SELF-IMPROVING SEMI-SUPERVISED MEDICAL IMAGE SEGMENTATION VIA CONSENSUS-ENHANCED PROTOTYPE-PROMPTED SEGMENT ANYTHING MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

The scarcity of expert-annotated data has driven the emergence of Semi-Supervised Medical Image Segmentation (SSMIS) as a promising approach. However, due to the absence of ground-truth supervision, current SSMIS methods cannot verify the correctness of self-consistency signals on unlabeled data, resulting in Blind Update. To address this, we propose Self-Improving Consensus-enhanced Prototype-prompted Segment Anything Model (SI-CP2SAM), the first self-improving SSMIS framework, which achieves reliable and accurate segmentation under limited annotation through self-improving mechanism. Specifically, it contains three key novel designs: Attention-guided Kolmogorov-Arnold Prototype Learning (AKaPL), which promotes reliable and debiased representations via Attention-guided Prototype Retrieval (APR) and Kolmogorov-Arnold enhanced Consensus Aggregation (KaCA); Consensus-guided Prototype Decoding (CPD), which yields complementary verifiable signals via prototype-conditioned prompting of SAM; and Anchor-Verified Gated Refinement (AVGR), which maintains a global unlabeled prototype memory and certifies its verifiable evolution via an accept-reject gating mechanism against trustworthy anchors. Extensive experiments on datasets demonstrate the effectiveness of SI-CP2SAM. Code is available at https://github.com/***.

1 INTRODUCTION

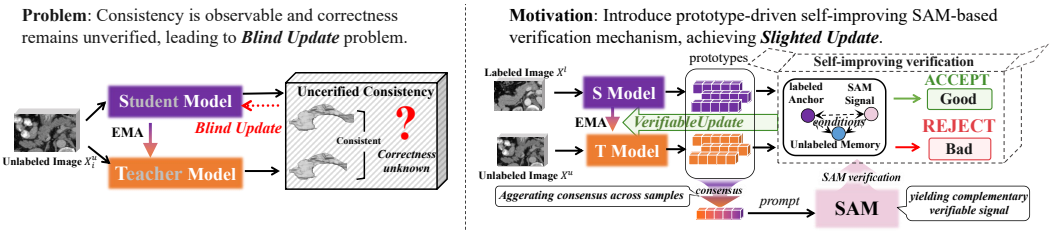
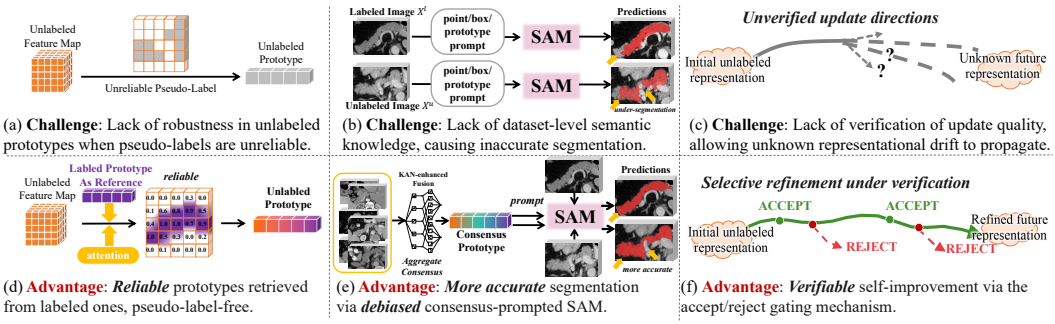


Figure 1: (a) Without reliable reference, the teacher-student consistency signal fails to certify prediction correctness on unlabeled data, causing Blind Update. (b) Our framework prompts SAM with consensus-enhanced prototypes as an complementary verification signal, gating model updates through an accept/reject condition to certify correctness and enable a Verifiable Update.

Dense pixel-level annotation is costly to acquire in medical imaging, making Semi-Supervised Medical Image Segmentation (SSMIS) a necessary solution (Moglia et al., 2025; Zhang et al., 2023). Among existing SSMIS approaches, Consistency Regularization (CR)-based methods (Yu et al., 2019; Luo et al., 2021; 2022) are the most widely adopted, enforcing consistent predictions on unlabeled data under different perturbations or views. Predictions on labeled data are well verified by ground truth; however, unlabeled data lacks a reliable reference, leaving the model unable to verify whether a given consistency signal is correct or not. Without this verification, an unlabeled consistency signal is propagated regardless of its correctness, rendering any embedded error irrecoverable once incorporated into the model. This unverifiable propagation of self-consistency leads to the **Blind Update** problem (Fig. 1(a)).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) **Challenge**: Lack of robustness in unlabeled prototypes when pseudo-labels are unreliable.
(b) **Challenge**: Lack of dataset-level semantic knowledge, causing inaccurate segmentation.
(c) **Challenge**: Lack of verification of update quality, allowing unknown representational drift to propagate.
(d) **Advantage**: ***Reliable*** prototypes retrieved from labeled ones, pseudo-label-free.
(e) **Advantage**: ***More accurate*** segmentation via ***debiased*** consensus-prompted SAM.
(f) **Advantage**: ***Verifiable*** self-improvement via the accept/reject gating mechanism.

Figure 2: Challenges of realizing prototype-driven self-improving verification (a)-(c) and the advantages of our SI-CP2SAM framework (d)-(f).

Prototype learning offers a natural interface for this verification (Liu et al., 2026a;b; Xu et al., 2022; Lu et al., 2023). Class-wise compact representations extracted from both labeled and unlabeled data can be directly compared. A segmentation foundation model such as SAM (Ma et al., 2024; Wang et al., 2025), endowed with strong zero-shot generalization from large-scale pretraining, provides a verification signal grounded in a separately pretrained representation, complementary to the SSMIS model's own feature space. Grounding this external signal alongside labeled prototypes as anchor into an explicit gating criterion transforms the correctness of a candidate update from an unverifiable presumption into a decidable condition, admitting only updates that satisfy it, constituting genuine self-improvement rather than the unconditional propagation underlying Blind Update(Fig. 1(b)). Realizing this mechanism, however, exposes three coupled limitations (Fig. 2(a)-(c)). (1) Lack of robustness in unlabeled prototype generation leaves it derived purely from the model's own prediction, as unreliable as the pseudo-label it originates from once decoupled from any labeled reference. (2) Lack of dataset-level semantic knowledge confines a single point, box, or prototype prompt to one sample's local evidence, leaving SAM's segmentation prone to under-segmentation. (3) Lack of verification of update quality admits a candidate update without an explicit decision mechanism, allowing representational drift to accumulate unchecked across iterations.

