $M^{(0)}_u \in \{0, \ldots, C-1\}^{H \times W}$ via argmax:

$$M^{(0)}_u[h,w] = \underset{c}{\text{argmax}}\ PL^{(0)}_u[c,h,w] \tag{5}$$

For $r > 1$, the input is $\mathcal{L}^{(r-1)} = \{M^{(r-1)}_u\}_{u=1}^N$ (potentially refined per step 4) from round $r-1$. Thus, $\mathcal{D}^{(r)}_{train} = \{(I_u, M^{(r-1)}_u)\}_{u=1}^N$.

2. **Model Initialization and Training:** A segmentation model $\mathcal{M}^{(r)}_{M2F}$ (Cheng et al. 2022) is initialized with LVD-142M pre-trained weights. Crucially, $\mathcal{M}^{(r)}_{M2F}$ is re-initialized at the start of each round $r \geq 1$. $\mathcal{M}^{(r)}_{M2F}$ is trained on $\mathcal{D}^{(r)}_{train}$ for E epochs, minimizing a multi-class segmentation loss $\mathcal{J}$ (e.g., combined cross-entropy and Dice loss). Training employs an optimizer $\mathcal{O}$ (e.g., AdamW (Loshchilov and Hutter 2017)) and learning rate scheduler $\mathcal{S}$ (e.g., Cosine Annealing (Loshchilov and Hutter 2016)).

3. **Pseudo-Label Generation and Uncertainty Estimation:** The trained $\mathcal{M}^{(r)}_{M2F}$ performs inference on all $I_u \in \mathcal{D}_U$. For each I_u, model outputs are raw logits $L^{(r)}_u \in \mathbb{R}^{C \times H \times W}$. These are converted to hard pseudo-labels $M^{(r)}_{u,\text{raw}} \in \{0, \ldots, C-1\}^{H \times W}$:

$$M^{(r)}_{u,\text{raw}}[h,w] = \underset{c}{\text{argmax}}\ L^{(r)}_u[c,h,w] \tag{6}$$

We use average pixel-wise entropy of the probability map $P^{(r)}_u = \text{softmax}(L^{(r)}_u)$ as the uncertainty metric:

$$\mathcal{U}^{(r)}_u = \frac{-1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=0}^{C-1} P^{(r)}_u[c,h,w] \log P^{(r)}_u[c,h,w] \tag{7}$$

The set of raw hard pseudo-labels is $\mathcal{M}^{(r)}_{\text{raw}} = \{M^{(r)}_{u,\text{raw}}\}_{u=1}^N$, and uncertainty scores are $\mathcal{U}^{(r)} = \{\mathcal{U}^{(r)}_u\}_{u=1}^N$.

4. **Pseudo-Label Refinement via KNN Propagation:** $M^{(r)}_{u,\text{raw}}$ are refined using $\mathcal{U}^{(r)}$ and FM features.

- *Identify Sets:* Using a pre-defined uncertainty quantile q_{unc} (e.g., 0.9), $\mathcal{D}_U$ is partitioned into an uncertain set $\mathcal{D}^{\text{unc}}_U = \{I_u \mid \mathcal{U}^{(r)}_u \geq \text{quantile}(\mathcal{U}^{(r)}, q_{unc})\}$ and a certain set $\mathcal{D}^{\text{cert}}_U = \mathcal{D}_U \setminus \mathcal{D}^{\text{unc}}_U$.
- *Extract Refinement Features:* DINOv2 features $F'_u = \Phi_{DINO}(I_u) \in \mathbb{R}^{D'}$ are extracted for all $I_u \in \mathcal{D}_U$, and L2-normalized to $\boldsymbol{f}_u = F'_u / \|F'_u\|_2$.
- *Build KNN Index:* A K-Nearest Neighbors index $\mathcal{K}$ is built on features $\{\boldsymbol{f}_u \mid I_u \in \mathcal{D}^{\text{cert}}_U\}$.
- *Propagate Labels:* For each $I_u \in \mathcal{D}^{\text{unc}}_U$:
  - Query $\mathcal{K}$ with $\boldsymbol{f}_u$ to find Q nearest neighbors $\{I_{n_1}, \ldots, I_{n_Q}\}$ from $\mathcal{D}^{\text{cert}}_U$ with features $\{\boldsymbol{f}_{n_1}, \ldots, \boldsymbol{f}_{n_Q}\}$.
  - Calculate similarity weights $w_j = \max(0, \boldsymbol{f}_u \cdot \boldsymbol{f}_{n_j})$ for $j = 1, \ldots, Q$.
  - Retrieve raw pseudo-labels $\{M^{(r)}_{n_j,\text{raw}}\}$ of neighbors, at resolution (H,W).
  - A refined soft label map $\bar{M}^{(r)}_u \in [0,1]^{C \times H \times W}$ is computed by weighted averaging of one-hot encoded (OHE) neighbor labels:

