Geometric Correspondence Constrained Pseudo-Label Alignment for Source-Free Domain Adaptive Fundus Image Segmentation

Zhouhongyuan Hu, Lei Zhang, Lituan Wang[*], Zhenwei Zhang, Minjuan Zhu, Zhenbin Wang
School of Computer Science, Sichuan University, Chengdu, China
hzhy@stu.scu.edu.cn, {leizhang, lituanwang}@scu.edu.cn, {zhangzw, zhuminjuan, wangzhenbin}@stu.scu.edu.cn

Abstract

Source-free unsupervised domain adaptation (SF-UDA), which relies only on a pre-trained source model and unlabeled target data, has gained significant attention. Pseudo-labeling, valued for its simplicity and effectiveness, is a key approach in SF-UDA. However, existing methods neglect the consistency priors of anatomical features across samples, leading them fail to revise of high-confidence noise in structurally inconsistent regions, ultimately manifesting as significant discrepancies in pseudo-labeled samples especially in limited source data scenarios. Motivated by this insight, we propose a novel Geometric Correspondence Constrained (GCC) pseudo-labeling framework. GCC first stratifies pseudo-labeled samples into high/low-quality subsets. It then refines low-quality samples by leveraging the anatomical features inherent in high-quality samples while injecting Gaussian perturbation to perturb high-confidence noise towards the decision boundaries. This process effectively mitigates high-confidence noise disruptive effect and preserves critical prior anatomical knowledge, making it particularly powerful for scenarios with limited source data. Experiments on cross-domain fundus image datasets demonstrate that our method achieves state-of-the-art performance.

Code — https://github.com/PHDhzhy/GCC

Introduction

While deep neural networks have achieved significant progress in medical image segmentation, especially for the analysis of fundus images such as optic cup/disc, these methods typically require extensive pixel-level annotations (Tajbakhsh et al. 2020). Furthermore, data collected across different hospitals often exhibit substantial distribution discrepancies (domain shift) due to variations in acquisition devices and scanning protocols (Wang et al. 2019), causing models trained on data from one source to generalize poorly to others. To simultaneously address the challenges of annotation scarcity and domain shift under privacy-preserving constraints (Chen et al. 2021), source-free unsupervised domain adaptation (SF-UDA) has emerged as a critical paradigm, which use only a pre-trained source model and unlabeled target data (Bateson et al. 2022; Niu et al. 2022; Yang et al. 2022; Zeng et al. 2024; Chen et al. 2021; Huai et al. 2023; Xu et al. 2022; Tang et al. 2023; Tian et al. 2023; Li, Zhou, and Yang 2024; Zhang et al. 2024; Zheng et al. 2025).

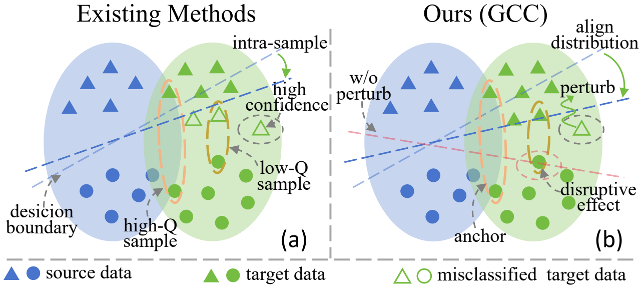


Figure 1: (a) Existing methods struggle to refine high-noise samples under unreliable boundaries, yielding pronounced anatomical discrepancies across pseudo-labeled samples. (b) GCC stratifies samples as high/low quality, aligns distributions using high-quality anchors, and optimizes alignment via Gaussian perturbation.

