This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3717225



# Conditional Flow Matching-based Semi-supervised Segmentation for Quality Assessment of Fundus Images

Runchang Zhou, Zhenquan Wu, Xinyu Zhao, Tiancheng Song, Youyi Song, Teng Zhou, Guoming Zhang and Hai Xie

***Abstract*— Axial misalignment in fundus imaging can severely distort anatomical geometry, frequently reducing the reliability of clinical diagnoses. The orientation of the Disc-Fovea Line (DFL) remains the best method for rotation quality control(QC), but it can be challenging to locate the Optic Disc (OD) and macula, which weakens its utility. The weak textures and few pixel-level labels bring great difficulties for macula annotation, resulting in a scarcity of available annotation data. To address this issue, we introduce FMSSL-GAN, an annotation-efficient Semi-Supervised Semantic Segmentation (SSSS) framework. In particular, a Conditional Flow Matching (CFM) module is added to turn rough, noisy predictions into anatomically correct structures by learning a continuous evolutionary vector field. Our CFM-based approach directly models the probability path, which makes inference much faster while still keeping high-precision structural refinement. This is different from traditional diffusion models, which have slow, multi-step iterative denoising. A Generative Adversarial Network (GAN) is added to impose topological constraints at the distribution level, which enables the spatial relationship between landmarks to remain consistent. For the assessment, our system automatically marks the images as non-compliant when the calculated OD-macula angle is more than $15^{\circ}$. Extensive experiments on a private dataset (i.e., the Szeye) and two public datasets (i.e., REFUGE and ORIGA) demonstrate that the proposed method performs well with a wide range of labeled data ratios and outperforms other compared semi-supervised methods. Our framework provides an annotation-efficient solution for objective fundus image quality assessment, demonstrating a high concordance with expert measurements.**

***Index Terms*— Conditional Flow Matching, GAN, Quality assessment, Semi-supervised segmentation**

## I. Introduction

COLOR fundus photography is an important tool for screening and monitoring eye diseases like diabetic retinopathy, glaucoma, and age-related macular degeneration. Although optical coherence tomography (OCT) provides depth-resolved structural information and has become indispensable for detailed retinal assessment, color fundus photography remains irreplaceable in many screening and follow-up scenarios because of its low cost, wide availability, non-invasive acquisition, large field of view, and compatibility with portable or teleophthalmology systems. More importantly, OCT and fundus photography provide complementary information rather than serving as mutually exclusive modalities: OCT characterizes cross-sectional retinal microstructures, whereas fundus photography offers en face color documentation of retinal anatomy and lesions. Therefore, reliable quality control of fundus images is still necessary, especially before downstream clinical interpretation, longitudinal comparison, and automated AI-based analysis. In this context, rotation and axial misalignment may distort the spatial relationship between the optic disc and macula, leading to unreliable anatomical measurements even when advanced imaging modalities are available. The quality of the images directly determines whether the lesion can be fully observed and how reliable quantitative biomarkers are [1]. Traditional approaches often quantify quality indices such as illumination, naturalness, and structure [2], whereas recent trends shift towards deep learning-based anatomical verification. Camera pose, fixation drift, media opacity, and device heterogeneity often cause axial misalignment and non-standard framing in routine clinics and large-scale screening programs. These problems make anatomical geometry and angle measurements less precise, which makes automated algorithms lack reliability and consistency in clinical interpretations. Hence, establishing standardized quality control (QC) protocols with measurable benchmarks and integrating automated decision-making systems are critical steps to ensure objective and reproducible results.

This work was supported partly by National Natural Science Foundation of China (Nos. 62376164, 62462021, 62406343, 82271103, 82401315, 82301269), Guangdong Basic and Applied Basic Research Foundation (Nos. 2024A1515011676, 2025A1515010197, 2022A1515012326), National Natural Science Foundation of Hainan Province (No. 725MS068 and 625RC716). The Philosophy and Social Sciences Planning Project of Zhejiang Province (No. 25JCXK006YB), the Hainan Province Higher Education Teaching Reform Project (No. HNJG2024ZD-16).(Runchang Zhou, Zhenquan Wu and Xinyu Zhao contributed equally to this work.)(Corresponding authors: Guoming Zhang and Hai Xie)

Runchang Zhou, Tiancheng Song, Teng Zhou, Hai Xie are with the School of Cyberspace Security (School of Cryptography), Hainan University, Haikou, Hainan, China(e-mail: 24220854120037@hainanu.edu.cn;25110839000007@hainanu.edu.cn; teng.zhou@hainau.edu.cn;xiehai@hainau.edu.cn).

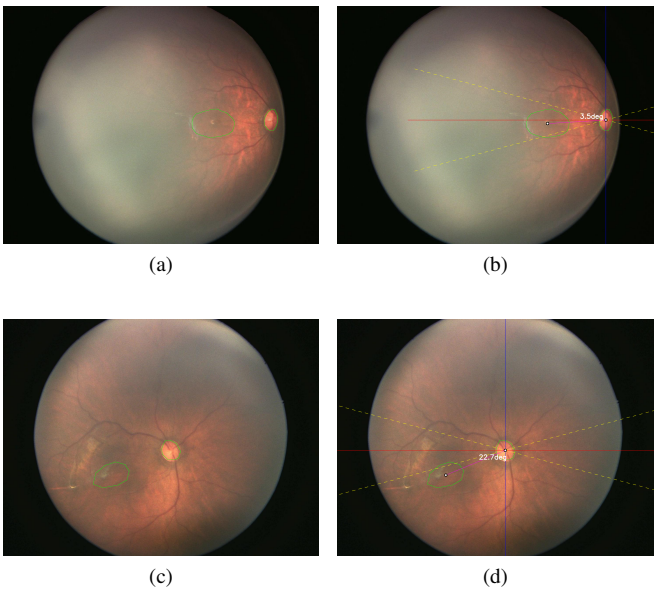
Zhenquan Wu, Xinyu Zhao, Guoming Zhang are now with the Shenzhen Eye Institute, Shenzhen Eye Hospital, Southern Medical University, Shenzhen, China(e-mail: wuzhenquan@sz-eyes.com;zhaoxinyu1@sz-eyes.com; zhangguoming@sz-eyes.com).

Youyi Song is with the School of Science, China Pharmaceutical University, Nanjing, China(e-mail: youyisong@cpu.edu.cn).

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on August 07,2026 at 07:33:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3717225






(a) (b) (c) (d)

Fig. 1. (a) and (c) show the visualization of the ground truth annotations for the optic disc and macula, (b) and (d) display the angle measurements. The red and blue lines represent the horizontal and vertical reference lines, respectively. The purple line denotes the Disc-Fovea Line (DFL) connecting the centroids of the optic disc and macula. The dashed line indicates the $\mathbf{15}^{\circ}$ angular threshold used for quality assessment.

In clinical applications, the optic disc (OD) and macula/fovea are used as key landmarks in fundus imaging for diagnosing diseases. The relative position of the optic disc and fovea, specifically the disc–fovea line orientation, which contains information about camera pose and fixation bias [3], offering a robust metric for assessing image suitability in downstream fundus analysis. Current quality control methods are usually based on human heuristics or fully supervised detection/segmentation. The first one is subjective and hard to repeat, and the second one requires a lot of pixel-level annotation, which is expensive and makes it hard to generalize across devices. Although fully supervised methods have achieved significant success, they are inherently limited by their dependence on comprehensive pixel-level annotations, in which a lot of expert knowledge and hard-effort annotations are required, especially in medical imaging. Therefore, real-world clinical datasets are usually distributed with ”few labels + a lot of unlabeled images”. Consequently, it is imperative to develop learning paradigms capable of leveraging the underlying distribution of unlabeled data to mitigate the dependency on scarce annotations [4].

To address the above challenges, a semi-supervised QC framework centered on anatomical geometry is designed. Given the scarcity of pixel-level annotations, we employ a semi-supervised network to jointly segment the optic disc and the macula under a setting of “few-labeled + many-unlabeled”. The core logic of QC is explicitly geometric, in which we extract the centroids of the segmented structures to compute the disc-fovea line and measure its angle relative to the horizontal axis, as shown in Fig. **1**. The images with an angle exceeding 15° are flagged as non-compliant and excluded from downstream analyses. This segmentation-based design can offer both clinical interpretability (by segmenting specific anatomical structures in fundus images) and engineering practicality (by reducing reliance on massively annotated datasets).

To this end, we propose a novel semi-supervised framework, FMSSL-GAN, which is applied to obtain the robust localization of these landmarks. A Pseudo-Label Refinement (PLR) module based on conditional flow matching is designed to address the issue that the macula pseudo-labels are often noisy or missing. This module rectifies coarse predictions in the latent space, yielding high-confidence pseudo-labels that serve as near-ground-truth supervision. To ensure the disc and macula are anatomically plausible in their relative positioning, we introduce an adversarial discriminator (GAN-style) to enforce structural consistency. In addition, in order to better withstand the very rotational distortions we aim to detect, we added elastic deformation to our data augmentation pipeline, which assists the model in performing better when faced with tricky viewpoint changes. Our contributions are summarized as follows.