$$\bar{M}^{(r)}_u[k,h,w] = \frac{\sum_{j=1}^{Q} w_j \cdot \text{OHE}(M^{(r)}_{n_j,\text{raw}})[k,h,w]}{\sum_{j=1}^{Q} w_j + \epsilon} \tag{8}$$

  - The refined hard mask is generated via argmax:

$$M^{(r)}_{u,\text{refined}}[h,w] = \underset{k}{\text{argmax}}\ \bar{M}^{(r)}_u[k,h,w] \tag{9}$$

- For $I_u \in \mathcal{D}_U^{\text{cert}}$, $M_{u,\text{refined}}^{(r)} = M_{u,\text{raw}}^{(r)}$.
- The resulting set of refined pseudo-labels is $\mathcal{L}^{(r)} = \{M_{u,\text{refined}}^{(r)}\}_{u=1}^{N}$.

5. **Prepare for Next Round:** Pseudo-labels $\mathcal{L}^{(r)}$ are saved and serve as input $M_u^{(r)}$ for round $r+1$ (Step 1). Steps 1-5 are repeated for R rounds.

After R rounds, MedSMILE outputs the trained segmentation model $\mathcal{M}_{M2F}^{(R)}$, and optionally the pseudo-labels $\mathcal{L}^{(R)}$.

Experiments

We conduct experiments to address two key questions: (1) How does the proposed MedSMILE compare with state-of-the-art methods, mostly semi-supervised approaches, for medical image segmentation? and (2) To what extent can MedSMILE approach the performance of fully supervised learning? We benchmark our method against a range of classical and recently proposed semi-supervised algorithms, including MT (Tarvainen and Valpola 2017), DAN (Zhang et al. 2017), UA-MT (Yu et al. 2019), URPC (Luo et al. 2021), CLCC (Zhao et al. 2022), SLC-Net (Liu, Desrosiers, and Zhou 2022), MC-Net+ (Wu et al. 2022), CDMA (Zhong et al. 2023), SCP-Net (Zhang et al. 2023), MCP (Wang et al. 2023), KnowSAM (Huang et al. 2025) and DEC-Seg (Gu et al. 2025). We further include a few-shot learning method (Khadka et al. 2022) and a recent generative AI-based approach (Zhang et al. 2025) for comparison.

The proposed MedSMILE operates over a total of four rounds: an initial Round 0 to obtain the initial pseudo-label masks, followed by three iterative refinement rounds ($R = 3$). We use a learning rate of 5×10^{-5}, a batch size of 4, and apply gradient clipping with a maximum norm of 1.0. For pseudo-label refinement, we use an uncertainty quantile threshold of 0.9, meaning that the top 10% most uncertain pseudo-label masks are selectively refined. K is set to 5 for the KNN propagation in the refinement step.

Tasks and Datasets

We evaluate MedSMILE and related methods on seven public medical image segmentation datasets covering endoscopy, dermoscopy, MRI, and ultrasound. To assess performance under extreme annotation scarcity, MedSMILE is trained using only a single labeled image per dataset, with all remaining training images treated as unlabeled.

Polyp Segmentation. Experiments are conducted on CVC-ColonDB (Tajbakhsh, Gurudu, and Liang 2015), Kvasir (Jha et al. 2020), CVC-ClinicDB (Bernal et al. 2015), and ETIS (Silva et al. 2014). Following (Fan et al. 2020), 1,450 images from CVC-ClinicDB and Kvasir form the training set, with only one image manually labeled and the remaining 1,449 unlabeled. For comparison, the competing semi-supervised learning methods use the setting with 10% labeled data (145 images). The test set includes the rest of CVC-ClinicDB and Kvasir, plus all the samples from ETIS and CVC-ColonDB.