Building upon this, we propose Self-Improving Consensus-enhanced Prototype-prompted SAM (SI-CP2SAM), the first self-improving SSMIS framework, which certifies every consistency-driven update against trustworthy anchors before propagation, achieving **Verifiable Update**. Specifically, it contains three key designs. (1) Attention-guided Kolmogorov-Arnold Prototype Learning (AKaPL) retrieves unlabeled prototypes through attention-guided cross-sample correspondence grounded in trustworthy labeled anchors, yielding reliable unlablled prototypes free of pseudo-label dependency, and aggregates consensus via a Kolmogorov-Arnold Network (KAN)-enhanced fusion layer to suppress source-specific bias. (2) Consensus-guided Prototype Decoding (CPD) converts the consensus prototype into a class-specifying semantic prompt that conditions SAM's mask decoder, yielding more accurate segmentation grounded in SAM's segmentation capability and a verification signal complementary to the prototype branch. (3) Anchor-Verified Gated Refinement (AVGR) gates each candidate prototype update against the trustworthy labeled anchor through an accept-reject gating mechanism, admitting only verified updates into a long-term memory prototype and yielding Verifiable self-improvement (Fig. 2(d)-(f)).

Our contributions are:

- We propose the first self-improving SSMIS framework, resolving the Blind Update problem by Consensus-enhanced Prototype-prompted SAM to achieve more reliable and accurate segmentation under limited annotation.
- Our novel Attention-guided Kolmogorov-Arnold Prototype Learning (AKaPL) achieves pseudo-label-free unlabeled prototype generation by retrieving directly from labeled anchors, generating more reliable representations, with KAN-enhanced fusion aggregating debiased consensus across sources.
- Our novel Consensus-guided Prototype Decoding (CPD) prompts SAM with a cross-sample consensus prototype, yielding complementary verifiable signals decorrelated from the prototype branch.

- Our novel Anchor-Verified Gated Refinement (AVGR) certifies verifiable improvement via an accept-reject gating mechanism against trustworthy anchors, generating a more verified representation free of unchecked drift.

2 Related Works

2.1 Foundation Models in Medical Image Segmentation

Vision foundation models such as SAM (Kirillov et al., 2023) achieve strong zero-shot generalization across imaging modalities and anatomical structures through large-scale pretraining, enabling segmentation via sparse prompts such as points or boxes. To improve adaptability to downstream tasks, a variety of adaptation strategies have been explored: MedSAM (Ma et al., 2024) performs full fine-tuning of SAM's image encoder and mask decoder on large-scale medical image-mask pairs; SAM-Med3D (Wang et al., 2025) restructures SAM's 2D components into a native 3D architecture. PGP-SAM (Yan et al., 2025) and DAPSAM (Wei et al., 2024) instead adapt SAM's prompting interface itself, learning prototypes through gradient-based optimization to generate prompts in place of manual points or boxes. However, these prototype prompts remain freely updated parameters within a single data source, lacking any mechanism to correct sample-specific bias. Our method instead retrieves a reliable, consensus-carrying prototype from cross-sample correspondence, yielding a more robust SAM prompt.

2.2 Semi-supervised Medical Image Segmentation

Consistency Regularization (CR) gains its popularity in SSMIS paradigm, built on the Mean-Teacher framework (Tarvainen & Valpola, 2017), enforcing consistent predictions on perturbed unlabeled data. UA-MT (Yu et al., 2019) weights consistency by Monte Carlo dropout uncertainty; DTC (Luo et al., 2021) enforces consistency between segmentation and signed distance field predictions. Prototype Learning-based methods(Xu et al., 2022; Lu et al., 2023; Liu et al., 2026b;a) bridge labeled and unlabeled sources through class-wise prototypes: CPCL (Xu et al., 2022) enforces cyclic consistency between them; UPCoL (Lu et al., 2023) integrates both sources into a single prototype; Multiple prototype-based methods (Wang et al., 2024; Bi et al., 2025; Liu et al., 2026b) pursue richer intra-class characteristics: BaPC (Wang et al., 2024) separates boundary and inner prototypes; MPCL (Liu et al., 2026b) aligns multiple prototypes with intensity-based subregions identified through GMM decomposition. More recently, foundation models have been incorporated into SSMIS pipelines: CPC-SAM (Miao et al., 2024) builds a dual-branch cross-prompting scheme in which each branch's prediction generates the prompt for SAM 2D segmentation on the other; SemiSAM (Zhang et al., 2024) injects SAM's output into consistency regularization as an additional supervisory signal; SemiSAM+ (Zhang et al., 2025) fixes SAM as a frozen generalist exchanging positional prompts with a trainable specialist for pseudo-label supervision. Despite these advances, the consistency or pseudo-label signal remains unverified against any external reference, allowing incorrectness to propagate unchecked. Our method instead verifies each candidate update against trustworthy labeled anchors before, achieving slighted update.

3 Method

SI-CP2SAM (Fig. 3) constructs a verification-gated self-improving framework via three key designs: AKaPL (Sec. 3.1) promotes more reliable and debiased representations via Attention-guided Prototype Retrieval (APR) and Kolmogorov-Arnold-enhanced Consensus Aggregation (KaCA), P2S (Sec. 3.2) yields more accurate grounded pseudo-labels via consensus-based prototype-prompted conditioning of SAM, and AVGR (Sec. 3.3) certifies verifiable self-improvement via anchor-gated accept-reject refinement of long-term prototype memory.