- **Semi-supervised framework.** A semi-supervised segmentation framework, called FMSSL-GAN, is designed to effectively extract features from unlabeled data via pseudo-labels.
- **PLR module.** A flow-matching–based PLR module is devised to refine low-quality pseudo-labels into training-ready labels with high confidence. Additionally, we incorporate a heterogeneous timestep strategy within this module.
- **Heterogeneous Timestep Strategy.** We design a Spatially Heterogeneous Timestep (SHT) strategy within the PLR module, which can decouple the learning dynamics of simple and complex regions to obtain accurate boundary delineation over the refinement process.
- **Adversarial supervision.** A GAN discriminator is introduced to regularize the realism and domain robustness of the predicted masks to boost the performance of the baseline model.

## II. RELATED WORK

### A. Fundus Image Quality Control and Disc–Macula Geometry

Automated quality control for fundus images initially relied on low-level indicators—illumination, contrast, sharpness—or heuristic rules based on visible anatomy [5] [6]. However, recent studies emphasized anatomical priors, especially the relative geometry between the optic disc center and the macula center in fundus images. For the reason that the disc–macula center line (DMCL) and its orientation angle, $\theta_{DMCL}$, are highly sensitive to camera pose and fixation drift, making them a key criterion for framing compliance [3]. However, most existing methods either require fully supervised keypoint/segmentation labels or depend on brittle hand-crafted thresholds, which limit scalability to large unlabeled cohorts and across devices. In this work, we adopt the DMCL orientation as the QC basis and make full use of the practical few-

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on August 07,2026 at 07:33:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3717225



label settings via semi-supervised learning to weaken the demand for annotation. By producing high-quality segmentation masks and computing centers, we obtain an interpretable and reproducible geometric metric, which supports a subsequent 15° threshold for automatic pass/fail decisions.

### *B. Semi-Supervised Segmentation in Medical Imaging*

In the real-world medical scene, the "few labels + many unlabeled" regime is ubiquitous. Semi-Supervised Semantic Segmentation (**SSSS**) seeks to exploit the structural distribution of unlabeled data under scarce pixel-level annotations to improve performance and cross-domain generalization [7]. Representative approaches fall into three broad categories: (1) Consistency regularization. The principle is to enforce prediction consistency for the same condition under input perturbations (data augmentation, noise injection, style/domain shifts) or model perturbations (e.g., dropout, parameter noise) [8] [9] [10]. A common practice is weak/strong augmentation pairing, minimizing the discrepancy between the two outputs to stabilize learning—simple to implement and effective for generalization. Beyond standard perturbations, recent works have extended consistency learning to geometric transformations [11]and cross-style domains to maximize the utilization of unlabeled data. (2) Pseudo-labeling / self-training. A current model generates pseudo-labels for unlabeled images. Then these labels supervise iterative training [12] [13] [14]. Despite its effectiveness, this paradigm is vulnerable to error accumulation from noisy pseudo-labels, especially for small or low-contrast structures. Hence, it is often combined with uncertainty estimation, temperature calibration, or cross-model supervision to curb noise propagation [15]. (3) Contrastive learning. According to the "pull similar, push dissimilar" principle, contrastive learning is applied at the pixel-level or region-level to strengthen representations. For example, U2PL [16] leveraged pseudo-label uncertainty to mitigate noise by treating unreliable pixels as negative samples. However, the overwhelming proportion of background pixels and severe class imbalance can dilute the contrastive objective, which limits practical gains in fundus image segmentation. Therefore, with the imbalance of pseudo-label noise and negative-sample imbalance in fundus images, we assume that combining consistency regularization with high-reliability pseudo-labels is more task-aligned and more likely to yield robust improvements than using contrastive learning alone.

### *C. Pseudo-Label Refinement and Generative Priors*

For semi-supervised segmentation, self-training with pseudo-labels is effective, yet its performance is often limited by pseudo-label noise, especially for small, low-contrast, or boundary-ambiguous structures where systematic bias can arise [17]. To mitigate this, prior works have progressed along three main avenues: post-hoc refinement, uncertainty-driven filtering/reweighting, and generative-prior–driven learnable refinement. (1) Post-hoc refinement. Typical techniques include Conditional Random Fields (CRF), morphological operations, boundary sharpening, and connected-component/topology constraints, which reduce noise and jagged edges via local smoothing or handcrafted priors [18]. These methods are simple and lightweight but tend to struggle in weak-texture/boundary-aliasing regions and lack robustness across domains. (2) Uncertainty & confidence control. Using temperature scaling, confidence thresholds, Monte Carlo dropout, or deep ensembles, one can filter low-confidence pixels or down-weight their losses [19]. While this curbs part of the noise accumulation, it often trades off recall for extremely small targets or very thin boundaries, leading to over-conservative behavior. (3) Generative priors & learnable refiners. Diffusion models can learn structured mappings from coarse pseudo-labels to high-quality masks through iterative generation [20]. In recent years, there has been a lot of interest in leveraging generative models for medical image segmentation tasks [21] [22] [23]. For example, [23] [24] demonstrate the feasibility of employing diffusion as a strong structural prior for segmentation. Unlike per-pixel post-processing, a learnable refiner jointly conditions on image features and coarse masks, a learnable refiner jointly conditions on image features and coarse masks, which can explicitly express shape, topology, and boundary consistency, resulting in more stable gains in cross-domain and weak-texture scenarios. Recent studies have also explored semi-supervised learning mode by modeling explicit shape. For instance, Liu et al. [4] presented a shape-aware network to enforce anatomical plausibility alongside consistency regularization. Complementarily, [25] shows the value of adversarial structural priors, where a discriminator is employed to align the predicted-mask distribution with that of ground-truth labels, injecting global shape plausibility and domain robustness.

Rather than relying on fixed post-processing, we instantiate a learnable refiner (PLR) built on a flow-matching generative prior and train it jointly on labeled and unlabeled data to improve pseudo-label quality at the source. Unlike traditional diffusion models that rely on stochastic denoising—which can be computationally intensive and may inadvertently erode the fine-grained structures of small targets—our flow-matching approach offers a deterministic transport path. This ensures more stable boundary preservation for the optic disc and macula while significantly reducing the inference overhead required for iterative refinement. The process of pseudo-label refinement in our framework can be interpreted as a form of feature-level matching between noisy predictions and anatomical priors. This aligns with recent advances in Optimal Transport (OT) based matching, which utilize transport-inspired paths to bridge distribution shifts [26]. Furthermore, our consistency regularization strategy across augmented views is deeply rooted in the principles of feature matching, as discussed in recent comprehensive reviews [27], which emphasize the importance of maintaining stable cross-view representations in challenging clinical imaging scenarios. In parallel, we attach a lightweight discriminator (GAN) to provide distribution-level structural constraints [28] [25]. These two parts are complementary, where PLR enhances pseudo-label usability and boundary/topology consistency and the adversarial supervision is adopted to strengthen global shape priors and cross-domain robustness, jointly reducing noise propagation and boosting

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on August 07,2026 at 07:33:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3717225



the segmentation performance.

## III. Method

FMSSL-GAN (Fig. 2) is a semi-supervised learning architecture that effectively integrates consistency regularization and self-training methods through the combination of three key components. These three components are composed of a base segmentation model that applies consistency regularization, a discriminator that imposes structural constraints, and a flow-matching refinement model for pseudo-label refinement. Fig. 2(c) represents the Pseudo Label Refinement Module, which leverages Flow Matching to rectify segmentation predictions through two parallel pathways during training dynamically:

Weak Augmentation Path: Conditioned on weakly augmented images, this path takes weak pseudo-labels as the initial state. For labeled data, the flow matching module learns the evolutionary vector field from predictions to the Ground Truth. For unlabeled data, the network generates refined weak pseudo-labels via inference, and these refined labels then serve as high-quality targets to compute the unsupervised consistency loss, guiding the update of the main segmentation network.

Strong Augmentation Path: Conditioned on strongly augmented images, the model takes strong pseudo-labels as input. It learns the trajectory mapping from strong pseudo-labels to the weak pseudo-labels, enforcing semantic alignment across different augmentation strategies.

Fig. 3 illustrates the analysis and discrimination mechanism for fundus images. First, the pre-trained Segmentation Diffusion Model (SDM) is employed to segment the optic disc and macula structures accurately. Next, based on the binary segmentation mask $I(x, y)$, the geometric centroid coordinates for each structure are calculated using the center of mass formula:

$$C = \frac{\sum_x \sum_y x \cdot I(x, y)}{\sum_x \sum_y I(x, y)} \tag{1}$$

Subsequently, using the centroid coordinates of the OD and Macula, the rotation angle $\theta_{\text{deg}}$ relative to the horizontal axis is derived as:

$$\theta_{\text{deg}} = \arctan\left(\frac{|y_{OD} - y_{Mac}|}{|x_{OD} - x_{Mac}|}\right) \times \frac{180}{\pi} \tag{2}$$

For the rotation angle $\theta_{deg}$ , $(x_{OD}, y_{OD})$ and $(x_{Mac}, y_{Mac})$ represent the geometric centroid coordinates of the optic disc and macula, respectively.