Skin Lesion Segmentation. We use the ISIC 2018 dataset (Codella et al. 2019), containing 2,594 training images with lesion annotations. Only one is labeled in our setting, with the rest treated as unlabeled. For comparison, the competing semi-supervised learning methods use the 10%-labeled setup. Evaluation is on the official test set.

Brain MRI Segmentation. The LGG dataset (Buda, Saha, and Mazurowski 2019) provides FLAIR MRI scans and expert-labeled masks for 110 patients. We split the data into 59 patients (2,017 slices) for training and 51 patients (1,912 slices) for testing. For our method, only one slice in the training set is labeled; the remaining 2,016 are used as unlabeled data.

BUSI Ultrasound Segmentation. The BUSI dataset (Al-Dhabyani et al. 2020) contains ultrasound images labeled as normal, benign, or malignant. We use the 570 normal and benign images for training (only one labeled), and the 210 malignant images as the test set. Only one image in the training set is labeled for our method.

Comparing to Semi-supervised Learning Methods

The results presented in Table 1 demonstrate the effectiveness of our proposed semi-supervised learning method for polyp segmentation tasks. Across all four datasets (CVC-ClinicDB, Kvasir, CVC-ColonDB, and ETIS), our approach achieves state-of-the-art performance, as evidenced by higher mDice and mIoU scores, while utilizing a remarkably small number of annotated samples. Specifically, our method requires only 1 labeled sample compared to 145 labeled samples used by all other contemporary methods. For instance, on the CVC-ClinicDB and Kvasir datasets, our model surpasses the next best performing method, DEC-Seg, achieving an mDice of 0.843 (vs. 0.836) and 0.868 (vs. 0.859) respectively. This trend of superior performance with significantly less supervision is even more pronounced on the more challenging CVC-ColonDB and ETIS datasets, where our model shows substantial gains, achieving an mDice of 0.826 (compared to DEC-Seg's 0.648) and 0.771 (compared to DEC-Seg's 0.592). MedSMILE's performance was evaluated over 5 runs with different random seeds, with results reported as the mean and standard deviation across these runs.

As shown in Table 2, our method achieves a mean Intersection-over-Union (mIoU) of 0.792 on the ISIC-2018 dataset using only a single labeled image, outperforming all existing methods trained with 259 labeled samples. Although DEC-Seg gives a marginally higher Dice score (0.871 vs. 0.866), our approach attains the highest mIoU while using 258 fewer labeled images, highlighting its superior labeling efficiency and practical utility. Notably, our approach significantly outperforms GenSeg, a recent method tailored for ultra-low-data scenarios, which uses 40 labeled samples (compared to our single labeled image) and does not leverage any unlabeled data. Additionally, the state-of-the-art few-shot learning method iMAML (Khadka et al. 2022), which leverages meta-learning and implicit gradients in a 5-shot setting (5 labeled images), yields a lower Dice score of 0.774. These results collectively demonstrate that

