A Local-Rotation-Driven Global Consistency Framework with Dual-View Decoding for Semi-Supervised Medical Image Segmentation

Zhen Yang[1], **Dongshuai Zhang**[1], **Yunliang Qi**[1], **Guidong Zhang**[1], **Shouliang Li**[1] and **Shuai Wu**[2*]

[1]School of Information Science & Engineering, Lanzhou University
[2]School of Computer Science and Technology, Xidian University
{zhenyang, zhangdsh2024, qiyunliang, zhanggd, lishoul}@lzu.edu.cn, shuaiwu9@gmail.com

Abstract

Medical image segmentation is still challenging, especially in semi-supervised scenarios where limited annotations are expected to support both accurate boundary delineation and coherent anatomical structures. We propose LR-GCF, a Local-Rotation-Driven Global Consistency Framework that couples strong local geometric perturbations with global consistency regularization in a dual-view architecture. Given an image, LR-GCF transforms it to different variants through patch-wise jigsaw-rotated perturbations. Such variants are decoded using parallel Swin Transformer decoders and perform consistency regularization by virtue of cross pseudo supervision. A rotation prediction head is designed to predict the rotation of each local patch. This makes locally discriminative representations robust to rotations and helps the model recognize global anatomical structure under strong perturbations. To further stabilize global modeling, we introduce a rotation-aligned attention consistency loss that encourages the inter-patch correlation across different variants to be as similar as possible. Moreover, we design a Position Encoding Duplicating scheme to propagate positional encodings from the low-resolution token map to high-resolution token maps, which helps alleviate the misalignment problem in rotation-aligned attention consistency. Extensive experiments on multiple medical image segmentation benchmarks demonstrate that LR-GCF consistently achieves superior performance under limited supervision, by jointly optimizing local detail extraction and stable global anatomical modeling. Source code is available at https://github.com/shuaiaihang/LR-GCF.

1 Introduction

Numerous clinical imaging workflows fundamentally rely on image segmentation. This capability enables computer-aided diagnosis, treatment planning, radiotherapy contouring, and lesion evaluation [Han *et al.*, 2024]. While deep learning has delivered substantial improvements, most high-performance segmenters still require fully supervised training paradigms, necessitating extensive pixel-wise annotations [Wang *et al.*, 2021]. Manual labeling creates a significant bottleneck: specialists must typically trace volumetric data slice-by-slice. Moreover, the resultant masks frequently exhibit variability across different observers and imaging protocol. By capitalizing on unlabeled data to lessen this manual burden, semi-supervised approaches in medical segmentation thus represent a pragmatic pathway forward.

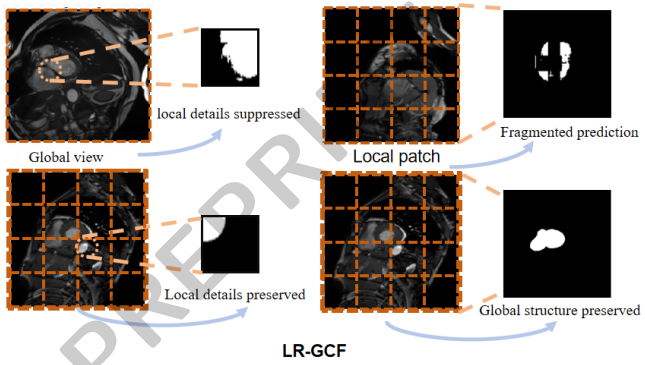

Figure 1: Global–local trade-off: global context may suppress fine details, whereas local patches can cause fragmented predictions. Our LR-GCF alleviates both issues.

The fundamental challenge in semi-supervised medical image segmentation lies in reconciling local precision with global anatomical plausibility. Fine-grained boundaries and textures depend largely on local evidence, yet clinically meaningful masks require coherent global context. With scarce annotations, models inevitably drift toward noisy local patterns or retreat into overly rigid global priors [Guan *et al.*, 2024; Zhang *et al.*, 2024]. Recent research tackles this balance from several architectural fronts. CNN-based approaches remain competitive through systematic extensions and meticulous re-benchmarking [Isensee *et al.*, 2023; Isensee *et al.*, 2024]. Hierarchical Transformers simultaneously strengthen long-range dependencies via structured attention mechanisms [Hatamizadeh *et al.*, 2022; He *et al.*, 2023]. Foundation and prompt-driven segmenters have further improved cross-domain generalization [Chen *et al.*,

*Corresponding author.

2023; Zhu *et al.*, 2024], while emerging state-space hybrids offer efficient alternatives for global modeling [Liu *et al.*, 2024; Ruan *et al.*, 2024].

Nevertheless, limited supervision leaves the model's internal reasoning sparsely constrained. Local features prove susceptible to artifactual cues, while learned global relations risk violating anatomical priors—particularly when local and global mechanisms remain tightly coupled [Guan *et al.*, 2024]. This challenge motivates architectures that explicitly regularize both shape encoding and relational coherence [Zhang *et al.*, 2024].

We therefore propose LR-GCF, a Local-Rotation-Driven Global Consistency Framework for semi-supervised medical segmentation. LR-GCF directly couples local patch-wise rotations with global structural constraints: from each image we derive two locally-rotated variants, encode them via a shared CNN backbone, and decode through parallel Swin Transformers. Consequently, learning under strong geometric perturbations becomes directly anchored by anatomical coherence.

To improve training with sparse annotations, we introduce complementary regularization strategies. First, an auxiliary Rotation Head operating on deep encoder features supervises patch-wise orientation, cultivating locally distinctive representations that resist rotational variations. This anchors local evidence to global anatomical priors despite aggressive perturbations. Second, we propose a rotation-synchronized attention consistency objective that constrains self-attention relationships across both views, fostering harmonized region-to-region reasoning beyond simple output agreement and bolstering anatomical soundness.

Additionally, independently learned multi-scale positional embeddings invite geometric drift—a difficulty amplified when inputs experience divergent local rotations. Our Positional Encoding Duplicating (PED) scheme counters this by deriving all positional encodings from a shared low-resolution anchor through scale-wise replication. This approach furnishes a consistent spatial reference throughout the decoding stack, calms attention fluctuations, and refines interview correspondence. In aggregate, LR-GCF harmonizes locally rotation-driven feature extraction with globally relational constraints, yielding resilient segmentation from limited supervision.