Preliminaries. SSMIS trains on labeled dataset $D_l = \{(X_i^l, Y_i^l)\}_{i=1}^{N_l}$ and unlabeled dataset $D_u = \{X_j^u\}_{j=1}^{N_u}$, where $N_u \gg N_l$. We adopt the Mean-Teacher framework (Fig. 3(a)), which consists of a student model trained via backpropagation and a teacher model updated via Exponential Moving Average (EMA) of the student's weights for stable predictions. Given a labeled image X^l and an unlabeled image X^u, the student and teacher encoders extract feature maps $\mathbf{F}^l$ and $\mathbf{F}^u$, respectively. Class-wise prototypes for labeled data are obtained via Masked Average Pooling (MAP) guided by

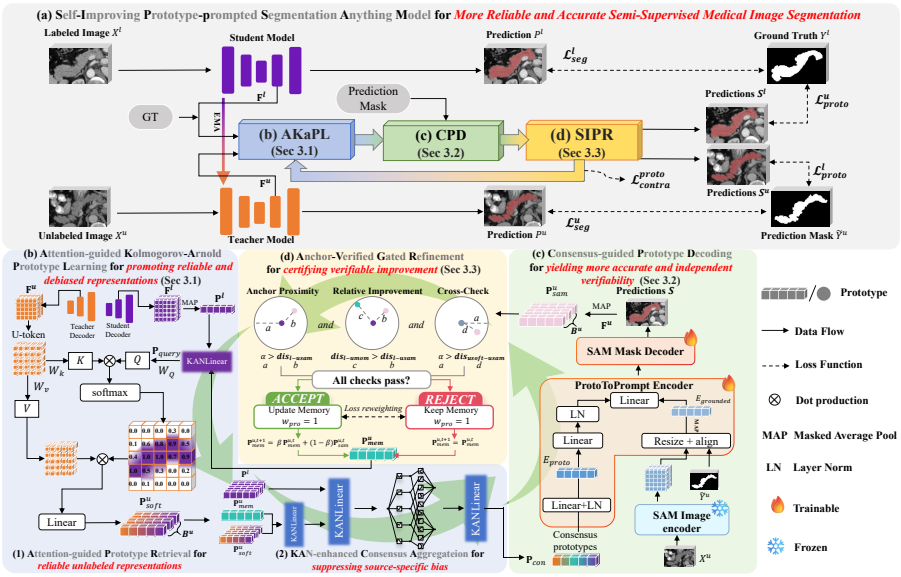


Figure 3: Overview of SI-CP2SAM, which achieves reliable and verifiable prototype self-improvement for SSMIS, consisting of (b) AKaPL promoting debiased representations, (c) CPD grounding complementary verifiability, and (d) AVGR certifying verifiable improvement via accept/reject gating against anchors.

the ground truth Y^l through $\mathbf{P}^l = \sum \mathbf{F}^l \cdot \mathbb{1}[Y^l = c] \,/\, \sum \mathbb{1}[Y^l = c]$, which are also referred to as the *labeled anchors*. For unlabeled data, we maintain a long-term memory prototype $\mathbf{P}^u_{mem}$, initialized from the labeled anchors as $\mathbf{P}^{u,0}_{mem} = \mathbf{P}^l$.

3.1 AKaPL for promoting reliable and debiased representations

AKaPL (Fig. 2(b)) employs ACR to generate reliable unlabeled prototype representations grounded in trustworthy labeled anchors, and further leverages KaCA to aggregate consensus across both labeled and unlabeled samples to get dataset-level semantic knowledge, jointly deriving more reliable and debiased prototype representations.

Attention-guided Prototype Retrieval (APR) novelly leverages attention-guided cross-sample prototype generation for unlabeled data based on trustworthy labeled anchors, eliminating the reliance on unreliable pseudo-labels required by conventional Masked Average Pooling (MAP), yielding more reliable unlabeled representations. Specifically, it flattens the unlabeled feature map $F^u \in \mathbb{R}^{D\times H\times W\times C}$ into a sequence of $N = D\times H\times W$ voxel tokens $F^u_{token} \in \mathbb{R}^{N\times C}$. The query anchor $\mathbf{P}_{query}$ is fused from the labeled anchor $\mathbf{P}^l$ and the long-term memory prototype $\mathbf{P}^u_{mem}$. It retrieves a soft-matched unlabeled prototype through cross-attention:

$$Q = \mathbf{P}_{query} W_Q, \quad K = F^u_{token} W_K, \quad V = F^u_{token} W_V, \tag{1}$$

$$\mathbf{P}^u_{soft} = \mathrm{Linear}\left(\mathrm{softmax}\left(\frac{QK^\top}{\sqrt{D}}\right)V\right), \tag{2}$$

where W_Q, W_K, W_V are learnable projections.

Kolmogorov-Arnold-enhanced Consensus Aggregation (KaCA) novelly performs learnable B-spline fitting ability of KAN (Liu et al., 2025) across prototypes from different source samples, adaptively aggregating consensus to suppress source-specific bias. Specifically, the labeled anchor $\mathbf{P}^l$ is first fused with the long-term memory prototype $\mathbf{P}^u_{mem}$ to form the query anchor $\mathbf{P}_{query}$ by:

$$\mathbf{P}_{query} = \mathrm{KANLinear}\left(\hat{\mathbf{P}}^l \oplus \hat{\mathbf{P}}^u_{mem}\right), \tag{3}$$

where $\hat{(\cdot)}$ denotes L2 normalization and $\oplus$ denotes concatenation. The retrieved prototype $\mathbf{P}^u_{soft}$ is then fused with $\mathbf{P}_{query}$ to yield the final consensus prototype $\mathbf{P}_{con}$:

$$\mathbf{P}_{con} = \mathrm{KANLinear}\left(\hat{\mathbf{P}}_{query} \oplus \hat{\mathbf{P}}^u_{soft}\right), \tag{4}$$

where $\mathrm{KANLinear}(x) = \sum_i h_i(x_i)$ is the KAN layer with learnable B-spline functions $h(\cdot)$.

Summary of Advantage: AKaPL novelly designs attention-guided cross-sample retrieval and KAN-enhanced consensus aggregation, promoting reliable and debiased prototype representations.

3.2 CPD FOR YIELDING COMPLEMENTARY VERIFIABILITY

CPD (Fig. 3(c)) converts the prototype containing consensus information into a class-specifying semantic prompt, driving SAM's automatic segmentation of unlabeled data and yielding more accurate pseudo-labels while providing a reliable verification signal complementary to the prototype branch.