Existing SF-UDA methods can be roughly divided into three categories: entropy minimization enhances prediction certainty on target data, source image approximation reconstructs source-domain characteristics without source data access, and pseudo-labeling leverages high-confidence predictions as self-training supervisory labels. Among these approaches, pseudo-labeling has gained prominence due to its direct supervisory mechanism, which focuses on discarding or correcting erroneous pseudo-labels to provide targeted training signals (Chen et al. 2021; Xu et al. 2022; Huai et al. 2023; Tang et al. 2023; Zhang et al. 2024). This characteristic makes it particularly suitable for medical applications where annotation scarcity necessitates effective unsupervised learning. However, current pseudo-labeling approaches focus primarily on intra-sample refinement while critically neglecting cross-sample anatomical consistency priors inherent in data distributions. Under source-scarce scenarios frequently caused by data access restrictions and expert annotation constraints, source models develop unreliable decision boundaries for target

[*] Corresponding Author

data (Zhao et al. 2022). Consequently, pseudo-labels exhibit substantial noise contamination. Sole reliance on intra-sample refinement proves inadequate for mitigating significant residual noise, particularly high-confidence noise persisting in structurally inconsistent regions, manifesting as pronounced anatomical discrepancies across samples (as illustrated in Figure 1(a)). These inconsistencies ultimately disseminate erroneous anatomical priors that degrade model performance. This inspire us to develop a geometric correspondence constrained alignment strategy to improve SF-UDA for medical image segmentation.

In this paper, as illustrated in Figure 1(b), we propose a novel Geometric Correspondence Constrained (GCC) pseudo-labeling framework for SF-UDA, addressing pseudo-labeled samples distribution discrepancies under limited source data. Specifically, we first develop a pseudo-labeled sample quality judgment strategy to stratify pseudo-labeled samples into high/low-quality (high/low-Q) subsets. Subsequently, high-Q samples serve as anchors to align the distribution of low-Q samples toward these anchors. At the same time, to mitigate the disruptive effects of high-confidence noise during alignment, we develop a Sample-Adaptive Perturbation Estimation (SAPE) mechanism that dynamically injects Gaussian perturbations to each sample, perturbing high-confidence noise toward decision boundaries, maximizing error correction while preserving correct predictions. While CPR (Huai et al. 2023) resolves local semantic consistency and IPLC (Zhang et al. 2024) relies on external SAM models, GCC establishes *global* anatomical constraints through unsupervised geometric correspondence without external supervision. Unlike prototype-based methods (Chen et al. 2021), our high-Q anchors preserve fine-grained structural priors lost in averaging. As a result, our proposed method enables effective propagation of anatomical priors through pseudo-label distribution alignment. To sum up, the main contributions of this paper are as follows:

- We presents a novel SF-UDA approach, GCC, to unify pseudo-labeled samples distribution for refinement while injecting SAPE-controlled perturbation for high-confidence noise correction.
- We establish the first anatomy-aware distribution alignment paradigm through unsupervised stratification of pseudo-labels and geometric knowledge transfer from high-Q anchors to low-Q samples, enabling cross-sample anatomical consistency propagation without external supervision.
- We present the first work to integrate adaptive perturbation into pseudo-label denoising, with rigorous theoretical proof of its critical role in alignment efficacy (see Supplement). Experimental results demonstrate the superior performance of our method over current SF-UDA approaches.

Related Works

Source-free unsupervised domain adaptation

Domain shift caused by heterogeneous imaging protocols and modalities presents a fundamental challenge in medical image analysis. SF-UDA has emerged as a critical paradigm to address this issue under clinical data privacy constraints, where only a pre-trained source model and unlabeled target data are accessible. Existing SF-UDA methods can be roughly divided into three categories: entropy minimization (Bateson et al. 2022, 2020; Wang et al. 2020; Niu et al. 2022, 2023), source image approximation (Yang et al. 2022; Zeng et al. 2024), and pseudo-labeling (Chen et al. 2021; Hou and Zheng 2021; Huai et al. 2023; Tang et al. 2023; Tian et al. 2023; Xu et al. 2022; Zhou, Ye, and Xiao 2022; Li, Zhou, and Yang 2024; Zhang et al. 2024; Zheng et al. 2025).