The main contributions are summarized as follows:

- We introduce LR-GCF, a dual-stream framework that integrates local rotational perturbations with global structural regularization for semi-supervised segmentation of medical images.
- We propose a localized rotation prediction module that cultivates orientation-robust, morphology-sensitive features, ensuring they preserve consistency with global anatomical priors under aggressive augmentations.
- We consolidate global relational regularization via rotation-synchronized attention consistency across self-attention matrices, coupled with PED-based propagation of multi-scale positional embeddings from a unified low-resolution anchor, jointly stabilizing Transformer decoding for all views and scales.
- Extensive experiments on multiple datasets demonstrate that LR-GCF consistently achieves state-of-the-art performance, with strong gains in both boundary accuracy and structural coherence under scarce-label settings.

2 Related Work

2.1 Semi-Supervised Medical Image Segmentation

Medical semi-supervised segmentation (SSMIS) confronts dense prediction tasks where sparse annotations coexist with abundant unlabeled scans, leveraging internally-generated predictions and auxiliary constraints as surrogate supervision [Jiao *et al.*, 2024]. Contemporary pipelines predominantly harness dual mechanisms—pseudo-label generation and consistency regularization—which operate synergistically to stabilize optimization when labeled data remains severely limited [Tarvainen and Valpola, 2018; Sohn *et al.*, 2020; Luo *et al.*, 2023].

For pseudo-label generation, temporal ensembling through mentor-learner architectures represents the dominant paradigm, wherein a gradually evolving teacher network synthesizes refined supervisory signals for unannotated volumes. Reliability enhancement typically incorporates uncertainty quantification or confidence-based filtering to suppress noisy pseudo-labels [Yu *et al.*, 2019; Sohn *et al.*, 2020]. Concurrently, mutual supervision across paired decoders mitigates self-reinforcement errors [Chen *et al.*, 2021b; Luo *et al.*, 2023], while region-mixing augmentations facilitate knowledge transfer from labeled to unlabeled instances [Olsson *et al.*, 2020]. Invariance-driven consistency constraints complement pseudo-labeling by enforcing stability against perturbations, with multi-scale implementation demonstrating particular efficacy for dense prediction tasks [Luo *et al.*, 2022]. Nevertheless, predominant reliance on output-space constraints biases optimization toward local texture patterns rather than long-range anatomical dependencies, yielding predictions that lack structural coherence under limited supervision [Jiao *et al.*, 2024].

2.2 Vision Transformers for Medical Image Segmentation

Self-attention mechanisms in transformer architectures capture distant dependencies while aggregating global context explicitly, making them increasingly attractive for medical segmentation tasks involving ambiguous boundaries or extensive anatomical regions [Shamshad *et al.*, 2023]. Initial implementations frequently embedded transformer encoders within U-shaped architectures such as TransUNet and UNETR to strengthen contextual inference [Chen *et al.*, 2021a; Hatamizadeh *et al.*, 2022]. For high-resolution imagery and volumetric data, window-based hierarchical attention mechanisms have emerged as viable solutions, spawning Swin-derived segmenters and subsequent refinements [Cao *et al.*, 2021; He *et al.*, 2023].

Complementing pure transformer models, extensive investigation targets CNN-Transformer amalgamation, merging convolutional spatial priors with transformer's global receptive field. Standard configurations integrate convolutional

feature extractors with transformer contextualization modules, subsequently fusing multi-scale information via skip pathways or stage-wise interactions [Zhang *et al.*, 2021; Xie *et al.*, 2021]. Although these hybrids boost performance, their local-global integration remains implicit. With sparse annotations, regional features frequently overwhelm optimization, producing texture-biased or disconnected predictions, while global reasoning depends on intermediate transformer representations (attention correlations and positional embeddings) prone to divergence across stages or augmented views.

Collectively, transformer-based and hybrid segmenters deliver powerful global modeling capabilities, yet preserving alignment between fine-grained boundary evidence and holistic anatomical consistency proves difficult under annotation scarcity. This challenge drives the need for direct regularization of intermediate transformer signals (attention patterns and positional correspondence) alongside shape-sensitive local representation learning, moving beyond sole reliance on output-level supervision.

3 Method

Given a labeled dataset $\mathcal{D}_L = \{(x_i^l, y_i^l)\}_{i=1}^N$ and an unlabeled dataset $\mathcal{D}_U = \{x_j^u\}_{j=1}^M$ with $M \gg N$, our framework aims to exploit strong local perturbations while preserving global anatomical consistency for semi-supervised segmentation. As illustrated in Fig. 2, for each input image x (either labeled or unlabeled), we construct two patch-wise jigsaw–rotated views $\{x^{(1)}, x^{(2)}\}$ via local patch rotation. Both views are fed into a shared CNN encoder to extract multi-scale features, which are then decoded by two parallel Swin Transformer decoders to produce segmentation predictions for each view. For labeled data, we apply standard supervised segmentation loss on both decoder outputs; for unlabeled data, we perform cross-pseudo supervision between decoders across the two views. To explicitly couple local and global learning, we attach a patch-wise Rotation Head on deep encoder features to predict each patch rotation state, and we further enforce rotation-aligned attention consistency on self-attention relation matrices to regularize global relational reasoning. In addition, we introduce **Positional Encoding Duplicating (PED)** to derive multi-scale positional embeddings from a shared low-resolution anchor, providing a unified geometric reference across decoder stages and stabilizing cross-view alignment.

3.1 Dual-Decoder Transformer Segmentation Framework

Segmentation Supervision

In each mini-batch, we generate two patch-wise rotated views for *all* samples. For a labeled pair $(x^l, y^l) \in \mathcal{D}_L$, we apply the same patch-wise rotation operator to the image and mask, obtaining two rotated pairs $\left(x^{l,(1)}, y^{l,(1)}\right)$ and $\left(x^{l,(2)}, y^{l,(2)}\right)$ with $v \in \{1, 2\}$. Pixel-wise supervision is computed *only* on the labeled subset.