Finally, the rotation deviation is quantitatively evaluated, where images are classified as unqualified if the deviation exceeds the threshold ($|\theta| > 15°$). This specific $15°$ boundary was established based on the consensus of three senior ophthalmologists. In clinical observation, $15°$ represents the critical point where axial misalignment begins to be visually perceptible. Adopting this strict standard allows the system to act as an early-warning mechanism, maximizing the prompt correction of camera pose. However, it is important to note that our framework calculates the rotation angle quantitatively via continuous centroid coordinates. Therefore, the threshold is not mathematically hardcoded but serves as a tunable hyperparameter. In real-world deployments, this tolerance can be flexibly calibrated (e.g., relaxed to $20°$) to adapt to different imaging protocols, device specificities, and patient populations.

### A. Adversarial Structural Constraint

Pixel-wise losses (e.g., CE/Dice) emphasize per-pixel alignment but provide insufficient constraints on global shape, connectivity, and boundary consistency. This weakness is particularly evident for small objects, weak boundaries, and cross-device/domain shifts, where the predicted masks may become "fragmented," "jagged," or contain spurious small false-positive islands [29] [30]. To mitigate this, we introduce an adversarial structural prior, where a discriminator is trained to model the distribution of real annotation masks, and the segmentor is encouraged to align its outputs with this distribution, thereby complementing the global structural constraints that the pixel-wise losses fail to capture.

We adopt a lightweight structural discriminator to model the distribution of real annotation masks. The discriminator $D$ is a lightweight CNN consisting of four $4 \times 4$ convolutional layers with a stride of 2. The feature channels progress through $\{64, 128, 256, 512\}$, each followed by Batch Normalization and LeakyReLU ($\alpha = 0.2$). Finally, an adaptive average pooling layer and a linear layer with Sigmoid activation are employed to output a global realism score. It operates purely in the mask space without direct access to the original image. Hence, it can focus on learning global shape and topological priors.

Given the inherent spatial consistency of fundus landmarks (i.e., the stable relative positioning of the optic disc and macula ), a lightweight discriminator is sufficient to capture global topological and geometric priors. This design avoids the risk of overfitting to local noise while effectively complementing the CFM-based refiner, which handles fine-grained structural details.

The hinge adversarial loss (a BCE variant can be used interchangeably with similar conclusions) is employed to restrain the output. Specifically, let $D(\cdot)$ denote the discriminator, $p_\theta(\cdot)$ the probability masks predicted by the segmentor, and y the ground-truth with one-hot mask. The objectives for the discriminator and the generator are, respectively:

Discriminator loss (to distinguish real vs. predicted mask distributions):

$$\mathcal{L}(D) = \mathbb{E}_y\left[\max(0, 1 - D(y))\right] + \mathbb{E}_z\left[\max(0, 1 + D(G(z)))\right] \tag{3}$$

Generator adversarial loss :

$$\mathcal{L}_{\text{adv}}^{(G)} = -\,\mathbb{E}_x\left[D(p_\theta(x))\right] \tag{4}$$

In these objectives, $\mathbb{E}$ denotes the expectation operator. $y$ is the one-hot encoded ground-truth mask, while $z$ represents the latent variable or the initial prediction state sampled from the distribution $p_z(z)$. $p_\theta(x)$ denotes the probability map predicted by the segmentor for a given input image $x$.

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on August 07,2026 at 07:33:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3717225



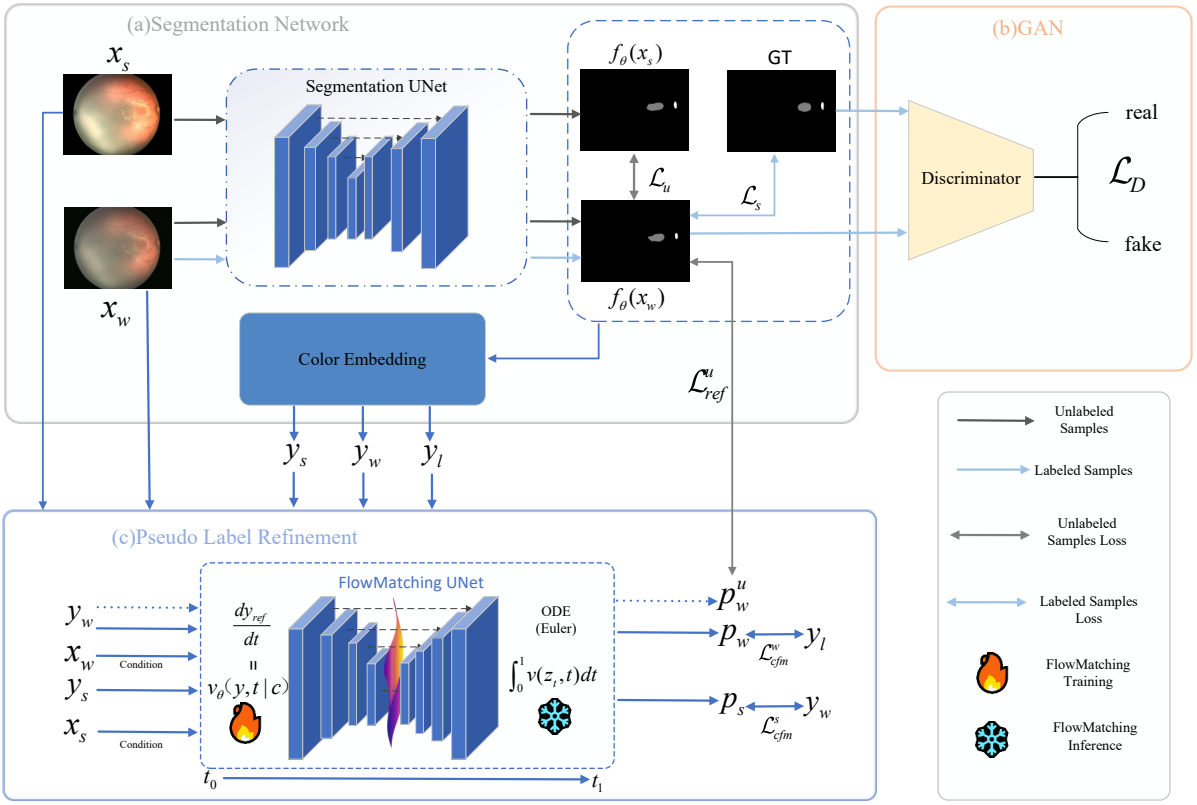



Fig. 2. The architecture diagram of the FMSSL-GAN model.Fig (c): Pseudo Label Refinement Module. This module dynamically rectifies predictions via Flow Matching. The Weak Augmentation Path, processing prediction $\boldsymbol{y_w}$ conditioned on $\boldsymbol{x_w}$, learns supervised velocity fields during Training (Flame) and generates rectified pseudo-labels $\boldsymbol{p_w^u}$ via ODE during Inference (Snowflake). The Strong Augmentation Path, processing prediction $\boldsymbol{y_s}$ conditioned on $\boldsymbol{x_s}$, learns the trajectory towards these targets $\boldsymbol{p_w}$ to enforce semantic consistency.

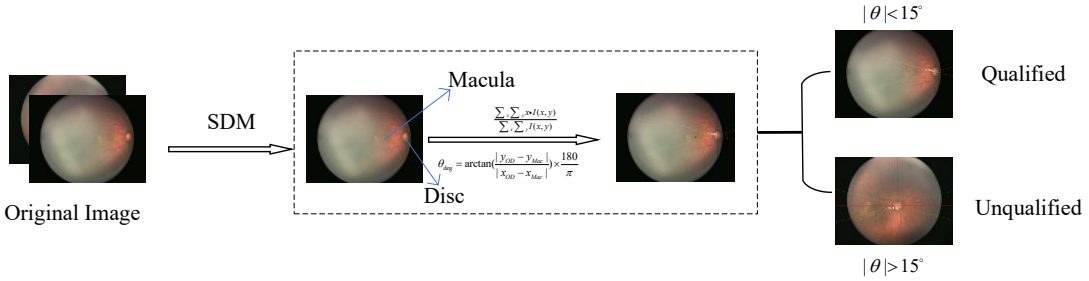



Fig. 3. Schematic of rotation angle calculation and quality screening. The post-processing module calculates the geometric centroids of the segmented Optic Disc and Macula. The inclination angle $\boldsymbol{\theta}$ is derived using the arctangent function. Images are classified as unqualified based on the criterion $|\boldsymbol{\theta}| > \mathbf{15}°$ to ensure geometric standardization.