	CVC-ClinicDB and Kvasir						CVC-ColonDB and ETIS				
	Method	#Lab.	#Unlab.	mDice ↑	mIoU ↑		Method	#Lab.	#Unlab.	mDice ↑	mIoU ↑
CVC-ClinicDB (Seen)	MT	145	1305	0.747	0.662	CVC-ColonDB (Unseen)	MT	145	1305	0.589	0.497
	DAN	145	1305	0.755	0.674		DAN	145	1305	0.620	0.517
	UA-MT	145	1305	0.749	0.676		UA-MT	145	1305	0.553	0.471
	URPC	145	1305	0.769	0.696		URPC	145	1305	0.556	0.480
	CLCC	145	1305	0.794	0.720		CLCC	145	1305	0.538	0.473
	SLC-Net	145	1305	0.752	0.689		SLC-Net	145	1305	0.595	0.525
	MC-Net+	145	1305	0.767	0.702		MC-Net+	145	1305	0.562	0.486
	CDMA	145	1305	0.759	0.676		CDMA	145	1305	0.507	0.419
	SCP-Net	145	1305	0.776	0.703		SCP-Net	145	1305	0.577	0.495
	MCF	145	1305	0.779	0.718		MCF	145	1305	0.566	0.498
	KnowSAM	145	1305	0.834	0.774		KnowSAM	145	1305	0.698	0.621
	DEC-Seg	145	1305	0.836	0.774		DEC-Seg	145	1305	0.648	0.565
	MedSMILE	1	1449	**0.8426** ±0.0072	**0.7810** ±0.0084		**MedSMILE**	1	1449	**0.8255** ±0.0101	**0.7603** ±0.0072
	Polyp-PVT	1450	0	0.937	0.889		Polyp-PVT	1450	0	0.808	0.727
	PraNet-V2	1450	0	0.931	0.881		PraNet-V2	1450	0	–	–
Kvasir (Seen)	MT	145	1305	0.814	0.722	ETIS (Unseen)	MT	145	1305	0.356	0.288
	DAN	145	1305	0.808	0.723		DAN	145	1305	0.437	0.358
	UA-MT	145	1305	0.799	0.713		UA-MT	145	1305	0.459	0.393
	URPC	145	1305	0.811	0.728		URPC	145	1305	0.420	0.356
	CLCC	145	1305	0.806	0.724		CLCC	145	1305	0.409	0.346
	SLC-Net	145	1305	0.840	0.773		SLC-Net	145	1305	0.431	0.374
	MC-Net+	145	1305	0.817	0.735		MC-Net+	145	1305	0.439	0.369
	CDMA	145	1305	0.786	0.695		CDMA	145	1305	0.309	0.254
	SCP-Net	145	1305	0.810	0.723		SCP-Net	145	1305	0.394	0.323
	MCF	145	1305	0.822	0.751		MCF	145	1305	0.425	0.367
	KnowSAM	145	1305	0.859	0.793		KnowSAM	145	1305	0.526	0.459
	DEC-Seg	145	1305	0.859	0.787		DEC-Seg	145	1305	0.592	0.511
	MedSMILE	1	1449	**0.8680** ±0.0170	**0.8043** ±0.0212		**MedSMILE**	1	1449	**0.7705** ±0.0164	**0.7164** ±0.0152
	Polyp-PVT	1450	0	0.917	0.864		Polyp-PVT	1450	0	0.787	0.706
	PraNet-V2	1450	0	0.915	0.861		PraNet-V2	1450	0	0.764	0.687

Table 1: Segmentation performance on four polyp datasets.

our method effectively harnesses unlabeled data, substantially alleviating the annotation burden typically required for medical image segmentation.

Comparing to Fully-supervised Models

In Table 1, we further compare MedSMILE against fully supervised, state-of-the-art models, Polyp-PVT (Dong et al. 2023) and PraNet-V2 (Hu et al. 2025), for the polyp segmentation. On the seen domains of CVC-ClinicDB and Kvasir, while the fully supervised models expectedly set the performance ceiling, MedSMILE proves remarkably effective. Specifically, it recovers 90% of the state-of-the-art Dice score on CVC-ClinicDB (0.843 vs. 0.937) and 95% on Kvasir (0.868 vs. 0.917). **The true strength of our method is revealed when evaluating generalization on the unseen domains of CVC-ColonDB and ETIS.** Samples from these two datasets were not used during the training of any model, making them a robust test of generalization. Strikingly, on CVC-ColonDB, MedSMILE not only closes the performance gap but outperforms the fully supervised Polyp-PVT, achieving a Dice score of 0.8255 versus 0.808. On the ETIS dataset, our method's performance (0.7705) is highly competitive with Polyp-PVT (0.787) and surpasses PraNet-V2 (0.764). These results strongly suggest that MedSMILE's semi-supervised approach, which leverages a large pool of unlabeled data, fosters the learning of more robust and generalizable representations that are less prone to overfitting on the source training domains.

When compared to state-of-the-art fully supervised models on the ISIC-2018 dataset, such as SkinFormer (Xu et al. 2024) and Ms RED (Dai et al. 2022), the efficiency of MedSMILE becomes evident (see Table 2). With only one labeled image, our method reaches 93% of the Dice score achieved by the fully supervised SkinFormer, demonstrating exceptional annotation efficiency.