Given a labeled view $x^{l,(v)}$, the dual-decoder network produces segmentation logits $\mathbf{S}_d^{l,(v)} \in \mathbb{R}^{K\times H\times W}$ for decoder $d \in \{1, 2\}$. We supervise each prediction using a composite loss combining cross-entropy and Dice:

$$\mathcal{L}_{\text{seg}}(\mathbf{S}, y) = \frac{1}{2}\Big(\mathcal{L}_{\text{CE}}(\mathbf{S}, y) + \mathcal{L}_{\text{Dice}}(\mathbf{S}, y)\Big). \tag{1}$$

The supervised loss is averaged over the two decoders and the two rotated views:

$$\mathcal{L}_{\text{sup}} = \frac{1}{4}\sum_{d\in\{1,2\}}\sum_{v\in\{1,2\}} \mathcal{L}_{\text{seg}}\Big(\mathbf{S}_d^{l,(v)},\, y^{l,(v)}\Big). \tag{2}$$

Cross-Pseudo Supervision

For unlabeled samples, we adopt a symmetric cross-pseudo supervision (CPS) strategy to exploit complementary predictions of the two decoders under two independently rotated views. Given an unlabeled image x^u, we construct two patch-wise rotated views $\{x^{u,(1)}, x^{u,(2)}\}$, and the dual-decoder network produces four segmentation logits $\mathbf{S}_d^{u,(v)} \in \mathbb{R}^{K\times H\times W}$, where $v \in \{1, 2\}$ indexes the view and $d \in \{1, 2\}$ indexes the decoder.

Since each view is generated by local patch rotations, we map the logits back to a canonical orientation using the recorded rotation index map $\mathbf{K}^{u,(v)}$:

$$\widetilde{\mathbf{S}}_d^{u,(v)} = \mathcal{T}_{\text{rot}}^{-1}\Big(\mathbf{S}_d^{u,(v)};\, \mathbf{K}^{u,(v)}\Big). \tag{3}$$

We then obtain hard pseudo-labels by

$$\widehat{y}_d^{u,(v)} = \underset{k\in\{0,\dots,K-1\}}{\text{argmax}}\ \text{softmax}\left(\widetilde{\mathbf{S}}_d^{u,(v)}\right)_k \tag{4}$$

We perform CPS in a symmetric cross-view and cross-decoder manner, yielding four equivalent supervision terms:

$$\begin{aligned}
\mathcal{L}_1^u &= \mathcal{L}_{\text{seg}}\Big(\widetilde{\mathbf{S}}_1^{u,(1)},\, \text{stopgrad}\Big(\widehat{y}_2^{u,(2)}\Big)\Big),\\
\mathcal{L}_2^u &= \mathcal{L}_{\text{seg}}\Big(\widetilde{\mathbf{S}}_1^{u,(2)},\, \text{stopgrad}\Big(\widehat{y}_2^{u,(1)}\Big)\Big),\\
\mathcal{L}_3^u &= \mathcal{L}_{\text{seg}}\Big(\widetilde{\mathbf{S}}_2^{u,(1)},\, \text{stopgrad}\Big(\widehat{y}_1^{u,(2)}\Big)\Big),\\
\mathcal{L}_4^u &= \mathcal{L}_{\text{seg}}\Big(\widetilde{\mathbf{S}}_2^{u,(2)},\, \text{stopgrad}\Big(\widehat{y}_1^{u,(1)}\Big)\Big),
\end{aligned} \tag{5}$$

where $\text{stopgrad}(\cdot)$ blocks gradients through the pseudo-label branch. In implementation, $\mathcal{L}_{\text{seg}}$ uses the same Dice+CE form as the supervised case, with pseudo-labels treated as hard targets.

Finally, the CPS loss on unlabeled data is computed as the average of the four terms:

$$\mathcal{L}_{\text{u}} = \frac{1}{4}\sum_{i=1}^{4}\mathcal{L}_i^u. \tag{6}$$

Patch-wise Rotation Prediction Loss

The rotation head is trained with *direct* supervision from the patch-wise rotations that we apply when constructing the two views; this loss is independent of cross-pseudo supervision. For each sample, we generate two rotated views $x^{(1)}$ and $x^{(2)}$ on a $P \times P$ grid, and record the corresponding rotation label maps $\mathbf{K}^{(1)}, \mathbf{K}^{(2)} \in \{0, 1, 2, 3\}^{P\times P}$, where label 0 indicates no rotation and labels $\{1, 2, 3\}$ correspond to