### *B. Pseudo-Label Refinement via Conditional Flow Matching*

To address the distribution misalignment issue of pseudo labels in semi-supervised learning, we design a Pseudo Label Refinement Module. The original work DiffRect [20] formulates its theoretical framework on latent diffusion models (Latent DDPM) and learns the transformation from coarse to refined pseudo-label distributions by simulating a denoising process. In the implementation, we adopt Conditional Flow Matching (CFM) [31] as a more direct and efficient generative modeling paradigm, while also incorporating the color embedding strategy from DiffRect [20]. CFM learns an ordinary differential equation (ODE) that explicitly encodes the continuous transition path between distributions, thereby avoiding the time-consuming iterative noise-injection procedure in DDPMs and enabling more efficient training and inference.

*1) Conditional Flow Matching (CFM) Framework :* Our goal is to learn a *conditional vector field* $\nu_\theta$ that defines a continuous flow from a source latent feature distribution $p_0(y)$ (e.g., noisy pseudo labels) to a target latent feature distribution $p_1(y)$ (e.g., latent representations corresponding to clean ground-truth labels). Formally, we consider two latent variables

$$y_0 \sim p_0(y), \qquad y_1 \sim p_1(y), \tag{5}$$

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on August 07,2026 at 07:33:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3717225



where $y_0$ in the latent space represents an initial pseudo label that may contain structural or boundary errors, while $y_1$ represents the expected latent code aligned with the ground-truth annotation.

To connect these two distributions, we define a continuous probability path from time $t = 0$ to $t = 1$. In this work, we adopt a simple **linear interpolation** path:

$$y_t = (1 - t)\, y_0 + t\, y_1, \qquad t \in [0, 1]. \tag{6}$$

The above interpolation induces a deterministic target vector field along the path. Differentiating Eq. (6) with respect to time $t$ yields the ground-truth velocity:

$$\frac{dy_t}{dt} = y_1 - y_0. \tag{7}$$

This velocity is constant along the path and describes how a sample should be transported from the source latent distribution toward the target latent distribution.

We therefore train a refinement network (i.e., the flow predictor) to directly regress this target vector field. Conditioned on auxiliary information $c$ (e.g., image features, task context, or the current segmentation prediction), the network predicts

$$\nu_\theta(y_t, t \mid c) \approx y_1 - y_0. \tag{8}$$

The learning objective is a mean-squared error (MSE) between the predicted conditional flow and the target flow, which is sampled at random time steps $t \sim \mathcal{U}(0, 1)$:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{y_0 \sim p_0,\, y_1 \sim p_1,\, t \sim \mathcal{U}(0,1)} \left[ \| \nu_\theta(y_t, t \mid c) - (y_1 - y_0) \|_2^2 \right], \tag{9}$$

where $y_t$ is computed by Eq. (6). In practice, $\nu_\theta(\cdot)$ is instantiated with a U-Net style architecture that serves as a vector-field predictor, and time-step embeddings are injected to make the model aware of continuous time $t$, and the current latent state $y_t$, the condition $c$. Besides, the time encoding is fused before being passed through the encoder–decoder to produce the predicted flow.

This conditional CFM formulation can *directly* encode the continuous transport from coarse/noisy pseudo labels to high-fidelity label-like latent features, avoiding the iterative noising–denoising procedure of diffusion-based refiners while retaining strong structural priors for segmentation.

*2) Spatially Heterogeneous Timestep Strategy :* In conventional CFM frameworks, the time step $t$ is typically sampled as a global scalar ($t \in \mathbb{R}$) for the entire input sample, implying that the transformation from the source distribution (noise/pseudo-label) to the target distribution (ground truth) progresses synchronously across all spatial locations. However, the complexity of the transformation varies significantly across the spatial domain for medical image segmentation. The "easy" uniform regions (e.g., the center of the optic disc or the background) may require minimal flow integration to converge, whereas the "hard" semantic boundaries require complex non-linear trajectories that are sensitive to specific time intervals in the flow.

To address this spatial imbalance, we introduce a Spatially Heterogeneous Timestep (SHT) mechanism. Instead of broadcasting a single scalar $t$ to the entire feature map, we sample a spatial time map $\mathbf{T} \in [0, 1]^{H \times W}$. This allows different pixels or local patches within the same image to undergo flow matching at independent temporal stages simultaneously. Formally, during training, the time map is sampled such that:

$$\mathbf{T}_{h,w} \sim \mathcal{U}(0, 1), \quad \forall (h, w) \in \Omega, \tag{10}$$

where $\Omega$ denotes the spatial domain. The interpolated state $y_{\mathbf{T}}$ is constructed using this spatially varying $\mathbf{T}$ via element-wise operations:

$$\mathcal{L}_{\text{SHT}} = \mathbb{E}_{\mathbf{T}, y_0, y_1} \left[ \frac{1}{|\Omega|} \| \nu_\theta(y_{\mathbf{T}}, \mathbf{T} | c) - (y_1 - y_0) \|_F^2 \right] \tag{11}$$

where $\odot$ denotes the Hadamard product. The flow matching objective is then computed over this pixel-wise time field:

$$y_{\mathbf{T}} = (1 - \mathbf{T}) \odot y_0 + \mathbf{T} \odot y_1, \tag{12}$$

By employing this heterogeneous timestep strategy, the network is forced to learn the vector field conditional on local temporal contexts rather than a global average. This effectively decouples the learning dynamics of simple and complex regions, preventing the dominant gradients from easy regions by suppressing the subtle flow adjustments required at the boundaries. Consequently, this process enables finer control over the refinement process, resulting in more accurate boundary delineation in the semi-supervised setting.

*3) Learning Two Key Distribution Transitions:* As mentioned, two complementary distribution transition paths need to learn to refine pseudo-labels and enforce consistency.

For labeled data, we can use them to train the refiner to correct initial predictions into accurate ground truth. We map the weak augmentation prediction$y_{rgb}^{\text{weak}}$(source) to the ground truth label map$y_{rgb}^{\text{gt}}$(target). The loss can be defined as:

$$\mathcal{L}_{\text{cfm}}^{\text{sup}} = \mathbb{E}_{\mathcal{D}} \left[ \frac{1}{|\Omega|} \left\| \nu_\theta(y_{\mathbf{T}}^{\text{W2G}}, \mathbf{T} | x^{\text{weak}}) - (y_{rgb}^{\text{gt}} - y_{rgb}^{\text{weak}}) \right\|^2 \right] \tag{13}$$

where $y_{\mathbf{T}}^{\text{W2G}}$ represents the interpolated state at time map $\mathbf{T}$, and $x^{\text{weak}}$ indicates the original images. The expectation is taken over the data distribution $\mathcal{D} = \{\mathbf{T}, y_{rgb}^{\text{weak}}, y_{rgb}^{\text{gt}}\}$.

For unlabeled data, we address the inconsistency between strong and weak augmentations. Specifically, we directly employ the prediction from the weakly augmented view as the target. The model then learns the trajectory from the noisy strong prediction $y_{rgb}^{strong}$ to this weak target:

$$\mathcal{L}_{\text{cfm}}^{\text{unsup}} = \mathbb{E}_{\mathcal{D}} \left[ \frac{1}{|\Omega|} \left\| \nu_\theta(y_{\mathbf{T}}^{\text{S2W}}, \mathbf{T} | x^{\text{strong}}) - (y_{rgb}^{\text{weak}} - y_{rgb}^{\text{strong}}) \right\|^2 \right] \tag{14}$$

This 'Strong-to-Weak' formulation ensures that the consistency loss is computed against the prediction from the weakly augmented view, serving as a stable target to enforce semantic consistency in the semi-supervised feedback loop. Wherein $y_{\mathbf{T}}^{\text{S2W}}$ represents the interpolated state at time map $\mathbf{T}$, and $x^{\text{strong}}$ is the input images. The expectation is taken over the data distribution $\mathcal{D} = \{\mathbf{T}, y_{rgb}^{\text{strong}}, y_{rgb}^{\text{weak}}\}$.

Finally, the two losses are combined to build the overall latent-flow objective loss:

$$\mathcal{L}_{\text{CFM}} = \mathcal{L}_{\text{cfm}}^{\text{sup}} + \lambda_{\text{unsup}}\, \mathcal{L}_{\text{cfm}}^{\text{unsup}}, \tag{15}$$

where $\lambda_{\text{unsup}}$ balances the contribution from unlabeled data.

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on August 07,2026 at 07:33:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3717225



*4) Pseudo-Label Refinement via ODE Solving:* During training, we leverage the learned W2G (and S2W) transport capability to *dynamically* refine pseudo labels. This refinement is realized by solving an ordinary differential equation (ODE) defined by the learned conditional vector field $\nu_\theta$.