On the LGG brain MRI dataset (see Table 3), our semi-supervised approach achieved an mDice score of 0.882 and an mIoU of 0.862. Unsurprisingly, a fully-supervised model trained with all 2017 samples labeled performed better, yielding an mDice of 0.955 and an mIoU of 0.941. It is noteworthy that our method achieved approximately 92.3% of the fully-supervised mDice (0.882 vs. 0.955) and 91.6% of the mIoU (0.862 vs. 0.941). A similar trend is observed on the BUSI ultrasound dataset (see Table 4). Our method achieves an mDice of 0.765 and an mIoU of 0.672, recovers approximately 95.1% of the fully supervised mDice and 91.5% of the mIoU.

Methods	#Lab.	#Unlab.	mDice ↑	mIoU ↑
MT	259	2335	0.822	0.738
DAN	259	2335	0.831	0.744
UA-MT	259	2335	0.839	0.752
URPC	259	2335	0.847	0.763
CLCC	259	2335	0.842	0.756
SLC-Net	259	2335	0.843	0.754
MC-Net+	259	2335	0.848	0.767
CDMA	259	2335	0.858	0.777
SCP-Net	259	2335	0.852	0.770
MCF	259	2335	0.860	0.780
KnowSAM	259	2335	0.865	0.785
DEC-Seg	259	2335	**0.871**	0.789
GenSeg	40	0	0.691	–
MedSMILE	1	2593	0.866	**0.792**
Ms RED	2594	0	0.914	0.845
SkinFormer	2594	0	0.932	0.876

Table 2: Segmentation Performance on ISIC-2018 dataset.

Methods	#Labeled	#Unlabeled	mDice ↑	mIoU ↑
MedSMILE	1	2016	0.882	0.862
Full (M2F)	2017	0	0.955	0.941

Table 3: Performance on the LGG brain MRI dataset: MedSMILE vs. model trained using fully labeled data set.

Methods	#Labeled	#Unlabeled	mDice ↑	mIoU ↑
MedSMILE	1	569	0.765	0.672
Full (M2F)	570	0	0.804	0.734

Table 4: Performance on BUSI ultrasound dataset: MedSMILE vs. model trained using fully labeled data set.

Dataset	**w/o Refinement**		**w Refinement**	
	mDice	mIoU	mDice	mIoU
CVC-ClinicDB	0.825	0.759	0.848	0.788
Kvasir	0.873	0.804	0.887	0.823
CVC-ColonDB	0.806	0.739	0.823	0.758
ETIS	0.748	0.698	0.773	0.719
ISIC	0.838	0.754	0.866	0.792
LGG	0.840	0.830	0.882	0.862
BUSI	0.691	0.603	0.765	0.672

Table 5: Ablation: Comparison of segmentation performance (mDice and mIoU) in MedSMILE with and without pseudo-label refinement.

Data	**Round 1**		**Round 2**		**Round 3**	
	mDice	mIoU	mDice	mIoU	mDice	mIoU
ISIC	0.855	0.780	0.868	0.794	0.866	0.792
LGG	0.856	0.841	0.868	0.849	0.882	0.862
BUSI	0.739	0.642	0.746	0.648	0.765	0.672

Table 6: Ablation: Segmentation performance (mDice and mIoU) improves as more rounds are applied in MedSMILE.

Ablation Study

Table 5 highlights the important role of the pseudo-label refinement step in enhancing overall segmentation performance. Moreover, for imaging modalities like ultrasound, which were not part of the original DINO-v2 training, self-supervised encoder fine-tuning on DINO-v2 using the ultrasound images (unlabeled training images) is particularly critical. Without this pre-training, performance drops significantly to a 0.259 Dice score and a 0.156 mIoU, because the initial encoder pre-trained on LVD-142M is not familiar with ultrasound images. Table 6 further shows that segmentation performance in MedSMILE generally improves with more training rounds, though minor fluctuations may occur.

Time Cost Analysis

The main computational cost is from Mask2Former training, which scales linearly with the number of rounds, epochs, and unlabeled images (N). Each refinement round adds a constant overhead due to three full $O(N)$ passes: (1) Mask2Former inference for label and uncertainty generation, (2) DINOv2 feature extraction, and (3) KNN-based label propagation. The initial DINOv2 inference in Round 0 incurs a one-time $O(N)$ cost. In practice, the framework remains efficient. For example, the full MedSMILE pipeline on the ISIC dataset completes in under an hour on a single consumer-grade GPU (RTX 4070 Ti SUPER).