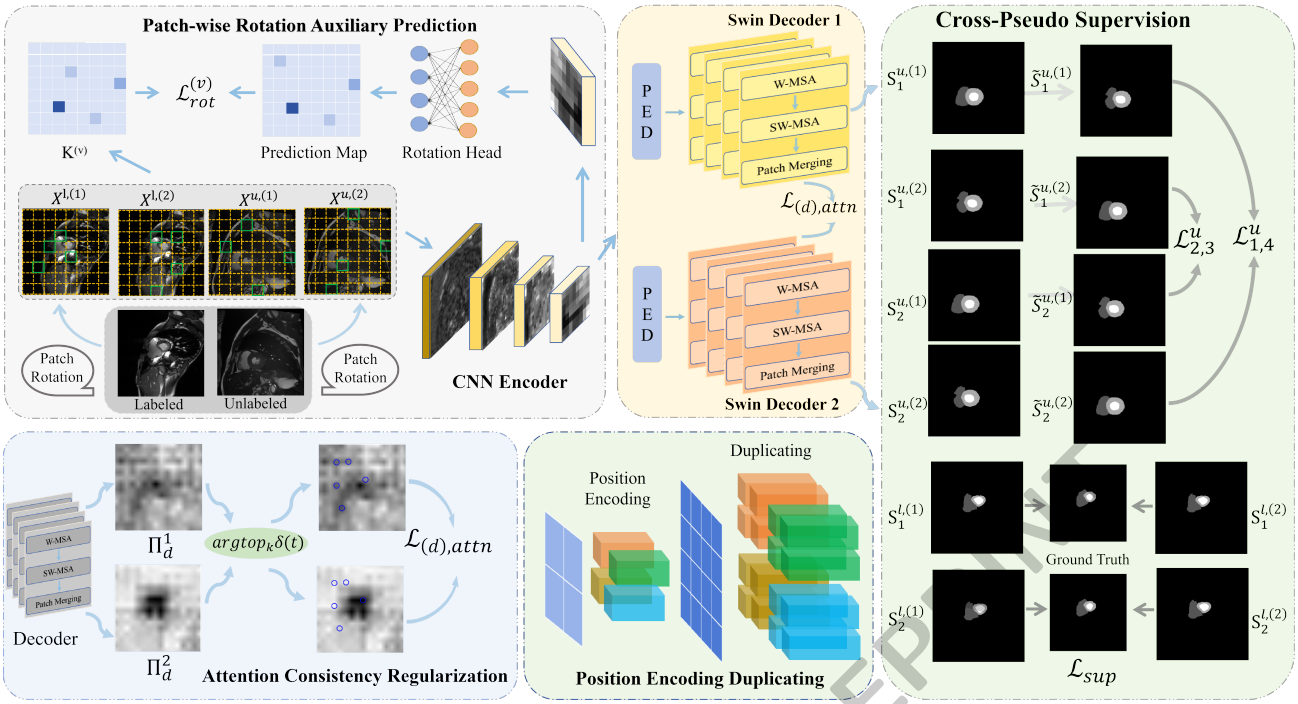

Figure 2: Overview of the proposed LR-GCF framework for semi-supervised medical image segmentation. For each input image, LR-GCF generates two patch-wise jigsaw-rotated views and feeds them into a shared CNN encoder to extract multi-scale representations. Two parallel Swin Transformer decoders produce view-specific segmentation predictions, which are optimized with supervised segmentation loss for labeled data and cross-pseudo supervision for unlabeled data. A patch-wise rotation head provides auxiliary orientation supervision to enhance rotation-aware local representation learning, while rotation-aligned attention consistency regularizes W-MSA/SW-MSA relations across views to preserve global anatomical coherence. The proposed Position Encoding Duplicating (PED) scheme further propagates a shared low-resolution positional anchor to multiple decoder scales, providing aligned multi-scale positional encodings for stable cross-view correspondence.

$\{90°, 180°, 270°\}$. Given view $v \in \{1, 2\}$, the rotation head predicts patch-level logits $\mathbf{R}^{(v)} \in \mathbb{R}^{4\times P\times P}$.

We supervise rotation prediction using patch-wise cross-entropy:

$$\mathcal{L}_{\text{rot}}^{(v)} = \frac{1}{P^2}\sum_{p=1}^{P}\sum_{q=1}^{P}\mathcal{L}_{\text{CE}}\left(\mathbf{R}_{:,p,q}^{(v)}, \mathbf{K}_{p,q}^{(v)}\right), \tag{7}$$

and average the loss over the two views:

$$\mathcal{L}_{\text{rot}} = \frac{1}{2}\sum_{v\in\{1,2\}}\mathcal{L}_{\text{rot}}^{(v)}. \tag{8}$$

Attention Consistency Regularization

We regularize the transformer behavior by enforcing attention consistency between the two patch-wise rotated views $X^{(1)}$ and $X^{(2)}$. This constraint is applied independently to each decoder $d \in \{1, 2\}$ and to all collected W-MSA / SW-MSA blocks.

For a given decoder d and one Swin attention block, let $A_d^{(v)} \in \mathbb{R}^{T\times T}$ be the head-averaged attention scores for view $v \in \{1, 2\}$, where T is the token number. We normalize attention into a probability distribution with temperature-scaled softmax:

$$\Pi_d^{(v)} = \text{softmax}\left(\frac{A_d^{(v)}}{\tau}\right) \in \mathbb{R}^{T\times T}, \tag{9}$$

where τ is the temperature. For each query token t, the vector $\Pi_d^{(v)}(t,:) \in \mathbb{R}^T$ represents how token t distributes its attention over all key tokens in the same block.

We measure the cross-view discrepancy for each query token by

$$\delta(t) = \left\|\Pi_d^{(1)}(t,:) - \Pi_d^{(2)}(t,:)\right\|_2, \quad t = 1, \ldots, T. \tag{10}$$

We then select a subset of query tokens with the largest discrepancies:

$$\mathcal{T}_{\text{top}} = \text{TopK}\big(\{\delta(t)\}_{t=1}^{T}, k\big), \qquad k = \max\big(1, \lfloor \rho T \rfloor\big), \tag{11}$$

where ρ is a fixed ratio (e.g., $\rho = 0.1$). The attention consistency loss for this block is defined as the mean squared error between the attention distributions of the selected query tokens:

$$\mathcal{L}_{\text{attn}}^{\text{blk}} = \frac{1}{|\mathcal{T}_{\text{top}}|}\sum_{t\in\mathcal{T}_{\text{top}}}\left\|\Pi_d^{(1)}(t,:) - \Pi_d^{(2)}(t,:)\right\|_2^2. \tag{12}$$

We compute $\mathcal{L}_{\text{attn}}^{\text{blk}}$ for all W-MSA and SW-MSA blocks across stages and average them to obtain the per-decoder loss $\mathcal{L}_{(d),\text{attn}}$. Finally, the overall attention regularization is the mean of the two decoders:

$$\mathcal{L}_{\text{attn}} = \frac{1}{2}\big(\mathcal{L}_{\text{attn}}^{(1)} + \mathcal{L}_{\text{attn}}^{(2)}\big). \tag{13}$$

Position Encoding Duplicating
To avoid geometric drift caused by learning independent positional embeddings at different decoder scales, we propose Position Encoding Duplicating. We learn a single absolute positional embedding on the coarsest token grid (e.g., 8×8 with 768 channels), and generate all higher-resolution encodings (e.g., $16 \times 16/32 \times 32/64 \times 64$ with $384/192/96$ channels) from this shared anchor. Concretely, the anchor embedding is reshaped to the 2D grid, its channel dimension is reduced by group-wise averaging to match the target scale, and it is then expanded to the target spatial resolution using nearest-neighbor duplication. The resulting positional encoding is added to the corresponding scale tokens before Swin attention blocks. By reusing the same anchor across all scales and both rotated views, PED provides a unified coordinate reference throughout decoding and stabilizes cross-view alignment under strong patch-wise rotations.

3.2 Overall Loss and Weight Scheduling

The overall training objective jointly optimizes the supervised segmentation loss (Sec. 3.1.1), the unlabeled cross-pseudo supervision loss (Sec. 3.1.2), the patch-wise rotation prediction loss (Sec. 3.1.3), and the attention consistency regularization (Sec. 3.1.4). The total loss for a mini-batch is defined as

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{rot}}\,\mathcal{L}_{\text{rot}} + \lambda_{\text{u}}\,\mathcal{L}_{\text{u}} + \lambda_{\text{attn}}\,\mathcal{L}_{\text{attn}}. \tag{14}$$

In our implementation, we set $\lambda_{\text{rot}} = 0.5$ and $\lambda_{\text{attn}} = 3$. The weight of the unlabeled loss, λ_{u}, is scheduled with a ramp-up strategy: it increases linearly from 0 to 1 as training proceeds, so that the model first learns from reliable labeled supervision and then gradually incorporates the unlabeled objective.