The consensus prototype $\mathbf{P}_{con}$ is first projected into a semantic embedding $E_{proto} = \mathrm{LN}(\mathrm{Linear}(\mathbf{P}_{con}))$. Since $\mathbf{P}_{con}$ resides in the prototype branch's own feature space and carries no spatial grounding within SAM's frozen image embedding space, we further align it with SAM's native representation. Specifically, the mask (Y^l for labeled data, $\tilde{Y}^u$ for unlabeled data) is interpolated to the resolution of the frozen image embedding F_{img}, denoted M', and F_{img} is averaged over the foreground region defined by M' to extract a grounded embedding:

$$E_{grounded} = \frac{\sum(F_{img} \odot M')}{\sum M'}. \tag{5}$$

The two embeddings are then fused into the sparse prompt embedding, aligning the prototype branch's consensus semantics with SAM's native spatial evidence:

$$E_{sparse} = \mathrm{LayerNorm}(\mathrm{GELU}(\mathrm{Linear}(E_{proto} \oplus E_{grounded}))). \tag{6}$$

E_{sparse} and the frozen image embedding F_{img} condition SAM's mask decoder to produce the segmentation prediction S^u, from which the SAM prototype $\mathbf{P}^u_{sam}$ is derived via masked average pooling over F^u. Only the mask decoder and the prototype-to-prompt projection remain trainable, while the image encoder and prompt encoder are frozen.

Summary of Advantage: CPD novelly leverages consensus semantics as the prompt, yielding more accurate grounded pseudo-labels and providing complementary verification signals.

3.3 AVGR FOR CERTIFYING VERIFIABLE SELF-IMPROVEMENT

AVGR (Fig. 3(d)) certifies each candidate prototype update against the trustworthy labeled anchor through a accept/reject gating mechanism admitting only verified updates into a long-term memory prototype that serves as the evolving benchmark for subsequent gating, guarding against blind updates while certifying verifiable self-improvement. Specifically, given the labeled anchor $\mathbf{P}^l$, the memory prototype $\mathbf{P}^u_{mem}$, the soft-matched prototype $\mathbf{P}^u_{soft}$, and the SAM-verified prototype $\mathbf{P}^u_{sam}$, we define the cosine distance $dis(\cdot,\cdot) = 1 - \cos(\cdot,\cdot)$ and compute three quantities:

$$\begin{aligned} dis_{l\text{-}usam} &= dis(\mathbf{P}^l, \mathbf{P}^u_{sam}), \qquad dis_{l\text{-}umem} = dis(\mathbf{P}^l, \mathbf{P}^u_{mem}), \\ dis_{usoft\text{-}usam} &= dis(\mathbf{P}^u_{soft}, \mathbf{P}^u_{sam}), \end{aligned} \tag{7}$$

where $dis_{l\text{-}usam}$ and $dis_{l\text{-}umem}$ denote the proximity of the SAM-verified prototype and the memory prototype to the labeled anchor, respectively, and $dis_{usoft\text{-}usam}$ denotes the agreement between two independently derived unlabeled prototypes. An update is accepted only if all three conditions hold jointly against a predefined threshold α:

$$\begin{aligned} &\mathrm{cond}_a : \alpha > dis_{l\text{-}usam}, \qquad \mathrm{cond}_b : dis_{l\text{-}umem} > dis_{l\text{-}usam}, \\ &\mathrm{cond}_c : \alpha > dis_{usoft\text{-}usam}. \end{aligned} \tag{8}$$

cond_a bounds the SAM-verified prototype's absolute deviation from the anchor; cond_b enforces strict improvement over the current memory; cond_c cross-validates against the separately retrieved

soft-matched prototype. When all three conditions are satisfied, the memory prototype is updated via an EMA rule that blends the current memory with the newly verified prototype:

$$\mathbf{P}_{mem}^{u,t+1} \leftarrow \beta\, \mathbf{P}_{mem}^{u,t} + (1-\beta)\, \mathbf{P}_{sam}^{u,t}, \tag{9}$$

where β controls the speed of memory update. When any condition fails, the update is rejected and the memory prototype is left unchanged.

To translate this verification outcome into optimization, the accept/reject decision reweights the prototype contrastive loss:

$$\mathcal{L}_{proto}^{contra} = w_{pro} \cdot (\mathcal{L}_{contra}(\mathbf{P}_l, \mathbf{P}_{sam}^u) + \mathcal{L}_{contra}(\mathbf{P}_{mem}^u, \mathbf{P}_{sam}^u)) + \mathcal{L}_{contra}(\mathbf{P}_l, \mathbf{P}_{mem}^u), \tag{10}$$

where $w_{pro} = 0.2$ if accepted and $w_{pro} = 1$ if rejected. For a prototype pair $(\mathbf{P}_a, \mathbf{P}_b)$, $\mathcal{L}_{contra}(\cdot,\cdot)$ is defined as:

$$\mathcal{L}_{contra}(\mathbf{P}_a, \mathbf{P}_b) = -\log \frac{\exp(\mathrm{sim}(\mathbf{P}_a, \mathbf{P}_b)/\tau)}{\sum_{\mathbf{P}_n \in \mathcal{N}} \exp(\mathrm{sim}(\mathbf{P}_a, \mathbf{P}_n)/\tau)}, \tag{11}$$

where $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity, τ is the temperature parameter, and $\mathcal{N}$ denotes the set of negative prototypes from other classes.

Summary of Advantage: AVGR novelly designs a accept/reject gating mechanism against the trustworthy labeled anchor and a self-tightening long-term memory, jointly certifying verifiable self-improving while guarding against blind updates.

3.4 Verification-Gated Joint Optimization

Structural verification knowledge is bridged between the self-improving consensus-enhanced prototype-prompted SAM branch and the segmentation network via consistency regularization, enabling the two branches to synergistically learn from each other. We define the hybrid loss as $\mathcal{L}_{hybrid} = \frac{1}{2}(\mathcal{L}_{Dice} + \mathcal{L}_{CE})$. For the segmentation network, we apply supervised loss on labeled data $\mathcal{L}_{seg}^l = \mathcal{L}_{hybrid}(P^l, Y^l)$ and consistency regularization on unlabeled data with pseudo-labels $\tilde{Y}^u$ from the teacher model $\mathcal{L}_{seg}^u = \mathcal{L}_{hybrid}(P^u, \tilde{Y}^u)$. We then enforce consistency between the segmentation network and the prototype-prompted SAM branch with $\mathcal{L}_{proto}^l = \mathcal{L}_{hybrid}(S^l, Y^l)$ and $\mathcal{L}_{proto}^u = \mathcal{L}_{hybrid}(S^u, \tilde{Y}^u)$. The final training objective combines all components:

$$\mathcal{L} = \mathcal{L}_{seg}^l + \mathcal{L}_{seg}^u + \mathcal{L}_{proto}^l + \lambda_{gs}\left(\mathcal{L}_{proto}^u + \mathcal{L}_{proto}^{contra}\right), \tag{12}$$