Limitations

When the foundation model (FM) feature encoder (e.g., DINOv2) fails to extract robust features, self-supervised fine-tuning on the unlabeled dataset becomes necessary. While such fine-tuning can improve domain adaptation and performance, it can be time-consuming. Furthermore, in fine-grained segmentation tasks that require distinguishing between visually similar structures, the encoder may still struggle to capture subtle differences even after fine-tuning. This can result in unreliable initial pseudo-labels, potentially limiting the effectiveness of MedSMILE.

Conclusion

We have introduced MedSMILE, a framework that successfully pushes the limits of medical image segmentation with minimal labeling effort. The primary novelty of our work lies not in the individual components (foundation models and self-training), but in their principled integration into a virtuous cycle of label generation and refinement, tailored for the extreme one-label setting. We demonstrated how a foundation model can serve a dual transductive role: first for bootstrapping initial labels and later for refining uncertain predictions by leveraging feature similarity across the dataset. This process provides robust supervision for an inductive segmentation model which, by being re-initialized each round, is forced to adapt to the progressively improving pseudo-labels without inheriting bias from its own past errors. This principled approach establishes a new, effective paradigm for annotation-efficient learning, showing that competitive performance is attainable even when supervision is reduced to a theoretical minimum. Future work will apply MedSMILE to more datasets and imaging modalities.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant 62201263) and Natural Science Foundation of Jiangsu Province (Grant BK20220949).

References

Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; and Fahmy, A. 2020. Dataset of Breast Ultrasound Images. *Data in Brief*, 28: 104863.

Alsaleh, A. M.; Albalawi, E.; Algosaibi, A.; Albakheet, S. S.; and Khan, S. B. 2024. Few-shot learning for medical image segmentation using 3D U-Net and model-agnostic meta-learning (MAML). *Diagnostics*, 14(12): 1213.

Awudong, B.; Li, Q.; Liang, Z.; Tian, L.; and Yan, J. 2024. Attentional adversarial training for few-shot medical image segmentation without annotations. *Plos one*, 19(5): e0298227.

Bernal, J.; Sanchez, F. J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; and Vilarino, F. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43: 99–111.

Buda, M.; Saha, A.; and Mazurowski, M. A. 2019. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in Biology and Medicine*, 109: 218–225.

Chen, C.; Qin, C.; Qiu, H.; Tarroni, G.; Duan, J.; Bai, W.; and Rueckert, D. 2020a. Deep Learning for Cardiac Image Segmentation: A Review. *Frontiers in Cardiovascular Medicine*, 7: 25.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention Mask Transformer for Universal Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1290–1299.

Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M. E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*.

Dai, D.; Dong, C.; Xu, S.; Yan, Q.; Li, Z.; Zhang, C.; and Luo, N. 2022. Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation. *Medical image analysis*, 75: 102293.

Das, A.; Gautam, C.; Cholakkal, H.; Agrawal, P.; Yang, F.; Savitha, R.; and Liu, Y. 2024. Decoupled Training for Semi-supervised Medical Image Segmentation with Worst-Case-Aware Learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 45–55. Springer.

Ding, H.; Sun, C.; Tang, H.; Cai, D.; and Yan, Y. 2023. Few-shot medical image segmentation with cycle-resemblance attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2488–2497.

Dong, B.; Wang, W.; Fan, D.-P.; Li, J.; Fu, H.; and Shao, L. 2023. Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers. *CAAI Artificial Intelligence Research*, 2.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020. PraNet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 263–273. Springer.

Gu, Y.; Zhou, T.; Zhang, Y.; Zhou, Y.; He, K.; Gong, C.; and Fu, H. 2025. Dual-scale enhanced and cross-generative consistency learning for semi-supervised medical image segmentation. *Pattern Recognition*, 158: 110962.

Guo, S.-C.; Liu, S.-K.; Wang, J.-Y.; Zheng, W.-M.; and Jiang, C.-Y. 2023. CLIP-driven prototype network for few-shot semantic segmentation. *Entropy*, 25(9): 1353.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, Q.; Yan, K.; Luo, Q.; Yi, D.; Wang, P.; Han, H.; and Liu, D. 2024. Exploring unlabeled data in multiple aspects for semi-supervised MRI segmentation. *Health Data Science*, 4: 0166.