4 Experiments

4.1 Dataset

ACDC dataset. The ACDC dataset is a four-class cardiac MRI segmentation benchmark, including background, right ventricle, left ventricle, and myocardium. It contains scans from 100 patients. Following the commonly-used split protocol, the dataset is divided into 70, 10, and 20 patients for training, validation, and testing, respectively.

ISIC 2018 dataset. The ISIC 2018 dataset is a widely used benchmark for dermoscopic skin lesion segmentation, containing 3,694 RGB images. These images cover various types of skin lesions, with a focus on melanoma and non-melanoma lesions, and include pixel-level segmentation masks. The dataset is split into 2,594 training images, 100 validation images, and 1,000 test images. The original image sizes range from 540×722 to 4499×6748 pixels.

BraTS 2019 dataset. The BraTS 2019 dataset, released for the Brain Tumor Segmentation challenge, consists of preoperative multimodal MRI scans from 335 glioma patients, including T1, contrast-enhanced T1 (T1Gd), T2, and T2-FLAIR modalities. It includes 259 high-grade glioma (HGG) cases and 76 low-grade glioma (LGG) cases. Following [Xu *et al.*, 2023], we adopt the T2-FLAIR modality and focus on whole tumor segmentation. The dataset is divided into 250 training samples, 25 validation samples, and 60 testing samples.

4.2 Evaluation Metrics

We evaluate segmentation performance using four widely adopted metrics: Dice score (%), Jaccard score (%), 95% Hausdorff distance (95HD), and average surface distance (ASD). For 2D datasets, these metrics are computed in pixel space, while for volumetric datasets they are computed in voxel space. Dice and Jaccard quantify the overlap between the predicted segmentation and the reference annotation, reflecting global region-level agreement. In contrast, 95HD and ASD emphasize boundary accuracy by measuring the discrepancy between the predicted and ground-truth contours or surfaces, where 95HD reports a robust maximum distance excluding extreme outliers and ASD computes the mean surface distance. Together, these metrics provide a comprehensive assessment of segmentation quality, capturing both overall overlap consistency and fine-grained boundary precision.

4.3 Implementation Details

We conduct all experiments on a single NVIDIA RTX 3060 GPU with fixed random seeds for reproducibility.

ACDC dataset. All slices are resized to 256×256 and the single-channel images are replicated to three channels to match the network input format. We adopt standard data augmentation to enlarge the effective training set, including random rotation, random flipping, and color jittering. In addition, for the proposed patch-wise rotation augmentation, each 256×256 slice is partitioned into an 8×8 grid of non-overlapping patches, and local patch rotations are applied on this grid during training.

ISIC 2018 dataset. All dermoscopic images are resized to 256×256 to ensure consistency in data and standardized inputs across the dataset. The images are kept in their original RGB format to align with the network's input requirements. To enhance the training dataset, we apply typical data augmentation techniques, including random rotations, random horizontal/vertical flipping, and color jittering, which adjusts brightness, contrast, saturation, and hue. Additionally, for the patch-wise rotation augmentation, each 256×256 image is divided into an 8×8 grid of non-overlapping patches, and local rotations are applied to these patches during training.

BraTS 2019 dataset. The BraTS 2019 dataset includes preoperative MRI scans of glioma patients with modalities such as T1, T1Gd, T2, and T2-FLAIR. We use the T2-FLAIR images for tumor segmentation, resampled to an isotropic resolution of $1 \times 1 \times 1$ mm³. Each image has a size of $155 \times 240 \times 240$ pixels. During preprocessing, 2D slices of size 224×224 are extracted from the 3D volumes. The slices are then converted to single-channel images. For patch-wise rotation augmentation, each 224×224 slice is divided into a 7×7 grid of non-overlapping patches, with local rotations applied during training and rotation labels used for supervision.

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
UA-MT [Yu *et al.*, 2019]			81.65	70.64	6.88	2.02
SASSNet [Li *et al.*, 2020]			84.50	74.37	5.42	1.86
DTC [Luo *et al.*, 2023]			84.29	73.92	12.81	4.01
URPC [Luo *et al.*, 2022]			83.10	72.41	4.84	1.53
MC-Net [Wu *et al.*, 2021]			86.44	77.04	5.50	1.84
MC-Net+[Wu *et al.*, 2022a]			87.1	-	6.68	2.00
SS-Net [Wu *et al.*, 2022b]			86.78	77.67	6.07	1.40
LCLPL [Chaitanya *et al.*, 2023]			89.10	-	5.11	1.81
BCP [Bai *et al.*, 2023]	7(10%)	63(90%)	88.82	80.62	3.98	**1.17**
ABD [Chi *et al.*, 2024]			88.52	79.79	5.06	1.43
UC-Seg [Huang *et al.*, 2025b]			88.86	80.66	**1.58**	-
ABD&BCP [Chi *et al.*, 2024]			89.81	81.95	-	-
DUCiSC [Wu *et al.*, 2026]			89.82	82.28	-	-
MFHS [Liu *et al.*, 2026b]			88.28	-	1.73	-
RSTAC [Liu *et al.*, 2026a]			89.99	82.30	3.72	**0.80**
LR-GCF			**90.10**	**82.34**	3.71	1.38
UA-MT [Yu *et al.*, 2019]			85.7	-	4.06	1.54
SASSNet [Li *et al.*, 2020]			87.1	-	5.84	2.15
DTC [Luo *et al.*, 2023]			86.3	-	6.14	2.11
URPC [Luo *et al.*, 2022]			85.1	-	4.26	1.77
MC-Net [Wu *et al.*, 2021]	14(20%)	56(80%)	87.8	-	3.91	1.52
MC-Net+[Wu *et al.*, 2022a]			88.5	-	4.35	1.54
LCLPL [Chaitanya *et al.*, 2023]			90.5	-	5.11	1.81
MFHS [Liu *et al.*, 2026b]			89.21	-	1.98	-
LR-GCF			**90.61**	**83.1**	3.35	**1.27**

Table 1: Comparison with the state-of-the-art semi-supervised medical image segmentation methods on ACDC dataset.