Entropy minimization, which assumes that more confident model predictions lead to better generalization, aims to reduce generalization error in the target domain by minimizing prediction entropy (Bateson et al. 2020). TENT (Wang et al. 2020) pioneered this direction by minimizing prediction entropy during test-time adaptation, dynamically updating batch normalization layers to reduce distribution shift. Subsequent innovations like ETTA (Niu et al. 2022) improved efficiency through sample selection, prioritizing reliable and non-redundant target instances, while SAR (Niu et al. 2023) enhanced stability by identifying and mitigating performance-degrading factors during adaptation. However, these methods risk reinforcing incorrect predictions when confronted with high-confidence errors inherent in medical images, particularly under significant domain gaps where entropy signals become unreliable.

Source approximation techniques bridge domain gaps by generating source-like target representations. FSM (Yang et al. 2022) leverages Fourier domain transformations to synthesize source-style images, enabling feature-space alignment without accessing source data. Recent advances incorporate generative models: RSA (Zeng et al. 2024) preserved anatomical structures through edge-guided diffusion with uncertainty filtering, while D2SFDA (Zhou et al. 2024) proposed diffusion perturbation flow to generate semantically consistent target views using pseudo-labels as conditional inputs for diffusion models. While effective in narrowing domain discrepancies, these methods often compromise fine-grained textures and class-specific variations during reconstruction, limiting their ability to capture the full complexity of medical imaging distributions.

Compared to entropy minimization and source approximation, pseudo-labeling has gained significant traction as the dominant paradigm in SF-UDA due to its intrinsic ability to directly provide supervisory signals for target domain adaptation without reconstructing source data or relying solely on prediction confidence. This characteristic makes it particularly well-suited for medical imaging tasks where anatomical structures exhibit inherent spatial regularity. Our GCC also aims to improve the performance of the model by dealing with pseudo-labels.

SF-UDA based Pseudo-labeling

Pseudo-labeling techniques mitigate domain shifts by generating target pseudo-labels for self-training, focusing on discarding or correcting erroneous labels induced by distribution discrepancies. Current research advances are centered around three primary strategies. The first paradigm enhances pseudo-label quality via denoising mechanisms: CPR (Huai et al. 2023) resolves anatomical incoherence through contextual similarity learning, which corrects ambiguous boundaries by enforcing local semantic consistency, while IPLC (Zhang et al. 2024) iteratively refines labels using entropy-weighted correction and robust prompts generated by the Segment Anything Model (SAM). The second paradigm screens unreliable pseudo-labels; notable implementations include DPL (Chen et al. 2021) dual-level denoising, which dynamically masks noisy pixels based on uncertainty and selects category-consistent labels via prototype alignment, and U-D4R (Xu et al. 2022) that employs Monte Carlo Dropout for uncertainty quantification with adaptive thresholding. The third paradigm designs robust training dynamics to counter error propagation: CBMT (Tang et al. 2023) mitigate class imbalance via global statistics-guided loss calibration, RIOG (Yan et al. 2025) align gradient directions and magnitudes across tasks to improve minority-class labels, and PLPB (Li, Zhou, and Yang 2024) enhance boundary robustness by combining adversarial training with pseudo-boundary constraints.

Despite their efficacy, existing methods are constrained to intra-sample information, neglecting rich geometric correspondences and anatomical consistency across samples. This oversight manifests as inconsistent pseudo-labeled samples, where low-Q samples exhibit anatomically implausible distortions misaligned with their high-Q counterparts. Such inconsistencies propagate segmentation errors during self-training. To address this limitation, our work bridges the gap by leveraging across-sample structural knowledge for pseudo-label refinement.

Methods

Let $f^s : \mathcal{X}^s \to \mathcal{Y}^s$ denote the model trained on the source domain $\mathcal{D}^s = (\mathcal{X}^s, \mathcal{Y}^s)$, and $\mathcal{D}^t$ denotes the target domain. The goal of SF-UDA is to construct a model $f^{s\to t}$ that performs well in the target domain with only the unlabeled dataset $\{x_i^t\}_{i=1}^{n^t}$, where $x_i^t \in \mathcal{X}^t$. Generally, the segmentation of the fundus image can be treated as a multi-label segmentation task with $x_i \in \mathbb{R}^{H\times W\times 3}$ and $y_i \in \{0,1\}^{H\times W\times C}$, where C denotes the number of classes.