Given an initial weakly pseudo label in latent space $y_w$ (at $t = 0$), we define the following ODE:

$$\frac{dy_w}{dt} = \nu_\theta(y_w, t \mid c), \qquad t \in [0, 1], \tag{16}$$

where $c$ is the condition (Weakly augmented image). By integrating the ODE from $t = 0$ to $t = 1$ with a numerical ODE solver, we obtain the refined latent features:

$$y_w^{ref} = y_w + \int_0^1 \nu_\theta(y_t, t \mid c)\, dt. \tag{17}$$

In practice, this integral is approximated by a finite-step solver:

$$y_w^{ref} \approx \text{ODESolve}\big(y_w, \nu_\theta, t = 0 \to 1\big), \tag{18}$$

While higher-order solvers like RK4 can be employed, we adopt the Euler method with a fixed number of steps $N = 6$ (corresponding to a step size of $\Delta t \approx 0.167$) to prioritize computational efficiency. Given that our linear interpolation path induces a constant ground-truth velocity field, this first-order approximation provides a sufficient balance between numerical accuracy and inference speed for pseudo-label refinement. This process deterministically "pushes" an arbitrary pseudo label toward its more accurate and label-like counterpart along the learned optimal path. The refined pseudo label $\tilde{y}$ is then decoded and reused as a *more reliable supervision signal* to compute the consistency loss using unlabeled data for the main segmentation network. Specifically, the loss $\mathcal{L}^u_{ref}$ is defined as follows:

$$\mathcal{L}^u_{\text{ref}} = \text{CE}(y_w^{ref}, y_w) + \text{Dice}(y_w^{ref}, y_w) \tag{19}$$

$ODESolve(\cdot)$ denotes a numerical ODE solver, such as the Euler method, used to deterministically transport predictions toward the target manifold. $CE$ and $Dice$ refer to the Cross-Entropy loss and Dice similarity coefficient loss, respectively.

By the way, we can effectively mitigate the confirmation bias induced by noisy pseudo-labels.

In summary, the training procedure for the PLR module follows three main steps: 1. Data Preparation: Generate weak and strong augmented views $(x_w, x_s)$ and obtain initial latent predictions $(y_w, y_s)$ from the base segmentor. 2. Vector Field Regression: Sample a spatially heterogeneous time map $T$ to learn the linear interpolation path between distributions. 3. Pseudo-Label Refinement: Solve the learned ODE during inference to 'push' noisy predictions toward the ground-truth manifold , which then provides high-confidence supervision for the consistency loop.

### C. Loss Function

In the proposed semi-supervised framework, the training objective comprises four components: a supervised segmentation loss, a consistency-driven unsupervised loss, an adversarial structural constraint, and a latent–flow–based refinement loss.

**Algorithm 1** FMSSL-GAN Training Procedure

**Require:** Labeled data $\mathcal{D}_L$, Unlabeled data $\mathcal{D}_U$, Max iterations $N$.
**Ensure:** Segmentor $f_\theta$, CFM Refiner $v_\phi$, Discriminator $D$.

```
for iter = 1 to N do
    // Step 1: Forward Propagation
    Obtain weak/strong predictions y_w, y_s via f_θ
    // Step 2: CFM Learning (Section III-B.3)
    Sample time map T ∈ [0,1]^{H×W} Eq. (10)
    Compute L_cfm^sup Eq. (13) and L_cfm^unsup Eq. (14)
    Calculate total L_CFM Eq. (15)
    // Step 3: Pseudo-Label Refinement (Section III-B.4)
    Solve ODE to get y_w^ref Eq. (18)
    // Step 4: Optimization (Section III-C)
    Calculate L_main Eq. (20)
    Update θ, φ, ψ by minimizing total loss L Eq. (21)
end for
```

This design follows the actual implementation, which can be defined as:

$$\mathcal{L}_{\text{main}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{con}}\, \mathcal{L}_{\text{unsup}} + \mathcal{L}^G_{\text{adv}} + \mathcal{L}^u_{\text{ref}}, \tag{20}$$

The final training loss is

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \mathcal{L}_{\text{CFM}}. \tag{21}$$

TABLE I
DATASET SPLIT AND LABELED DATA DISTRIBUTION FOR SEMI-SUPERVISED LEARNING

| Dataset | Train | Val | Test | Labeled Training Samples | | |
|---|---|---|---|---|---|---|
| | | | | 10% | 20% | 50% |
| Szeye | 548 | 78 | 155 | 56 | 112 | 280 |
| REFUGE | 840 | 120 | 240 | 84 | 168 | 420 |
| ORIGA | 455 | 65 | 130 | 46 | 92 | 230 |

## IV. EXPERIMENTS AND RESULTS

### A. Experimental setup

*1) Datasets:* We evaluate the proposed method on one private dataset and two publicly available benchmarks. Szeye dataset: this dataset was collected by Shenzhen Eye Hospital and contains 781 fundus images with manual annotations of the optic disc and the macula. For the Szeye dataset, manual annotations were provided by three senior ophthalmologists with over 5 years of clinical experience.All images were collected with the explicit consent of the participants. REFUGE dataset [32]: this dataset originates from the MICCAI 2018 Retinal Fundus Glaucoma Challenge and comprises 1,200 fundus images with annotations of the optic disc and optic cup. ORIGA dataset [33]: the ORIGA-650 dataset is provided by the Singapore National Eye Centre and consists of 650 color fundus images, each with segmentation masks of the optic disc and optic cup.

As detailed in Table I, all three datasets are randomly partitioned into training, validation, and test sets following a

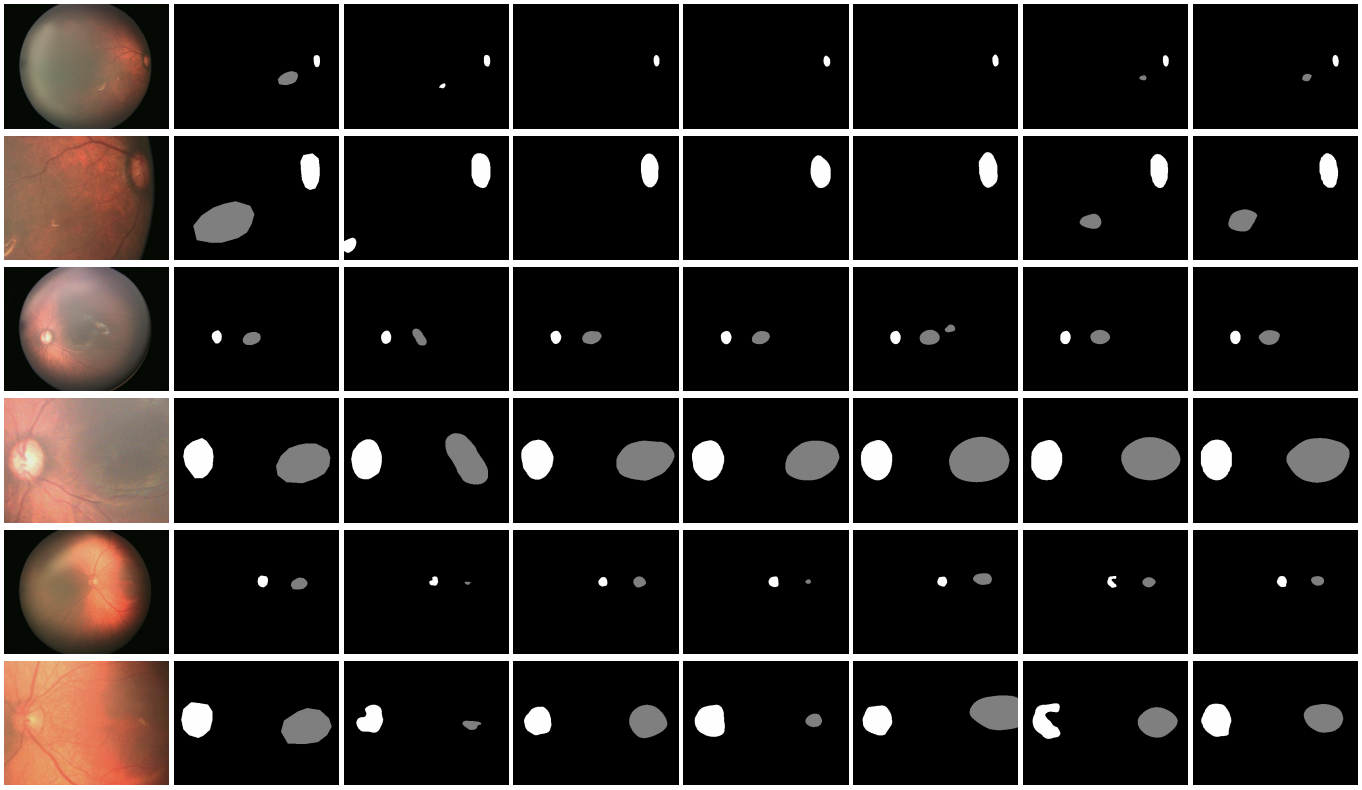
Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on August 07,2026 at 07:33:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3717225



TABLE II
COMPARISON OF DIFFERENT SEMI-SUPERVISED METHODS UNDER 10%, 20%, AND 50% LABELED RATIOS. THE EVALUATION METRIC IS THE DICE SCORE.