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
U-Net	129(5%)	0	60.75	49.41	82.28	38.69
U-Net	259(10%)	0	76.16	66.55	43.76	17.21
U-Net	518(20%)	0	81.18	71.54	35.73	14.22
U-Net	2594(All)	0	86.84	78.81	24.58	9.90
UA-MT [Yu *et al.*, 2019]			77.88	68.76	46.15	17.80
FixMatch [Sohn *et al.*, 2020]			83.42	73.77	27.28	10.98
DTC [Luo *et al.*, 2023]			79.29	69.36	44.49	18.46
UbiMatch [Yang *et al.*, 2023]			84.19	75.19	25.57	10.08
CrossMatch [Zhao *et al.*, 2025]	259(10%)	2335(90%)	84.71	75.81	25.56	9.91
BS-NET [He *et al.*, 2024]			81.70	73.68	-	-
UC-Seg [Huang *et al.*, 2025b]			86.37	**78.84**	-	-
KnowSAM [Huang *et al.*, 2025a]			86.51	78.49	-	-
LR-GCF			**86.59**	76.9	**20.28**	**8.41**
UA-MT [Yu *et al.*, 2019]			82.21	72.47	39.46	16.09
FixMatch [Sohn *et al.*, 2020]			84.01	75.67	26.98	9.99
DTC [Luo *et al.*, 2023]	518(20%)	2076(80%)	83.07	73.86	31.49	12.37
UbiMatch [Yang *et al.*, 2023]			84.76	75.00	25.44	10.24
CrossMatch [Zhao *et al.*, 2025]			85.43	76.68	23.77	9.34
LR-GCF			**87.05**	**77.57**	**19.25**	**7.92**

Table 2: Comparison with the state-of-the-art semi-supervised medical image segmentation methods on ISIC 2018 dataset.

Method	Scans used		Metrics			
	Labeled	Unlabeled	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
DTC [Luo *et al.*, 2023]			80.01	69.78	11.56	1.94
SASSNet [Li *et al.*, 2020]			77.44	66.24	21.55	6.77
UA-MT [Yu *et al.*, 2019]			82.82	72.42	14.01	3.75
URPC [Luo *et al.*, 2022]			83.61	73.52	9.92	2.42
MC-Net [Wu *et al.*, 2021]			83.33	73.29	9.92	2.42
MC-Net+ [Wu *et al.*, 2022a]			82.18	72.18	10.81	1.74
AC-MC[Xu *et al.*, 2023]	25(10%)	225(90%)	81.60	71.04	10.76	2.06
CAML[Gao *et al.*, 2023]			83.91	74.18	13.23	3.54
ML-RPL[Su *et al.*, 2024]			81.60	71.23	11.63	2.99
Diff-CL[Guo *et al.*, 2025]			84.63	75.05	12.08	3.73
LR-GCF			**84.77**	**75.21**	**10.58**	**2.58**
DTC [Luo *et al.*, 2023]			81.17	70.96	11.51	2.16
SASSNet [Li *et al.*, 2020]			81.58	71.44	10.74	1.97
UA-MT [Yu *et al.*, 2019]			81.21	71.38	11.96	3.73
URPC [Luo *et al.*, 2022]			84.05	70.72	11.17	2.45
MC-Net [Wu *et al.*, 2021]			83.16	73.47	**8.68**	2.16
MC-Net+[Wu *et al.*, 2022a]			84.10	74.56	9.99	2.68
AC-MC [Xu *et al.*, 2023]	50(20%)	200(80%)	83.28	73.26	9.87	**1.79**
CAML[Gao *et al.*, 2023]			81.94	72.17	9.86	2.85
ML-RPL[Su *et al.*, 2024]			80.49	70.07	15.65	4.73
Diff-CL [Guo *et al.*, 2025]			85.45	**75.88**	9.05	2.6
LR-GCF			**86.53**	76.49	9.74	3.1

Table 3: Comparison with the state-of-the-art semi-supervised medical image segmentation methods on BraTs 2019 dataset.

PatchRot	RotHead	AttnCons	PED	Metrics			
				Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
-	-	-	-	85.37	75.01	5.33	2.13
✓	-	-	-	86.79	77.16	5.4	1.9
✓	✓	-	-	88.21	79.25	4.16	1.55
✓	✓	✓	-	89.31	81.31	4.01	**1.17**
✓	✓	✓	✓	**90.1**	**82.34**	**3.71**	1.38

Table 4: Ablation study of LR-GCF on the ACDC dataset under **10% labeled** supervision. We evaluate the contributions of four key components: **PatchRot** denotes the two-view patch-wise rotation augmentation, **RotHead** denotes the auxiliary patch-wise rotation prediction head with cross-entropy supervision, **AttnCons** denotes the attention consistency regularization between the two parallel Swin-based decoders, and **PED** denotes the proposed Positional Encoding Duplicating scheme. **Baseline** uses two independently perturbed views generated by standard noise/augmentations, while disabling PatchRot, RotHead, and AttnCons, and adopting standard absolute positional encoding.

4.4 Comparison with State-of-the-Art Methods

ACDC dataset. We compare LR-GCF on the ACDC dataset with a series of competitive semi-supervised baselines, including UA-MT [Yu *et al.*, 2019], SASSNet [Li *et al.*, 2020], DTC [Luo *et al.*, 2023], and SS-Net [Wu *et al.*, 2022b], as well as recent strong augmentation/self-training based methods such as BCP [Bai *et al.*, 2023] and other competitors listed in Table 1. Following the commonly used protocol, we conduct semi-supervised experiments under two labeled ratios, i.e., 10% and 20%, while treating the remaining scans as unlabeled. For fair comparison, the results of competing methods are taken from their reported numbers under the same experimental settings when available. As shown in Table 1, LR-GCF achieves the highest Dice and Jaccard scores under both labeled settings, while maintaining competitive boundary-distance results in terms of 95HD and ASD. These results indicate that LR-GCF can produce accurate and anatomically plausible segmentations under limited supervision.

ISIC 2018 dataset. We further evaluate LR-GCF on the ISIC 2018 skin lesion segmentation benchmark, comparing it with representative semi-supervised methods listed in Table 2. Following the standard split, we use 10% and 20% labeled images and treat the remaining training images as unlabeled. As shown in Table 2, LR-GCF achieves the highest Dice score under both settings and obtains the best overall results under the 20% labeled setting. Under the 10% setting, it also remains competitive on Jaccard and boundary metrics. These results indicate that LR-GCF generalizes well to dermoscopic image segmentation and effectively leverages unlabeled data under limited supervision.

BraTS 2019 dataset. As shown in Table 3, LR-GCF achieves the highest Dice and Jaccard scores on BraTS 2019 under both 10% and 20% labeled settings, demonstrating strong region-level segmentation performance in the scarce-label regime. Compared with prior semi-supervised baselines, LR-GCF produces more complete and consistent tumor regions, while maintaining competitive boundary distances in terms of 95HD and ASD. The gains become more evident with 20% labeled data, suggesting that LR-GCF can effectively exploit unlabeled scans while benefiting from additional supervision.