As illustrated in Figure 2, our Geometric Correspondence Constrained (GCC) pseudo-labeling framework consists of three stages. In this section, we first introduce the pseudo-labeled sample quality judgment strategy. Next, we propose the pseudo-label alignment framework. The training procedures with the refined pseudo-labels for the model adaption are finally described.

Pseudo-labeled Sample Quality Judgment

To ensure distribution consistency among all generated pseudo-labeled samples, we focus on the low-Q pseudo-labeled sample judgment and the refinement. As illustrated in in Figure 2(a), high-entropy pseudo-labeled samples demonstrate reduced structural rationality due to quasi-circular anatomical prior of optic cup/disc, exhibiting characteristic low Iso-Perimetric Quotient (Li, Goodchild, and Church 2013) compared to low-entropy counterparts. Therefore, we utilize entropy values as the basis for assessing the quality of pseudo-labeled sample. Given a target image x^t and source model f^s, we obtain the context-aware probability p_v for each pixel following CPR (Huai et al. 2023).

Recognizing that pseudo-label quality is inherently relative, we establish a sample-wise ranking criterion achieved by computing the mean entropy for each probability map p. The entropy for the i-th sample in pseudo-labeled sample set of size N is calculated as:

$$E_i = -\frac{1}{H\times W}\sum_{v=1}^{H\times W} p_v(i)\log p_v(i). \tag{1}$$

Samples are then sorted in descending order of entropy ($E_1 \geq E_2 \geq \cdots \geq E_N$), establishing an inverse correlation between entropy rank and pseudo-label quality: higher-ranked samples with greater entropy exhibit lower quality.

To accommodate the evolving alignment capability of the refinement network during training, we implement a dynamic stratification strategy. This strategy progresses from initial conservative filtering, where low-Q assignments are limited to the lowest entropy samples in early stages, to a gradual expansion that applies increasingly strict quality thresholds as the network's alignment competence improves. This progression is formalized through a time-dependent selection function:

$$r(iter) = \alpha + (1-\alpha)\cdot\frac{iter}{Iter}, \tag{2}$$

where α define the proportion of samples designated as low-Q during the initial training phases, $iter$ is the current epoch, and $Iter$ is the total epochs. The quality assignment is then determined for each sample as follows:

$$p(i) \in \begin{cases}\mathcal{P}_{low} & \text{if } \mathrm{rank}(E_i) \leq \lceil r(iter)\cdot N\rceil, \\ \mathcal{P}_{high} & \text{otherwise,}\end{cases} \tag{3}$$

with $\mathrm{rank}(E_i) = k$ indicating $E_i = E_k$ in the sorted sequence.

Pseudo-label Alignment

After partitioning pseudo-labeled samples into high-Q ($\mathcal{P}_{high}$) and low-Q ($\mathcal{P}_{low}$) subsets, refining low-Q pseudo-labeled samples is naturally achieved by aligning them with high-Q counterparts through geometric correspondence. We introduce a GAN with generator G and discriminator D to enforce structural and semantic consistency (Figure 2(b)). However, high-confidence noise in structurally inconsistent regions presents a critical challenge where enforcing anatomical consistency may corrupt correct pseudo-labels adjacent to high-confidence noisy areas. To address this limitation, our pseudo-label alignment implements a dual-stage refinement strategy. First, randomized perturbations establish foundational robustness and smoothness through anatomical consistency learning, building basic refinement capability. Second, sample-specific perturbation injection expands variance of pseudo-labels, perturbing high-confidence noise toward the correct side of decision boundaries while preserving semantically correct information with maximal possibility.