| Method | Labeled ratio | Szeye | | | REFUGE | | | ORIGA | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | OD | MA | Avg | OD | OC | Avg | OD | OC | Avg |
| FixMatch | 10% | 90.14 | 57.90 | 74.02 | 87.36 | 76.98 | 82.17 | 84.33 | 77.20 | 80.76 |
| UA-MT | | 87.39 | 56.27 | 71.82 | 93.07 | 85.85 | 89.46 | 94.29 | 81.54 | 87.91 |
| SLC-Net | | 89.27 | 68.23 | 78.75 | 93.80 | 85.77 | 89.78 | 94.30 | 84.00 | 89.15 |
| URPC | | 87.05 | 56.04 | 71.54 | 93.13 | 85.29 | 89.21 | 93.66 | 82.84 | 88.25 |
| DiffRect | | 90.45 | 71.84 | 81.14 | 94.52 | 86.30 | 90.41 | 95.23 | 84.89 | 90.06 |
| Ours | | **90.77** | **73.68** | **82.22** | **95.18** | **87.34** | **91.26** | **95.62** | **85.96** | **90.79** |
| FixMatch | 20% | 89.91 | 65.40 | 77.66 | 89.47 | 86.04 | 87.76 | 87.82 | 82.88 | 85.10 |
| UA-MT | | 90.97 | 71.31 | 81.14 | 93.97 | 86.13 | 90.05 | 95.20 | 83.24 | 89.22 |
| SLC-Net | | 89.79 | 70.69 | 80.24 | 92.71 | 85.88 | 89.80 | 94.94 | 83.78 | 89.36 |
| URPC | | 89.50 | 67.91 | 78.71 | 93.93 | 84.92 | 89.42 | 94.44 | 82.34 | 88.39 |
| DiffRect | | 90.65 | 75.69 | 83.17 | 95.00 | 86.92 | 90.96 | 94.81 | 84.92 | 89.86 |
| Ours | | **92.23** | **77.44** | **84.84** | **95.40** | **87.94** | **91.67** | **96.02** | **86.34** | **91.18** |
| FixMatch | 50% | 91.36 | 65.26 | 78.31 | 94.88 | 85.14 | 90.01 | 92.98 | 80.95 | 86.96 |
| UA-MT | | 90.88 | 70.88 | 80.88 | 94.27 | 86.27 | 90.27 | 93.39 | 83.12 | 88.28 |
| SLC-Net | | 90.84 | 68.49 | 79.66 | 95.50 | 87.18 | 91.34 | 94.92 | 83.93 | 89.43 |
| URPC | | 90.20 | 68.25 | 79.23 | 94.28 | 87.27 | 90.78 | 94.29 | 84.22 | 89.26 |
| DiffRect | | 92.07 | 77.15 | 84.61 | 95.49 | 88.47 | 91.98 | 95.28 | 85.69 | 90.49 |
| Ours | | **92.52** | **77.89** | **85.21** | **95.70** | **88.83** | **92.26** | **95.72** | **87.68** | **91.70** |



(a) Original (b) True label (c) FixMatch (d) UA-MT (e) URPC (f) SLC-Net (g) DiffRect (h) Ours

Fig. 4. Comparison of segmentation results. (a) original image, (b) ground truth, (c)-(g) compared methods, (h) proposed FMSSL-GAN.

7:1:2 ratio. For the semi-supervised experiments, we further divide the training data into labeled and unlabeled subsets based on labeled ratios of 10%, 20%, and 50%. The specific number of labeled samples corresponding to these ratios for each dataset is also summarized in this table. While the Szeye dataset is used for the full DFL-based QC task , REFUGE and ORIGA serve exclusively as benchmarks for generalized segmentation performance on the optic disc and cup, as they lack macula annotations.

*2) Evaluation metric:* We adopt the UNet as the backbone network, and the Dice coefficient is employed to assess the segmentation performance quantitatively.

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on August 07,2026 at 07:33:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3717225



*3) Implementation details:* We conducted the experiments following the method described in [34]. Our training process follows a dual-path semi-supervised framework that integrates the base segmentor with the proposed CFM-based refinement module . For each iteration, labeled and unlabeled images are sampled concurrently, resized to $256 \times 256$ pixels, and normalized to the range of $[0, 1]$. We apply weak perturbations, including random flipping and rotation, to generate initial predictions; these are subsequently processed by the PLR module to produce high-confidence pseudo-labels through ODE solving. These refined signals serve as targets to supervise the network's predictions for strongly augmented views—subjected to Gaussian blur, elastic deformations, and photometric adjustments such as contrast, sharpness, and brightness variations—thereby enforcing cross-augmentation semantic consistency. The model is optimized using stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 0.0001. The initial learning rate is set to 0.01.

## B. Comparison with State-of-the-art Methods

We compare our proposed method against five representative approaches: FixMatch [8], UA-MT [19], URPC [35], SLC-Net [4], and DiffRect [20]. From Table II, we can see that our method consistently achieves superior performance under all labeled ratios, which verifies the effectiveness of the proposed approach across all three datasets. Compared to the state-of-the-art methods, our method yields significant performance gains on the test set. FMSSL-GAN consistently outperforms the strong generative baseline, DiffRect, across all labeled ratios. Notably, on the Szeye dataset, our method achieves significant performance gains, surpassing the runner-up by up to 1.72% (at the 20% ratio) and 1.08% in the extremely low-data regime (10% ratio). For the REFUGE and ORIGA datasets, FMSSL-GAN maintains a steady lead with consistent but more incremental improvements, further validating the robustness of our flow-based refinement strategy across diverse clinical scenarios. These results strongly indicate that even when labeled samples are extremely scarce, the model effectively maintains robust segmentation performance by accurately modeling pseudo-label distribution propagation and generating refined masks.

To visually evaluate the segmentation performance, Fig. 4 illustrates the segmentation results of five compared models on representative samples from the test set. As observed in Fig. 4, our proposed model can achieve more precise boundary delineation for the optic disc and macula than the compared methods, resulting in structurally more complete segmentation predictions.

## C. Ablation Study

To verify the effectiveness of each module in the proposed method, we conduct a group of ablation studies using 20% labeled data. The results are reported in Table III. Whereas FM denotes the Pseudo Label Refinement module, GAN represents the discriminator, and Tmap refers to the heterogeneous time steps within the refinement module. As we can see that the incorporation of the FM module yields a Dice score of 84.08%, representing a 6.42% improvement over the baseline. Further introducing the GAN discriminator and heterogeneous time steps brings an additional 0.76% gain in Dice. These results demonstrate that each module is necessary and contributes positively to the overall segmentation performance.

TABLE III
ABLATION STUDY ON THE SZEYE DATASET USING 20% LABELED DATA

| Components | | | | Szeye (Dice score, %) | | |
|---|---|---|---|---|---|---|
| Base | FM | Tmap | GAN | OD | Macula | Avg |
| ✓ | × | × | × | 89.91 | 65.40 | 77.66 |
| ✓ | ✓ | × | × | 91.99 | 76.17 | 84.08 |
| ✓ | ✓ | ✓ | × | 92.13 | 76.72 | 84.43 |
| ✓ | ✓ | ✓ | ✓ | 92.23 | 77.44 | 84.84 |

TABLE IV
STATISTICAL SIGNIFICANCE ANALYSIS OF THE SHT MECHANISM (TMAP) OVER THREE INDEPENDENT RUNS ON THE SZEYE DATASET

| Method | Run 1 | Run 2 | Run 3 | Mean ± SD (%) |
|---|---|---|---|---|
| Base + FM | 84.08 | 83.88 | 83.85 | $83.94 \pm 0.13$ |
| Base + FM + Tmap | 84.36 | 84.67 | 84.43 | $\mathbf{84.49 \pm 0.16}$ |

The improvement of $+0.55\%$ was validated using a paired t-test ($p = 0.065$). Results demonstrate a consistent performance gain across all independent runs.

To further confirm the SHT mechanism's reproducibility and statistical significance, we performed a specific analysis across three separate experimental runs. The model with Tmap consistently got higher Szeye Dice scores than the one without it, as shown in Table IV (marked in red to show the change). Specifically, the mean performance improved from $83.94 \pm 0.13\%$ to $84.49 \pm 0.16\%$. A paired t-test confirms a strong positive trend ($p = 0.065$). Given the small sample size, this consistent superiority across all trials demonstrates that the SHT module provides a reliable and stable refinement for the anatomically challenging macula region, rather than fluctuating with random weight initialization.

TABLE V
T-TEST RESULTS COMPARISON

| PreDeg | TrueDeg | DiffDeg | T | P |
|---|---|---|---|---|
| $8.12 \pm 6.08$ | $8.26 \pm 6.08$ | $-0.14 \pm 0.69$ | $-0.20$ | 0.84 |

PreDeg denotes the predicted rotation angle; TrueDeg denotes the true rotation angle; DiffDeg denotes the difference between predicted and true values; T and P denote the T-value and P-value, respectively.