He, S.; Zhang, R.; Liu, D.; Yan, P.; Li, J.; and Wang, G. 2023. OneSeg: A Simple Framework for Interactive One-Shot Segmentation of 3D Medical Images. arXiv:2309.13671.

Hu, B.-C.; Ji, G.-P.; Shao, D.; and Fan, D.-P. 2025. PraNet-V2: Dual-Supervised Reverse Attention for Medical Image Segmentation. *arXiv preprint arXiv:2504.10986*.

Hua, L.; Luo, Y.; Qi, Q.; and Long, J. 2024. Medicalclip: Anomaly-detection domain generalization with asymmetric constraints. *Biomolecules*, 14(5): 590.

Huang, K.; Zhou, T.; Fu, H.; Zhang, Y.; Zhou, Y.; Gong, C.; and Liang, D. 2025. Learnable prompting sam-induced knowledge distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*.

Huang, S.; Chen, Z.; Wang, L.; Zhou, F.; and Zhang, Y. 2023a. CLIP in medical imaging: A comprehensive survey. arXiv:2312.07353.

Huang, S.; Xu, T.; Shen, N.; Mu, F.; and Li, J. 2023b. Rethinking few-shot medical segmentation: a vector quantization view. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3072–3081.

Huang, S.-C.; Jensen, M.; Yeung-Levy, S.; Lungren, M. P.; Poon, H.; and Chaudhari, A. S. 2024. Multimodal Foundation Models for Medical Imaging-A Systematic Review and Implementation Guidelines. *medRxiv*, 2024–10.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; de Lange, T.; Johansen, D.; and Johansen, H. D. 2020. Kvasir-SEG: A segmented polyp dataset. In *International Conference on MultiMedia Modeling*, 451–462. Springer.

Jiang, Y.; Liu, J.; Chen, P. H. C.; Yan, C.; and Liu, H. 2024. ProtoSAM: One-Shot Medical Image Segmentation With Foundational Models. arXiv:2407.07042.

Kato, S.; and Hotta, K. 2023. One-shot and partially-supervised cell image segmentation using small visual prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4295–4304.

Khadka, R.; Jha, D.; Hicks, S.; Thambawita, V.; Riegler, M. A.; Ali, S.; and Halvorsen, P. 2022. Meta-learning with implicit gradients in a few-shot setting for medical image segmentation. *Computers in Biology and Medicine*, 143: 105227.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Konwer, A.; Yang, Z.; Bas, E.; Xiao, C.; Prasanna, P.; Bhatia, P.; and Kass-Hout, T. 2025. Enhancing SAM with Efficient Prompting and Preference Optimization for Semi-supervised Medical Image Segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 20990–21000.

Li, H.; Xu, X.; Liu, Z.; Xia, Q.; and Xia, M. 2024. Low-Quality Sensor Data-Based Semi-Supervised Learning for Medical Image Segmentation. *Sensors (Basel, Switzerland)*, 24(23): 7799.

Lin, Y.; Chen, Y.; Cheng, K.-T.; and Chen, H. 2023. Few shot medical image segmentation with cross attention transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 233–243. Springer.

Liu, J.; Desrosiers, C.; and Zhou, Y. 2022. Semi-supervised medical image segmentation using cross-model pseudo-supervision with shape awareness and local context constraints. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 140–150. Springer.

Liu, Y.; Luo, G.; and Zhu, Y. 2024. Fedfms: Exploring federated foundation models for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 283–293. Springer.

Long, J.; Yang, C.; Ren, Y.; and Zeng, Z. 2023. Semi-supervised medical image segmentation via feature similarity and reliable-region enhancement. *Computers in Biology and Medicine*, 167: 107668.

Loshchilov, I.; and Hutter, F. 2016. SGDR: Stochastic Gradient Descent with Warm Restarts. arXiv:1608.03983.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Luo, X.; Liao, W.; Chen, J.; Song, T.; Chen, Y.; Zhang, S.; Chen, N.; Wang, G.; and Zhang, S. 2021. Efficient Semi-Supervised Gross Target Volume of Nasopharyngeal Carcinoma Segmentation via Uncertainty Rectified Pyramid Consistency. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*, volume 12903 of *Lecture Notes in Computer Science*, 318–329. Springer.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Qiu, Z.; Gan, W.; Yang, Z.; Zhou, R.; and Gan, H. 2024. Dual uncertainty-guided multi-model pseudo-label learning for semi-supervised medical image segmentation. *Math. Biosci. Eng.*, 21: 2212–32.