where λ_{gs} is a time-dependent Gaussian warming-up function that gradually increases consensus learning contribution during training.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on three datasets: **(a)** *NIH-PAN* (Roth et al., 2016), 80 abdominal contrast-enhanced CT scans (60 training/20 validation), following (Chen et al., 2024; Lu et al., 2023). **(b)** *MSD-PAN* (Antonelli et al., 2022), 281 pancreas CT scans with pancreas and tumor merged into one foreground class as in (Zhang et al., 2021; Gou et al., 2025) (168/57/56 for train/val/test). **(c)** *BraTS2019* (Menze et al., 2015), 335 multimodal MRI scans of glioblastoma; we use T2-FLAIR for whole tumor segmentation (250/25/60 for train/val/test), following (Xu et al., 2022). **Implementation Details.** SI-CP2SAM is implemented in PyTorch on a single RTX 3090 GPU, with V-Net (Milletari et al., 2016) as the backbone for all methods. We use Adam (lr=0.001), batch size 4 (2 labeled, 2 unlabeled), and augment via random cropping and copy-paste (Bai et al., 2023; Yun et al., 2019). All images are cropped into $96 \times 96 \times 96$ patches following (Luo et al., 2022; Lu et al., 2023). We report Dice, Jaccard, HD95, and ASD, capturing both region overlap and boundary accuracy.

4.2 Results and Analysis

Comparison with State-of-the-Arts (Table 1, Fig. 4). Table 1 compares our SI-CP2SAM against other state-of-the-art SSMIS methods on three datasets under different labeled settings. Our method

Table 1: Comparison of different methods on NIH-PAN, MSD-PAN and BraTS2019 datasets demonstrating the advantages of our method. Results are reported as mean±std. Best results are in red, second-best in blue.

Method	Type	NIH-PAN 5% (3 Labeled / 57 Unlabeled)				NIH-PAN 10% (6 Labeled / 54 Unlabeled)			
		Dice ↑	Jaccard ↑	HD95 ↓	ASD ↓	Dice ↑	Jaccard ↑	HD95 ↓	ASD ↓
V-Net	SL	24.64±21.18	15.84±14.91	42.74±42.43	7.00±6.22	53.76±17.02	38.58±15.79	20.35±14.62	3.17±5.20
UA-MT	CR	34.37±19.25	22.46±14.13	49.19±24.29	14.69±14.84	67.66±14.73	52.78±14.91	14.64±11.90	2.84±1.97
BCP	CR	74.90±7.39	60.39±8.91	11.38±11.96	2.00±1.04	82.59±4.36	70.57±6.10	6.10±4.05	1.73±1.06
AD-MT	CR	67.46±5.48	51.15±6.07	13.70±6.91	1.93±0.60	80.73±6.52	68.18±8.77	7.03±6.66	1.39±0.58
CPCL	PL	50.88±21.96	36.72±17.84	33.16±23.76	10.19±14.51	69.09±13.89	54.23±14.60	17.19±15.45	3.69±3.76
UPCoL	PL	51.77±18.84	36.86±15.41	32.90±24.42	3.32±3.03	72.78±13.41	58.72±14.45	14.00±15.52	4.07±3.87
KP2L	PL	57.78±20.69	43.02±16.66	23.95±17.70	5.19±8.92	77.29±11.61	64.17±13.33	10.90±15.75	1.68±0.99
BaPC	PL	41.59±18.19	27.88±14.30	51.30±30.85	3.86±5.58	75.10±11.18	61.24±12.50	12.26±15.82	3.26±4.49
MPCL	PL	59.01±17.77	43.72±14.95	23.85±17.55	3.16±3.84	76.46±9.99	62.81±11.48	9.93±14.33	1.80±0.85
Ours	PL	79.54±5.33	66.33±6.96	6.97±4.39	1.78±0.67	83.70±4.00	72.16±5.73	4.86±2.00	1.43±0.64

Method	Type	MSD-PAN 10% (17 Labeled / 151 Unlabeled)				MSD-PAN 20% (34 Labeled / 134 Unlabeled)			
		Dice ↑	Jaccard ↑	HD95 ↓	ASD ↓	Dice ↑	Jaccard ↑	HD95 ↓	ASD ↓
V-Net	SL	32.42±30.18	23.61±23.83	30.55±32.77	6.34±8.64	56.51±21.50	42.28±19.55	20.29±18.89	4.57±4.56
UA-MT	CR	40.15±26.01	28.55±21.20	33.53±16.67	12.72±8.85	55.62±23.37	41.88±20.90	29.30±31.19	2.25±1.80
BCP	CR	57.63±18.02	42.67±17.48	21.58±13.26	7.05±4.45	62.63±17.53	47.77±17.27	19.72±25.49	2.04±1.44
AD-MT	CR	59.74±21.58	45.57±19.60	20.86±26.00	3.22±4.33	62.14±23.61	48.67±21.27	22.96±31.38	2.66±4.16
CPCL	PL	29.95±30.39	21.86±23.90	40.57±40.89	6.19±9.77	50.64±26.84	38.00±23.01	36.38±36.32	2.84±3.84
UPCoL	PL	43.35±30.51	32.46±24.78	41.42±37.80	8.79±11.55	59.56±21.21	45.34±19.61	25.92±29.53	2.78±4.20
KP2L	PL	46.46±28.71	34.61±23.50	35.98±34.29	8.28±10.72	61.32±19.81	46.86±18.69	22.59±25.98	3.17±3.96
BaPC	PL	42.93±30.06	32.04±24.83	44.00±40.07	6.93±11.43	56.15±25.33	42.85±21.91	32.83±34.81	3.30±5.24
MPCL	PL	52.23±23.55	38.56±20.37	32.39±32.53	4.80±8.74	62.25±18.41	47.56±17.87	23.16±26.98	2.06±1.25
Ours	PL	62.76±16.80	47.76±16.81	17.26±20.74	2.62±2.66	65.84±17.00	51.20±16.95	15.70±23.00	2.25±2.23