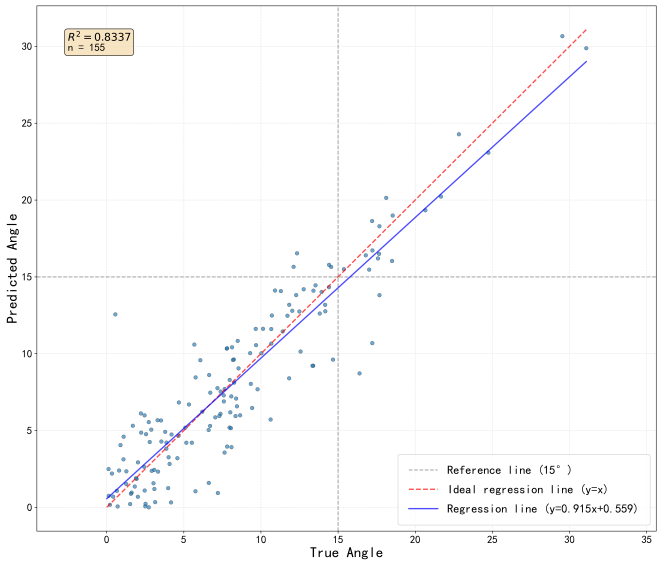
## D. Measurement results of the ocular torsion angle

Following the segmentation of the optic disc and macula using 20% labeled data, the ocular torsion angle is measured by computing the angle of the line connecting their centroids. Fig. 6 provides a qualitative visualization of the angle measurement. Fig. 6(a) and Fig. 6(c) display the ground truth angles, and Fig. 6(b) and Fig. 6(d) show the corresponding predicted angles. The figure includes examples of both compliant and non-compliant images, which is utilized to demonstrate minimal discrepancy between the true and predicted values.

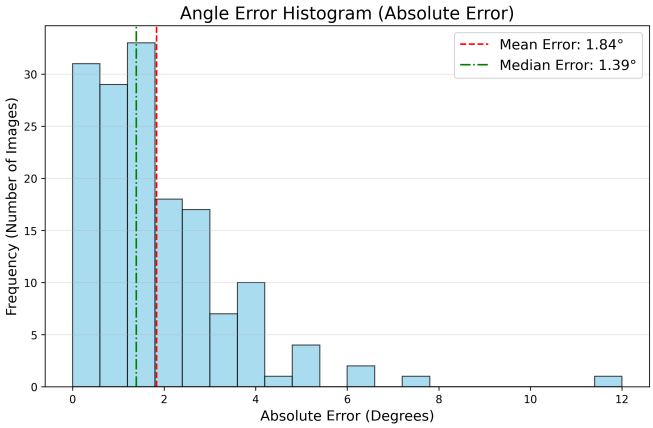
Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on August 07,2026 at 07:33:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3717225

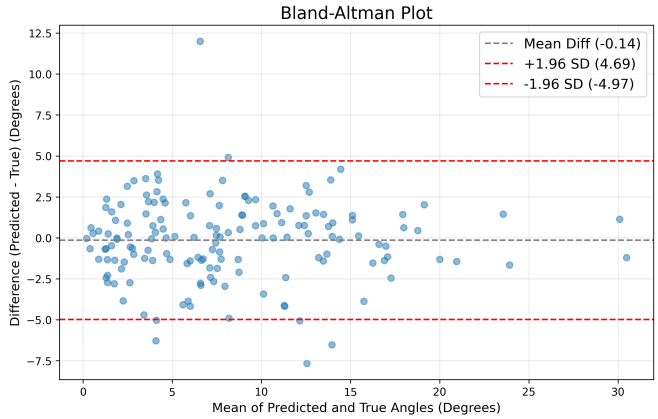




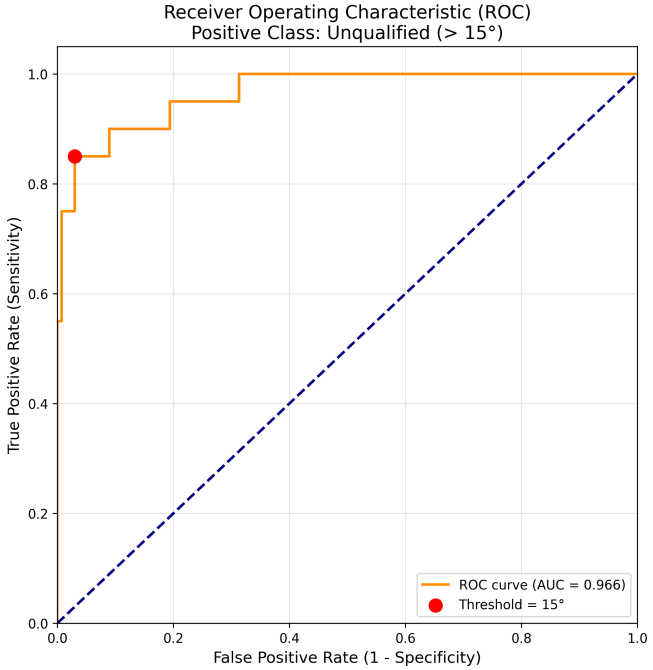

(a) Linear Regression Analysis of Predicted vs. True Angles




(b) Histogram of Absolute Angular Prediction Errors




(c) Bland-Altman Analysis of Prediction Agreement




(d) ROC Curve for Rotation Quality Classification

Fig. 5. Quantitative evaluation results of the proposed method in terms of rotation angle estimation and quality control. (a) Linear regression analysis between predicted and true angles, showing a high correlation ($\boldsymbol{R^2 = 0.8337}$) with the ideal line ($\boldsymbol{y = x}$). (b) Histogram of absolute angular prediction errors with a mean error of $\mathbf{1.84}°$ and median error of $\mathbf{1.39}°$. (c) Bland-Altman plot for assessing the agreement between predicted and true values; most points fall within the 95% limits of agreement ($\boldsymbol{\pm 1.96 SD}$). (d) Receiver Operating Characteristic (ROC) curve for the quality classification task (threshold $\boldsymbol{> 15}°$), achieving an AUC of 0.966.

To comprehensively evaluate quantitative performance, we conduct experiments on a test subset of 155 samples (Fig. 5). First, the linear regression analysis in (Fig. 5(a)) illustrates a high degree of concordance, where the scatter points are tightly clustered around the ideal regression line ($y = x$). The model achieved a coefficient of determination ($R^2$) of 0.8337, indicating a strong correlation with expert annotations.

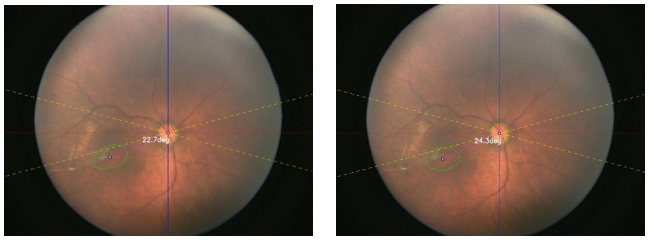
Second, the error distribution is presented in detail by the histogram in (Fig. 5(b)). The errors are predominantly concentrated in the low-error range ($0°$–$2°$), demonstrating high precision. Quantitatively, the method yielded a Mean Absolute Error (MAE) of $1.84°$ and a median error of $1.39°$.

Furthermore, we employ a Bland-Altman analysis (Fig. 5(c)) to assess the agreement indicator between the predicted and true values. The plot reveals a negligible mean difference (bias) of $0.14°$, suggesting the absence of significant systematic error. The majority of sample points fall within the 95% limits of agreement ($\pm 1.96$ SD), further confirming

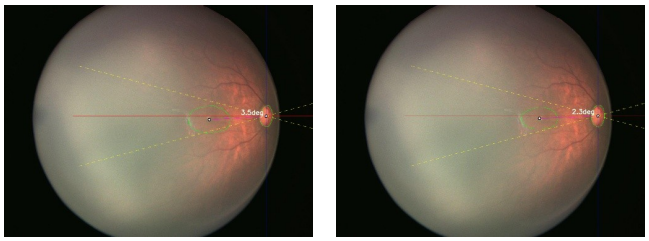
Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on August 07,2026 at 07:33:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3717225





(a) True Sample 1 (b) Test Sample 1



(c) True Sample 2 (d) Test Sample 2

Fig. 6. Examples of eyeball rotation angle measurement. (a)-(b) show a Qualified case (compliant with the $\mathbf{15}^{\circ}$ threshold); (c)-(d) show an Unqualified case (non-compliant but accurately measured).

the reliability of the proposed method.

Finally, regarding the $15^{\circ}$ clinical threshold for quality control, the system demonstrates robust classification capabilities. As illustrated in the Receiver Operating Characteristic (ROC) curve (Fig. 5(d)), the proposed method achieves an Area Under the Curve (AUC) of 0.966. To further validate the consistency of the measurements, we also perform a T-test. The statistical analysis is reported in Table V, which shows that the P-value exceeds 0.05, indicating no statistically significant difference between the results of the proposed method and the expert measurements.