4.5 Ablation Study

Table 4 presents the ablation study of LR-GCF's components on the ACDC dataset under 10% labeled supervision. The

top-k	Metrics			
	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
0	89.55	81.36	3.76	1.45
10%	**90.10**	**82.34**	**3.71**	1.38
50%	89.79	81.78	3.73	**1.32**
100%	89.33	81.04	3.8	1.46

Table 5: Ablation on the token selection ratio for attention consistency regularization on the ACDC dataset with 10% labeled data. "top-k" denotes the percentage of query tokens with the largest attention discrepancy selected to compute the MSE-based attention consistency loss. Due to the lower bound $k = \max(1, \lfloor \rho T \rfloor)$, $\rho = 0$ corresponds to selecting only the single most discrepant token.

baseline model shows the weakest performance, while introducing PatchRot improves the results by augmenting the data with patch-wise rotations. Adding the RotHead further improves Dice, Jaccard, 95HD, and ASD, demonstrating the benefit of auxiliary rotation prediction for enhancing high-level representation learning. The introduction of AttnCons brings additional gains, especially in Dice, Jaccard, and ASD, by regularizing attention consistency between the two decoders. Finally, the full LR-GCF model, which combines all components, achieves the best Dice, Jaccard, and 95HD scores, while maintaining competitive ASD. These results demonstrate the effectiveness of jointly combining patch-wise rotations, rotation prediction, attention consistency, and positional encoding for improving segmentation performance under limited supervision.

Table 5 presents the ablation study on the token selection ratio used in attention consistency regularization on the ACDC dataset with 10% labeled data. Here, "top-k" denotes the percentage of query tokens with the largest attention discrepancy selected for the MSE-based consistency loss. Due to the lower bound $k = \max(1, \lfloor \rho T \rfloor)$, $\rho = 0$ corresponds to selecting only the single most discrepant token. As shown, selecting the top 10% discrepant tokens achieves the best Dice, Jaccard, and 95HD scores, while the 50% setting obtains the lowest ASD. Considering the overall balance across metrics, we adopt the top 10% setting in LR-GCF. Selecting too few tokens or using a much larger proportion leads to less effective regularization and slightly degraded overall performance.

5 Visualization

Fig. 3 presents a qualitative comparison on the ACDC dataset against several competitive baselines, such as MT, CPS, DTC, and BCP. LR-GCF consistently yields masks that exhibit superior morphological fidelity and boundary precision. This advantage is particularly evident in challenging slices where other methods produce over-segmented regions or disconnected structures due to low contrast. By contrast, our model preserves the inherent anatomical topology. Fig. 4 further illustrates the impact of different modules on the ISIC dataset. The baseline is prone to fragmented predictions and spurious blobs around weak boundaries. While PatchRot encourages more stable local cues, the addition of RotHead and AttnCons progressively refines the results, suppressing noise and enhancing structural integrity. In aggregate, LR-GCF produces the most coherent segmentations that closely align with the ground truth.

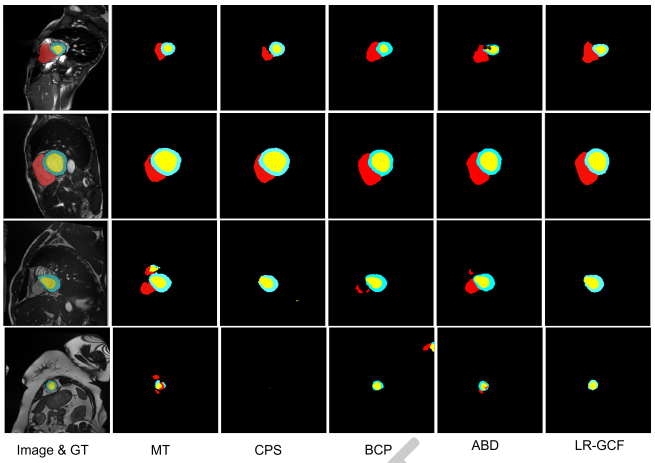

Figure 3: LR-GCF better preserves anatomical structures and reduces fragmented predictions in these examples.

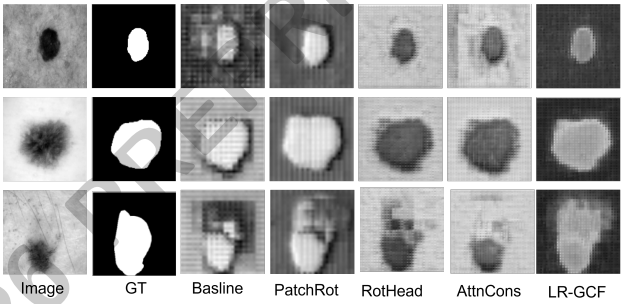

Figure 4: Visualization of different components for LR-GCF on ISIC dataset.

6 Conclusion

This work presents LR-GCF, a robust semi-supervised framework for medical image segmentation that integrates local patch-wise rotations with global consistency regularization. By combining rotation prediction, rotation-aligned attention consistency, and the proposed PED scheme, LR-GCF improves both local feature learning and global anatomical coherence. Extensive experiments on ACDC, ISIC 2018, and BraTS 2019 demonstrate its superiority over strong baselines, especially under limited annotation settings.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62302378 in part by the Innovation and Entrepreneurship Talent Project of Qin Chuang Yuan under Grant QCYRCXM2023-115 and Talent Scientific Fund of Lanzhou University.

References

[Bai *et al.*, 2023] Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation, 2023.

[Cao *et al.*, 2021] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation, 2021.

[Chaitanya *et al.*, 2023] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. *Medical Image Analysis*, 87:102792, 2023.

[Chen *et al.*, 2021a] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation, 2021.

[Chen *et al.*, 2021b] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision, 2021.

[Chen *et al.*, 2023] Cheng Chen, Juzheng Miao, Dufan Wu, Zhiling Yan, Sekeun Kim, Jiang Hu, Aoxiao Zhong, Zhengliang Liu, Lichao Sun, Xiang Li, Tianming Liu, Pheng-Ann Heng, and Quanzheng Li. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation, 2023.

[Chi *et al.*, 2024] Hanyang Chi, Jian Pang, Bingfeng Zhang, and Weifeng Liu. Adaptive bidirectional displacement for semi-supervised medical image segmentation, 2024.