Method	Type	BraTS2019 5% (13 Labeled / 237 Unlabeled)				BraTS2019 10% (25 Labeled / 225 Unlabeled)			
		Dice ↑	Jaccard ↑	HD95 ↓	ASD ↓	Dice ↑	Jaccard ↑	HD95 ↓	ASD ↓
V-Net	SL	71.40±29.12	62.03±29.09	16.62±18.57	3.69±5.65	76.94±20.44	66.18±22.71	13.41±15.45	2.06±2.54
UA-MT	CR	77.73±20.92	67.41±22.82	15.78±18.89	4.45±7.04	80.84±16.40	70.44±19.39	15.73±20.38	4.65±6.13
BCP	CR	81.79±14.01	71.30±17.90	16.61±19.15	4.54±5.62	79.29±17.95	68.75±20.93	16.00±18.33	4.55±5.69
AD-MT	CR	81.29±17.37	71.39±20.13	9.86±12.40	1.48±1.04	82.84±13.84	72.70±17.04	10.77±13.07	2.98±4.10
CPCL	PL	77.62±21.96	67.60±23.66	14.52±21.09	6.26±14.61	82.60±14.65	72.61±18.53	10.82±13.94	3.41±5.05
UPCoL	PL	77.69±18.94	66.89±21.98	13.73±14.94	2.17±2.50	80.21±18.37	70.08±21.07	13.07±18.50	3.14±11.30
KP2L	PL	78.12±20.76	68.05±23.46	13.05±15.02	1.45±0.98	83.94±12.45	74.05±16.51	9.69±12.36	1.35±0.80
BaPC	PL	78.17±22.63	68.59±24.18	13.32±15.90	1.99±2.75	80.22±17.26	69.91±20.65	10.76±12.63	2.29±2.99
MPCL	PL	80.10±15.87	69.35±19.33	11.86±13.63	1.97±2.33	83.40±13.23	73.45±17.12	9.83±13.17	1.70±2.25
Ours	PL	82.22±14.72	72.06±18.24	10.30±13.23	1.94±2.58	85.37±10.43	75.77±14.29	9.13±12.87	1.80±2.64

consistently achieves superior performance across all metrics, with the gains most pronounced on NIH-PAN under extreme supervision scarcity: with only 3 labeled cases, SI-CP2SAM surpasses the second-best method by 4.64% in Dice and reduces HD95 by 4.41, demonstrating that verification-gated self-improving effectively mitigates the Blind Update problem and yields more reliable pseudo-labels under severe label scarcity. On the overall segmentation metrics, Dice and Jaccard, our method outperforms all competing approaches across all three datasets. On the boundary-sensitive metrics, HD95 and ASSD, SI-CP2SAM surpasses most competing methods, reflecting more accurate boundary delineation under limited supervision. Moreover, our method attains the lowest standard deviation across most comparisons, indicating superior robustness and consistency across samples. Fig. 4 further visualizes representative cases, where SI-CP2SAM preserves topological completeness and achieves more accurate boundary delineation compared to competing methods.

Ablation on Key Components (Table 2) Table 2 reports the ablation results validating the effectiveness of each proposed key component. Comparing rows 2 and 3, incorporating AKaPL improves Dice by 2.87%, demonstrating that cross-sample retrieval with KAN-enhanced consensus aggregation produces more reliable and debiased prototypes than MAP. Comparing rows 3 and 4, CP2SAM without gating slightly worsens 95HD and ASD despite improving Dice, an instability later resolved by AVGR's gating in row 7. Comparing rows 3, 4 and 5 exposes the gate's reliance on SAM: without CP2SAM, the gate lacks a decorrelated verification signal and Dice degrades by 1.21%, confirming that SAM-based verification is indispensable for the self-improving mechanism to function correctly. Comparing rows 6 and 7, both incorporating AKaPL and CP2SAM, the three-condition accept/reject gate yields a modest yet critical refinement, lifting Dice from 82.81% to 83.30%, confirming that the gating mechanism alone provides a precise improvement in suppressing the Blind Update problem. The full model attains the best results across all four metrics, indicating that debiased prototype

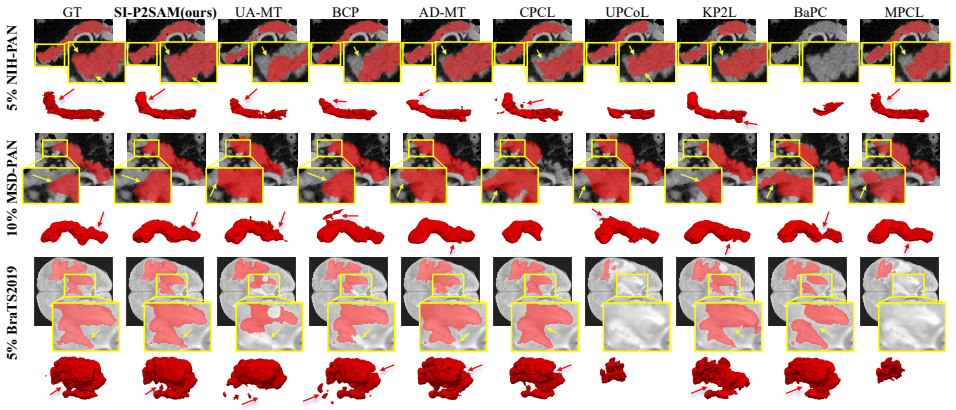


Figure 4: 2D slice and 3D visualizations on different datasets illustrate our method's superiority in preserving topological completeness and achieving near-ground-truth segmentation accuracy.

Table 2: Ablation study of key components on 10% NIH-PAN, validating the effectiveness of our proposed designs. Reported improvements are measured relative to the MT+PL baseline.

#	MT	PL	AKaPL	CPD	AVGR		Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
					w/o Gate	w/ Gate				
1	✓						66.60	52.55	18.97	6.12
2	✓	✓					77.69	64.16	9.15	2.09
3	✓	✓	✓				80.56 (↑2.87)	67.77 (↑3.61)	5.79 (↓3.36)	1.56 (↓0.53)
4	✓	✓	✓	✓			82.51 (↑4.82)	70.43 (↑6.27)	6.41 (↓2.74)	1.89 (↓0.20)
5	✓	✓	✓			✓	81.30 (↑3.61)	68.74 (↑4.58)	6.23 (↓2.92)	1.82 (↓0.27)
6	✓	✓	✓	✓	✓		82.81 (↑5.12)	70.88 (↑6.72)	5.69 (↓3.46)	1.68 (↓0.41)
7	✓	✓	✓	✓		✓	**83.30** (↑5.61)	**71.57** (↑7.41)	**5.42** (↓3.73)	**1.62** (↓0.47)

learning, decorrelated verification, and certified self-improvement are complementary and jointly necessary for optimal performance.