### *E. Discussion on Mathematical Interpretability*

The superior performance of FMSSL-GAN in fundus images, particularly in weak-texture areas like the macula, can be attributed to the mathematical properties of Conditional Flow Matching (CFM). Unlike traditional diffusion models that rely on stochastic Langevin dynamics, CFM learns a deterministic probability path. Specifically, the linear interpolation path $y_t = (1-t)y_0 + ty_1$ induces a constant target velocity $v_t = y_1 - y_0$. In regions with weak textures, standard segmentation networks often suffer from vanishing gradients or local optima due to the lack of distinct intensity transitions. However, the CFM refiner operates in the latent space by regressing the vector field $v_\theta(y_t, t|c) \approx y_1 - y_0$. This allows the model to "transport" noisy, fragmented predictions toward the high-fidelity ground-truth distribution through a direct geometric path, effectively bypassing the ambiguity of pixel-level intensity gradients.

Furthermore, the Spatially Heterogeneous Timestep (SHT) strategy further enhances this interpretability by decoupling the refinement dynamics. By sampling a spatial time map $T \in [0,1]^{H\times W}$, the model handles varying complexity across the image. In "easy" uniform regions, the flow integration converges rapidly, while at the weak boundaries of the macula, the SHT forces the network to learn sophisticated vector field dynamics across different temporal stages. This ensures that even when local image contrast is low, the global structural prior imposed by the flow field maintains strict anatomical consistency.

## V. Conclusion

In this paper, we propose FMSSL-GAN, a semi-supervised segmentation framework aimed at fundus image quality assessment under conditions of few annotations and low quality. Specifically, a conditional flow-matching refiner is incorporated into the weak/strong consistency pipeline to explicitly learn continuous transport paths between coarse pseudo-label distributions and label-like representations. This design restricts gradient propagation to refined areas confirmed as high-confidence, which can effectively boost the robustness against noisy pseudo-labels. In parallel, we imposed an adversarial structural constraint in the mask space to recover global shape plausibility and enforce topological consistency, thereby mitigating fragmented masks, jagged boundaries, and small false positives that commonly arise in low-contrast fundus images. Experiments on the Szeye, REFUGE, and ORIGA datasets demonstrate that FMSSL-GAN achieves a promising segmentation performance across three different labeled data ratios. Furthermore, ablation studies confirm that CFM refinement, the heterogeneous timestep strategy, and the GAN discriminator are all necessary and complementary components. Future directions include extending the flow-matching refiner to multi-class ophthalmic segmentation, jointly learning confidence scheduling with refinement, and evaluating distribution-level mask priors in cross-device/domain generalization scenarios. While the current framework primarily focuses on geometric standardization via rotation angle measurement with an adaptable classification threshold, we acknowledge that clinical quality control also encompasses dimensions such as occlusion, blurring, and illumination consistency. The flexibility of the continuous angle regression allows our method to be easily calibrated for diverse clinical deployment scenarios. Our FMSSL-GAN model is designed with high robustness to handle the weak textures and noisy boundaries often caused by these factors. In future work, we plan to extend this framework by integrating auxiliary quality scoring heads to provide a multi-dimensional assessment of fundus image quality.

## References

[1] Y. Luo, K. Chen, L. Liu, J. Liu, J. Mao, G. Ke, and M. Sun, "Dehaze of cataractous retinal images using an unpaired generative adversarial network," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3374–3383, 2020.

[2] F. Shao, Y. Yang, Q. Jiang, G. Jiang, and Y.-S. Ho, "Automated quality assessment of fundus images via analysis of illumination, naturalness and structure," *IEEE Access*, vol. 6, pp. 806–817, 2018.

[3] Q. Hou, P. Cao, L. Jia, L. Chen, J. Yang, and O. R. Zaiane, "Image quality assessment guided collaborative learning of image enhancement and classification for diabetic retinopathy grading," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1455–1466, 2023.

[4] J. Liu, C. Desrosiers, D. Yu, and Y. Zhou, "Semi-supervised medical image segmentation using cross-style consistency with shape-aware and local context constraints," *IEEE Transactions on Medical Imaging*, vol. 43, no. 4, pp. 1449–1461, 2024.

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on August 07,2026 at 07:33:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3717225




[5] H. Fu, B. Wang, J. Shen, S. Cui, Y. Xu, J. Liu, and L. Shao, "Evaluation of retinal image quality assessment networks in different color-spaces," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, ser. Lecture Notes in Computer Science, vol. 11764. Springer, 2019, pp. 48–56.

[6] E. Volkov and A. Averkin, "Fundus image quality assessment: a brief review of techniques," in *International Conference on Information Processes and Systems Development and Quality Assurance (IPS)*, 2024, pp. 50–54.

[7] R. Jiao, Y. Zhang, D. Wei, Y. Wu, Y. Tian, C. Ma, X. Yang, and Y. Xu, "Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, no. 5, pp. 1684–1718, 2024.

[8] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 596–608.

[9] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 7236–7246.

[10] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 1195–1204.

[11] Y. Liu, Y. Tian, C. Wang, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Translation consistent semi-supervised segmentation for 3d medical images," *IEEE Transactions on Medical Imaging*, vol. 44, no. 2, pp. 952–968, 2025.

[12] Y. Zou, Z. Yu, B. V. K. Vijaya Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 297–313.

[13] Y. Zhu, Z. Zhang, C. Wu, Z. Zhang, T. He, H. Zhang, R. Manmatha, M. Li, and A. Smola, "Improving semantic segmentation via efficient self-training," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1589–1602, 2024.

[14] Y. Yang, L. Zhang, and L. Ren, "Semi-supervised medical image segmentation via cross teaching between MobileNet and MobileViT," *Image and Vision Computing*, vol. 150, p. 105196, 2024.

[15] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2613–2622.

[16] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 4238–4247.

[17] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," *arXiv preprint arXiv:2101.06329*, 2021.

[18] H. Kervadec, J. Dolz, É. Granger, and I. Ben Ayed, "Curriculum semi-supervised segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Shenzhen, China, 2019, pp. 568–576.

[19] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Shenzhen, China, 2019, pp. 605–613.

[20] X. Liu, W. Li, and Y. Yuan, "DiffRect: Latent diffusion label rectification for semi-supervised medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Marrakesh, Morocco, 2024, pp. 56–66.

[21] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models in medical imaging: A comprehensive survey," *Medical Image Analysis*, vol. 88, p. 102846, 2023.

[22] J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, and Y. Xu, "MedSegDiff-V2: Diffusion-based medical image segmentation with transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 6030–6038.

[23] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, and Y. Xu, "MedSegDiff: Medical image segmentation with diffusion probabilistic model," in *Medical Imaging with Deep Learning (MIDL)*, ser. Proceedings of Machine Learning Research, vol. 227. PMLR, 2024, pp. 1623–1639.

[24] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin, "Diffusion models for implicit image segmentation ensembles," in *Proc. Int. Conf. Med. Imag. Deep Learn. (MIDL)*, Lübeck, Germany, 2021, pp. 1–13.

[25] W. Chen, T. Zhang, and X. Zhao, "Semantic segmentation using generative adversarial network," in *Proc. 40th Chin. Control Conf. (CCC)*, Shanghai, China, 2021, pp. 8492–8495.

[26] W. Liu, C. Zhang, H. Ding, T.-Y. Hung, and G. Lin, "Few-shot segmentation with optimal transport matching and message flow," *IEEE Transactions on Multimedia*, vol. 25, pp. 5130–5141, 2023.

[27] W. Liu, W. Zhou, J. Liu, P. Hu, J. Cheng, J. Han, and W. Lin, "Modality-aware feature matching: A comprehensive review of single- and cross-modality techniques," *arXiv preprint arXiv:2507.22791*, 2025.

[28] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, "SegAN: Adversarial network with multi-scale $L_1$ loss for medical image segmentation," *Neuroinformatics*, vol. 16, no. 3, pp. 383–392, 2018.

[29] S. Shit, J. C. Paetzold, A. Sekuboyina, I. Ezhov, A. Unger, A. Zhylka, J. P. W. Pluim, B. H. Menze, and G. Tetteh, "clDice - a novel topology-preserving loss function for tubular structure segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 560–16 569.

[30] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. Ben Ayed, "Boundary loss for highly unbalanced segmentation," *Medical Image Analysis*, vol. 67, p. 101851, 2021.

[31] A. Tong, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, K. Fatras, G. Wolf, and Y. Bengio, "Conditional flow matching: Simulation-free dynamic optimal transport," *arXiv preprint arXiv:2302.00482*, 2023.

[32] M. Ortega *et al.*, "REFUGE: Retinal fundus glaucoma challenge," distributed by Grand Challenge, 2018, [Online]. Available: https://refuge.grand-challenge.org/. Accessed: 2024.

[33] Z. Zhang *et al.*, "ORIGA: Online retinal fundus image database for glaucoma analysis," distributed by NIMTE, 2010, [Online]. Available: https://imed.nimte.ac.cn/dataoforiga.html. Accessed: 2024.

[34] X. Luo, "SSL4MIS," GitHub repository, 2020, [Online]. Available: https://github.com/HiLab-git/SSL4MIS.

[35] X. Luo, G. Wang, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, D. N. Metaxas, and S. Zhang, "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Med. Image Anal.*, vol. 80, p. 102517, 2022.


Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on August 07,2026 at 07:33:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.