Shao, Q.; Kang, J.; Chen, Q.; Li, Z.; Xu, H.; Cao, Y.; Liang, J.; and Wu, J. 2024. Enhancing semi-supervised learning via representative and diverse sample selection. *Advances in Neural Information Processing Systems*, 37: 111199–111226.

Silva, J.; Histace, A.; Romain, O.; Dray, X.; and Granado, B. 2014. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, 9(2): 283–293.

Tajbakhsh, N.; Gurudu, S. R.; and Liang, J. 2015. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2): 630–644.

Tang, S.; Yan, S.; Qi, X.; Gao, J.; Ye, M.; Zhang, J.; and Zhu, X. 2025. Few-shot medical image segmentation with high-fidelity prototypes. *Medical Image Analysis*, 100: 103412.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Wang, T.; Zhang, X.; Chen, Y.; Zhou, Y.; Zhao, L.; Tan, T.; and Tong, T. 2024. Synergy-Guided Regional Supervision of Pseudo Labels for Semi-Supervised Medical Image Segmentation. *arXiv preprint arXiv:2411.04493*.

Wang, Y.; Xiao, B.; Bi, X.; Li, W.; and Gao, X. 2023. MCF: Mutual Correction Framework for Semi-Supervised Medical Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15651–15660. IEEE.

Wu, Y.; Ge, Z.; Zhang, D.; Xu, M.; Zhang, L.; Xia, Y.; and Cai, J. 2022. Mutual Consistency Learning for Semi-Supervised Medical Image Segmentation. *Medical Image Analysis*, 81: 102530.

Xu, R.; Wang, C.; Zhang, J.; Xu, S.; Meng, W.; and Zhang, X. 2024. Skinformer: Learning statistical texture representation with transformer for skin lesion segmentation. *IEEE Journal of Biomedical and Health Informatics*, 28(10): 6008–6018.

Yao, W.; Bai, J.; Liao, W.; Chen, Y.; Liu, M.; and Xie, Y. 2024. From cnn to transformer: A review of medical image segmentation models. *Journal of Imaging Informatics in Medicine*, 37(4): 1529–1547.

Yu, L.; Wang, S.; Li, X.; Fu, C.-W.; and Heng, P.-A. 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *International conference on medical image computing and computer-assisted intervention*, 605–613. Springer.

Zhang, L.; Jindal, B.; Alaa, A.; Weinreb, R.; Wilson, D.; Segal, E.; Zou, J.; and Xie, P. 2025. Generative AI enables medical image segmentation in ultra low-data regimes. *Nature Communications*, 16(1): 6486.

Zhang, Y.; Yang, L.; Chen, J.; Fredericksen, M.; Hughes, D. P.; and Chen, D. Z. 2017. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International conference on medical image computing and computer-assisted intervention*, 408–416. Springer.

Zhang, Z.; Ran, R.; Tian, C.; Zhou, H.; Li, X.; Yang, F.; and Jiao, Z. 2023. Self-aware and cross-sample prototypical learning for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 192–201. Springer.

Zhao, X.; Fang, C.; Fan, D.-J.; Lin, X.; Gao, F.; and Li, G. 2022. Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation. In *2022 IEEE 19th international symposium on biomedical imaging (ISBI)*, 1–5. IEEE.

Zheng, H.; Zhang, Y.; Yang, L.; Wang, C.; and Chen, D. Z. 2020. An annotation sparsification strategy for 3D medical image segmentation via representative selection and self-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 6925–6932.

Zheng, X.; Zhang, Y.; Zhang, H.; Liang, H.; Bao, X.; Jiang, Z.; and Lao, Q. 2024. Curriculum prompting foundation models for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, 487–497. Springer.

Zhong, L.; Liao, X.; Zhang, S.; and Wang, G. 2023. Semi-supervised pathological image segmentation via cross distillation of multiple attentions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 570–579. Springer.

Zhou, H. Y.; Xu, H.; Xu, J.; Wu, S.; Wang, J.; and Zhou, S. K. 2023. Segment Anything Model for medical image analysis: an experimental study. *Medical Image Analysis*, 89: 102918.