Ablation on fusion strategy Design (Fig. 5) To validate the design of AKaPL in aggregating consensus across samples, we compare three fusion variants: (1) Average Fusion, naively averaging retrieved prototypes; (2) Linear Fusion, replacing KAN aggregation with a standard linear layer; and (3) KAN Fusion (Ours), the full AKaPL design. Attention-driven cross-sample correspondence with a learnable fusion layer narrows the labeled-unlabeled prototype gap over naive averaging, and replacing the linear layer with KAN-based aggregation narrows it further, confirming the gain stems from KAN's fitting capacity rather than the fusion layer alone.

Ablation on SAM Prompt Design (Fig. 6) To validate the effectiveness of the prototype-conditioned prompt strategy in CP2SAM, we compare it against MedSAM, MedSAM3D, and two ablated variants: Ours (w/o Mask), which removes mask-guided alignment between prototypes and SAM features, and Ours (w/o Proto), which retains only the pseudo-label as prompt. MedSAM and MedSAM3D rely on geometric prompts that offer coarse spatial cues and are sensitive to pseudo-label noise. The comparison shows that aligning the two feature spaces is essential, giving SAM sample-specific semantic context that geometric cues cannot provide; our full prototype-conditioned prompt leverages this context to strengthen SAM's zero-shot segmentation, yielding more accurate and consistent results.

Ablation on Sensitivity of Prototype Contrastive Loss Weight w_{pro} (Fig. 7). To validate the effectiveness of w_{pro} in regulating the accept-reject mechanism over prototype contrastive supervision, we examine Dice performance under $w_{pro} \in \{0.1, 0.2, 0.3, 0.5, 0.8, 1.0\}$ across three datasets. Dice remains stable and near-optimal for $w_{pro} < 0.5$, then declines progressively as over-weighted contrastive supervision interferes with the verification signal. At $w_{pro} = 1$ (mechanism effectively disabled, contrastive loss applied unconditionally), performance is lowest across all datasets, confirming that unfiltered supervision degrades segmentation without verification gating. We set $w_{pro} = 0.2$ for all main experiments.

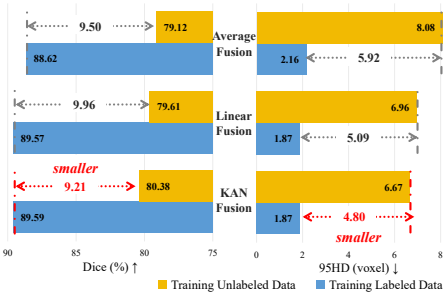


Figure 5: Comparison of different fusion strategy between labeled and unlabeled data on 10% NIH-PAN demonstrating KAN fusion aggreagetes consensus across samples.

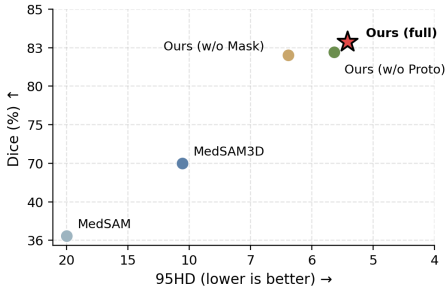


Figure 6: Comparison of MedSAM, MedSAM3D and our different settings on SAM demonstrating the effectiveness of our designed prototype-prompted SAM strategy.

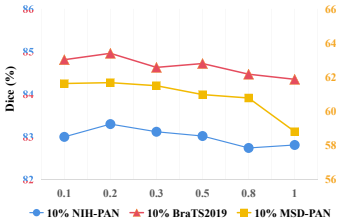


Figure 7: Sensitivity of prototype contrastive loss weight w_{pro}, validating its effect in regulating the accept-reject mechanism.

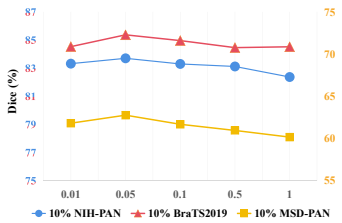


Figure 8: Sensitivity of conditional threshold α, validating its effectiveness in regulating gating strictness within the accept-reject mechanism.

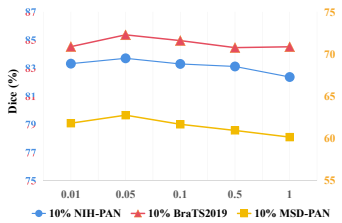


Figure 9: Sensitivity of memory update momentum β, validating the necessity of regulated update pacing for stable long-term prototype memory.

Ablation on Sensitivity of Conditional Threshold α (Fig. 8). To validate the effectiveness of α in regulating gating strictness within the accept-reject mechanism, we examine Dice performance under $\alpha \in \{0.01, 0.05, 0.1, 0.5, 1.0\}$ across three datasets, where α jointly governs cond_a and cond_c. Performance remains stable and near-optimal for $\alpha \leq 0.1$, since the gate has already filtered low-confidence updates within this range; Dice declines progressively as α increases to 0.5 and 1.0, as more unreliable prototype updates are admitted into memory, weakening AVGR's verification guarantee. We set $\alpha = 0.05$ as the default, balancing filtering strictness against unnecessary rejection of valid updates.

Ablation on Sensitivity of Long-Term Memory Update Momentum β (Fig. 9). To validate the necessity of regulated update pacing for stable maintenance of the long-term unlabeled prototype memory, we examine Dice performance under $\beta \in \{0.80, 0.90, 0.95, 0.99\}$ across three datasets. Performance follows a non-monotonic trend, peaking at $\beta = 0.95$: a smaller β updates the memory too aggressively, absorbing prototypes before consensus stabilizes and leaving it susceptible to noise; a larger β over-smooths the memory, delaying incorporation of verified updates and causing it to lag behind the evolving model. We set $\beta = 0.95$ as the default, balancing update stability and responsiveness.

5 Conclusion

In this paper, we propose SI-CP2SAM, the first self-improving SSMIS framework, which couples prototype learning with SAM for verification-gated self-improvement. By admitting prototype updates through an accept-reject gate against reliable anchors grounded in labeled data, SI-CP2SAM turns self-improvement from an unverified process into one that is provably grounded in the ground-truth distribution. Extensive experiments on three datasets show that SI-CP2SAM consistently outperforms existing SSMIS methods, confirming the value of verification-gated self-improving as a general principle for semi-supervised learning.