[Gao *et al.*, 2023] Shengbo Gao, Ziji Zhang, Jiechao Ma, Zihao Li, and Shu Zhang. Correlation-aware mutual learning for semi-supervised medical image segmentation, 2023.

[Guan *et al.*, 2024] Xi Guan, Qi Zhu, Liang Sun, Junyong Zhao, Daoqiang Zhang, Peng Wan, and Wei Shao. Global–local consistent semi-supervised segmentation of histopathological image with different perturbations. *Pattern Recognition*, 155:110696, 2024.

[Guo *et al.*, 2025] Xiuzhen Guo, Lianyuan Yu, Ji Shi, Na Lei, and Hongxiao Wang. Diff-cl: A novel cross pseudo-supervision method for semi-supervised medical image segmentation, 2025.

[Han *et al.*, 2024] Kai Han, Victor S Sheng, Yuqing Song, Yi Liu, Chengjian Qiu, Siqi Ma, and Zhe Liu. Deep semi-supervised learning for medical image segmentation: A review. *Expert Systems with Applications*, 245:123052, 2024.

[Hatamizadeh *et al.*, 2022] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, 2022.

[He *et al.*, 2023] Yufan He, Vishwesh Nath, Dong Yang, Yucheng Tang, Andriy Myronenko, and Daguang Xu. Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 416–426, Cham, 2023. Springer Nature Switzerland.

[He *et al.*, 2024] Along He, Tao Li, Juncheng Yan, Kai Wang, and Huazhu Fu. Bilateral supervision network for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(5):1715–1726, 2024.

[Huang *et al.*, 2025a] Kaiwen Huang, Tao Zhou, Huazhu Fu, Yizhe Zhang, Yi Zhou, Chen Gong, and Dong Liang. Learnable prompting sam-induced knowledge distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(5):2295–2306, 2025.

[Huang *et al.*, 2025b] Kaiwen Huang, Tao Zhou, Huazhu Fu, Yizhe Zhang, Yi Zhou, and Xiao-Jun Wu. Uncertainty-aware cross-training for semi-supervised medical image segmentation, 2025.

[Isensee *et al.*, 2023] Fabian Isensee, Constantin Ulrich, Tassilo Wald, and Klaus H Maier-Hein. Extending nnu-net is all you need. In *BVM workshop*, pages 12–17. Springer, 2023.

[Isensee *et al.*, 2024] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 488–498. Springer, 2024.

[Jiao *et al.*, 2024] Rushi Jiao, Yichi Zhang, Le Ding, Bingsen Xue, Jicong Zhang, Rong Cai, and Cheng Jin. Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, 169:107840, 2024.

[Li *et al.*, 2020] Shuailin Li, Chuyu Zhang, and Xuming He. *Shape-Aware Semi-supervised 3D Semantic Segmentation for Medical Images*, page 552–561. Springer International Publishing, 2020.

[Liu *et al.*, 2024] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Yizhou Yu, Yong Liang, Guangming Shi, Shaoting Zhang, Hairong Zheng, and Shanshan Wang. Swin-umamba: Mamba-based unet with imagenet-based pretraining, 2024.

[Liu *et al.*, 2026a] Xingyao Liu, Fan Li, Yueqin Diao, Sichen Tao, and Yujie Yang. Semi-supervised medical image segmentation via region stitching and teacher-assistant collaboration. *Biomedical Signal Processing and Control*, 112:108429, 2026.

[Liu *et al.*, 2026b] Xuejun Liu, Zhaichao Tang, Yonghao Wu, Ruixiang Zhai, Xuanhe Dong, Zikang Du, and Shujun Cao. Mfhs: Mutual consistency learning-based foundation model integrates hypergraph for semi-supervised medical image segmentation. *Pattern Recognition*, 172:112721, 2026.

[Luo *et al.*, 2022] 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.

[Luo *et al.*, 2023] Xiangde Luo, Jieneng Chen, Tao Song, Yinan Chen, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation through dual-task consistency, 2023.

[Olsson *et al.*, 2020] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning, 2020.

[Ruan *et al.*, 2024] Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation, 2024.

[Shamshad *et al.*, 2023] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, 88:102802, 2023.

[Sohn *et al.*, 2020] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence, 2020.

[Su *et al.*, 2024] Jiawei Su, Zhiming Luo, Sheng Lian, Dazhen Lin, and Shaozi Li. Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. *Medical Image Analysis*, 94:103111, 2024.

[Tarvainen and Valpola, 2018] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, 2018.

[Wang *et al.*, 2021] Shanshan Wang, Cheng Li, Rongpin Wang, Zaiyi Liu, Meiyun Wang, Hongna Tan, Yaping Wu, Xinfeng Liu, Hui Sun, Rui Yang, et al. Annotation-efficient deep learning for automatic medical image segmentation. *Nature communications*, 12(1):5915, 2021.

[Wu *et al.*, 2021] Yicheng Wu, Minfeng Xu, Zongyuan Ge, Jianfei Cai, and Lei Zhang. Semi-supervised left atrium segmentation with mutual consistency training, 2021.

[Wu *et al.*, 2022a] Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai. Mutual consistency learning for semi-supervised medical image segmentation, 2022.

[Wu *et al.*, 2022b] Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation, 2022.

[Wu *et al.*, 2026] Han Wu, Chong Wang, and Zhiming Cui. Dual cross-image semantic consistency with self-aware pseudo labeling for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 45(1):177–189, 2026.

[Xie *et al.*, 2021] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation, 2021.

[Xu *et al.*, 2023] Zhe Xu, Yixin Wang, Donghuan Lu, Xiangde Luo, Jiangpeng Yan, Yefeng Zheng, and Raymond Kai yu Tong. Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. *Medical Image Analysis*, 88:102880, 2023.

[Yang *et al.*, 2023] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation, 2023.

[Yu *et al.*, 2019] 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, 2019.

[Zhang *et al.*, 2021] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation, 2021.

[Zhang *et al.*, 2024] Chong Zhang, Lingtong Wang, Guohui Wei, Zhiyong Kong, and Min Qiu. A dual-branch and dual attention transformer and cnn hybrid network for ultrasound image segmentation. *Frontiers in Physiology*, 15:1432987, 2024.

[Zhao *et al.*, 2025] Bin Zhao, Chunshi Wang, and Shuxue Ding. Crossmatch: Enhance semi-supervised medical image segmentation with perturbation strategies and knowledge distillation. *IEEE Journal of Biomedical and Health Informatics*, 29(9):6780–6792, 2025.

[Zhu *et al.*, 2024] Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2, 2024.