AI USE STATEMENT

In this work, we only used generative AI tools for polishing and refining the English writing of the manuscript. We have reviewed all AI-assisted work manually and verified that the AI-polished text faithfully reflects our original ideas and intended meaning. We take responsibility for the final content of this work.

REFERENCES

Michela Antonelli, Annika Reinke, Spyridon Bakas, et al. The medical segmentation decathlon. *Nature Communications*, 13:4128, 2022. doi: 10.1038/s41467-022-30695-9.

Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11514–11524, 2023.

Yali Bi, Enyu Che, Yinan Chen, Yuanpeng He, and Jingwei Qu. Multi-prototype-based embedding refinement for medical image segmentation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889249.

Yantao Chen, Yong Ma, Xiaoguang Mei, Lin Zhang, Zhigang Fu, and Jiayi Ma. Triple-task mutual consistency for semi-supervised 3d medical image segmentation. *Computers in Biology and Medicine*, 175:108506, 2024. doi: 10.1016/j.compbiomed.2024.108506.

Yue Gou, Yuming Xing, Shengzhu Shi, and Zhichang Guo. Diffusion probabilistic multi-cue level set for reducing edge uncertainty in pancreas segmentation. *Biomedical Signal Processing and Control*, 106:107744, 2025. doi: 10.1016/j.bspc.2025.107744.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023. doi: 10.1109/ICCV51070.2023.00371.

Yuqi Liu, Yufei Chen, Wei Fu, Xiaodong Yue, and Thierry Denœux. Kp2l: Knowledge-driven pyramid prototype learning for semi-supervised medical image segmentation. *Knowledge-Based Systems*, 340:115662, 2026a. doi: 10.1016/j.knosys.2026.115662.

Yuqi Liu, Yufei Chen, Wei Fu, Xiaodong Yue, and Shuo Li. Embracing intra-class heterogeneity for semi-supervised medical image segmentation: From diversity to precision. *Medical Image Analysis*, 113:104199, 2026b. doi: 10.1016/j.media.2026.104199.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov–arnold networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wenjing Lu, Jiahao Lei, Peng Qiu, Rui Sheng, Jinhua Zhou, Xinwu Lu, and Yang Yang. Upcol: uncertainty-informed prototype consistency learning for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 662–672. Springer, 2023. doi: 10.1007/978-3-031-43901-8_63.

Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8801–8809, 2021. doi: 10.1609/aaai.v35i10.17066.

Xiangde Luo, Guotai Wang, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Dimitris N Metaxas, and Shaoting Zhang. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, 80:102517, 2022. doi: 10.1016/j.media.2022.102517.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. doi: 10.1038/s41467-024-44824-z.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34 (10):1993–2024, 2015. doi: 10.1109/TMI.2014.2377694.

Juzheng Miao, Cheng Chen, Keli Zhang, Jie Chuai, Quanzheng Li, and Pheng-Ann Heng. Cross prompting consistency with segment anything model for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 167–177. Springer, 2024. doi: 10.1007/978-3-031-72120-5_16.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. IEEE, 2016.

Andrea Moglia, Matteo Cavicchioli, Luca Mainardi, and Pietro Cerveri. Deep learning for pancreas segmentation on computed tomography: a systematic review. *Artificial Intelligence Review*, 58 (8), 2025.

Holger R. Roth, Amal Farag, Evrim B. Turkbey, Le Lu, Jiamin Liu, and Ronald M. Summers. Data from pancreas-ct, 2016. The Cancer Imaging Archive dataset.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30:1195–1204, 2017.

Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, et al. Sam-med3d: a vision foundation model for general-purpose segmentation on volumetric medical images. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):17599–17612, 2025. doi: 10.1109/TNNLS.2025.3586694.

Yongchao Wang, Bin Xiao, Xiuli Bi, Weisheng Li, and Xinbo Gao. Boundary-aware prototype in semi-supervised medical image segmentation. *IEEE Transactions on Image Processing*, 33: 5456–5467, 2024. doi: 10.1109/TIP.2024.3463412.

Zhikai Wei, Wenhui Dong, Peilin Zhou, Yuliang Gu, Zhou Zhao, and Yongchao Xu. Prompting segment anything model with domain-adaptive prototype for generalizable medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 533–543. Springer, 2024. doi: 10.1007/978-3-031-72111-3_50.

Zhe Xu, Yixin Wang, Donghuan Lu, Lequan Yu, Jiangpeng Yan, Jie Luo, Kai Ma, Yefeng Zheng, and Raymond Kai-yu Tong. All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3174–3184, 2022. doi: 10.1109/JBHI.2022.3162043.

Zhonghao Yan, Zijin Yin, Tianyu Lin, Xiangzhu Zeng, Kongming Liang, and Zhanyu Ma. Pgp-sam: Prototype-guided prompt learning for efficient few-shot medical image segmentation. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2025. doi: 10.1109/ISBI60581.2025.10980911.

Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part II 22*, pp. 605–613. Springer, 2019. doi: 10.1007/978-3-030-32245-8_67.

Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6022–6031, 2019. doi: 10.1109/ICCV.2019.00612.

Yichi Zhang, Jin Yang, Yuchen Liu, Yuan Cheng, and Yuan Qi. Semisam: Enhancing semi-supervised medical image segmentation via sam-assisted consistency regularization. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3982–3986. IEEE, 2024. doi: 10.1109/BIBM62325.2024.10821951.

Yichi Zhang, Bohao Lv, Le Xue, Wenbo Zhang, Yuchen Liu, Yu Fu, Yuan Cheng, and Yuan Qi. Semisam+: Rethinking semi-supervised medical image segmentation in the era of foundation models. *Medical Image Analysis*, 106:103733, 2025. doi: 10.1016/j.media.2025.103733.

Yue Zhang, Jiong Wu, Yilong Liu, Yifan Chen, Wei Chen, Ed X Wu, Chunming Li, and Xiaoying Tang. A deep learning framework for pancreas segmentation with multi-atlas registration and 3d level-set. *Medical Image Analysis*, 68:101884, 2021. doi: 10.1016/j.media.2020.101884.

Zheyuan Zhang, Lanhong Yao, Elif Keles, Yury Velichko, and Ulas Bagci. Deep learning algorithms for pancreas segmentation from radiology scans: A review. *Advances in Clinical Radiology*, 5(1): 31–52, 2023.