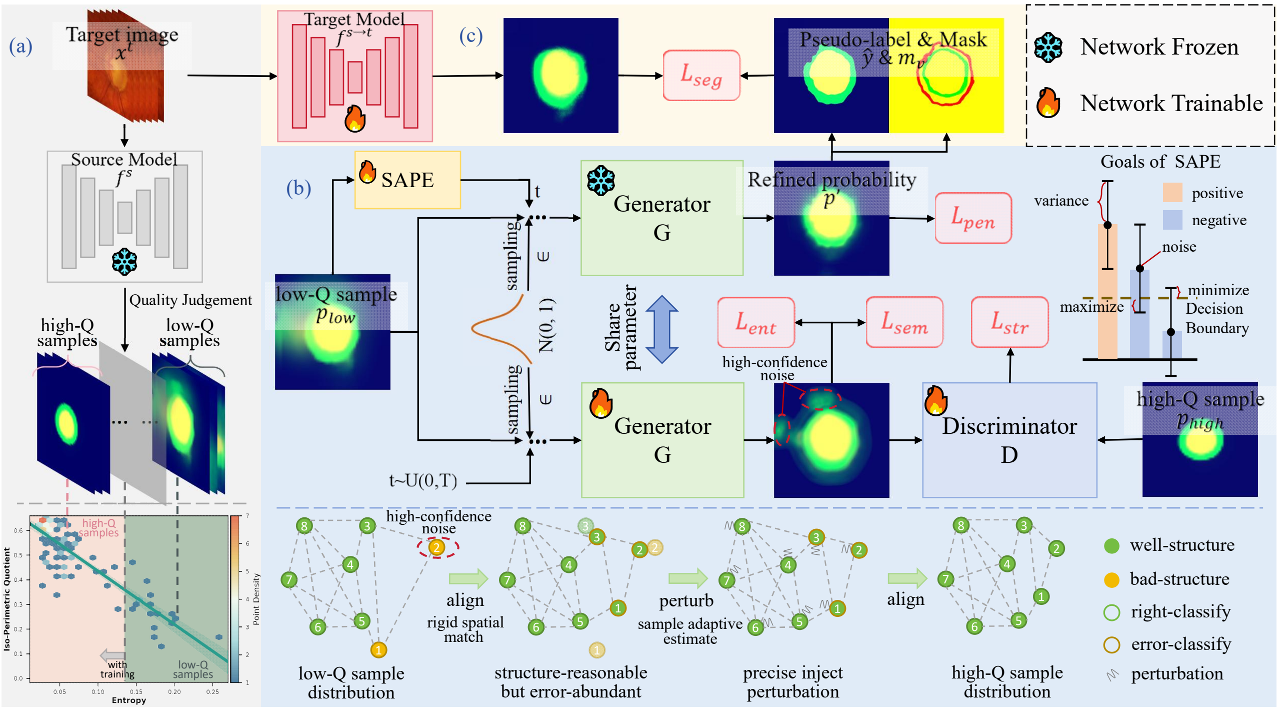


Figure 2: The overview of Geometric Correspondence Constrained (GCC) pseudo-labeling framework. It consists of three stages: (a) Generating probability maps and judging the quality of probability maps; (b) The Pseudo-label Alignment, where low-Q samples are aligned after injecting with adaptive Gaussian perturbation; (c) Model adaptation with refined pseudo-labels.

Geometric correspondence learning. To develop fundamental noise immunity, this stage enhances decision boundary stability and smoothness through controlled randomization. During training, we apply randomized Gaussian perturbation to low-Q probability maps. For each $p_{low} \in \mathcal{P}_{low}$, we generate perturbed versions via linear interpolation:

$$\begin{aligned} p^t_{low} &= (1-r)\cdot p_{low} + r\cdot\epsilon, \quad \epsilon\sim\mathcal{N}(0,1), \\ r &= \frac{t}{T}, \quad t\sim U(0,T), \end{aligned} \tag{4}$$

where t is uniformly sampled from discrete timesteps. The generator G learns to denoise these perturbed inputs under triple composite objectives:

$$\begin{aligned} \mathcal{L}_{str} &= \mathbf{E}[\log D(p_{high})] + \mathbf{E}[\log(1-D(G(p^t_{low})))], \\ \mathcal{L}_{sem} &= \mathbf{E}[\mathrm{e}^{-p_{low} log(p_{low})}\cdot\|G(p^t_{low}) - p_{low}\|_1], \\ \mathcal{L}_{ent} &= \mathbf{E}[G(p^t_{low}) log(G(p^t_{low}))], \end{aligned} \tag{5}$$

minimizing the composite loss $\mathcal{L}_{gan} = \mathcal{L}_{str} + \lambda_{sem}\mathcal{L}_{sem} + \lambda_{ent}\mathcal{L}_{ent}$. $\mathcal{L}_{sem}$ enforces semantic consistency with uncertainty weighting, while $\mathcal{L}_{ent}$ minimizes prediction entropy to sharpen outputs. This adversarial framework simultaneously enforces structural consistency through discrimination, establishing fundamental refinement capability that reliably revises relative low-confidence errors under moderate perturbation conditions. However, for high-confidence noise, the semantic consistency constraint and structural consistency constraint designed herein conflict with each other, compromising the refinement effectiveness.

Sample-adaptive perturbation estimation. To specifically combat high-confidence noise in structurally inconsistent regions, this module learns optimal perturbation intensities that maximally correct labeling errors. In essence, we implement a ”tipping point” perturbation philosophy that pursues the Goldilocks Zone of intensit, neither excessive nor insufficient, to identify the precisely calibrated strength for each sample. This optimal intensity flips persistent noise while preserving valid anatomical structures. Relevant proofs are provided in the supplement.

Specifically, following GAN training, we freeze the generator’s parameters to stabilize the decision boundary and train the Sample-Adaptive Perturbation Estimation (SAPE) module to discover sample-specific optimal noise ratios. The SAPE architecture combines a feature extractor F with Gumbel-Softmax (Jang, Gu, and Poole 2016) to estimate gradients for discrete distributions (Pang et al. 2022), mapping p_{low} to optimized timesteps $t \in (0,T)$. To efficiently navigate the expansive search space, we implement a hierarchical decomposition strategy:

$$t = \underbrace{z^{(2)}\odot\mathbf{N2}}_{hundreds}\times 100 + \underbrace{z^{(1)}\odot\mathbf{N2}}_{tens}\times 10 + \underbrace{z^{(0)}\odot\mathbf{N1}}_{units} \tag{6}$$

where $z = argmax(GS(F(p_{low})))$ via differentiable categorical sampling, with $\mathbf{N}_1 = [1,...,9]^T$ and $\mathbf{N}_2 = [0,...,9]^T$. This hierarchical encoding dramatically reduces the perturbation intensity search space dimensionality. The resulting t determines noise ratio $r = t/T$ in Eq.(4), enabling precise perturbation calibrated to each sample’s noise characteristics.

This optimal perturbation strategically transitions erroneous pseudo-labels across decision boundaries while maintaining correct labels within their original classification regions. SAPE parameters are optimized end-to-end using an entropy-aware penalty:

$$\mathcal{L}_{pen} = \mathbf{E}[p_{low} log(p_{low}) \cdot \mathrm{e}^{G(p^t_{low}) - p_{low}}], \tag{7}$$

where the exponential term amplifies penalties for high-confidence misalignments, driving the module toward maximally effective denoising configurations.

Model Adaptation with Refined Pseudo-Labels

After training, the reliabel pseudo-label $\hat{y}_v$ is calculated by:

$$\begin{aligned} t = SAPE(p), p \in P_{low} \cup P_{high}, \\ \hat{y}_v = \mathbb{1}[p'_v \geq \gamma], p'_v = G(p^t_v). \end{aligned} \tag{8}$$

Considering that the injection of Gaussian perturbation inevitably disrupts semantic information, filtering reliable pseudo-labels at the pixel-level to constrain model training to reduce the error impact caused by Gaussian perturbation:

$$m_v = \mathbb{1}(p'_v < \gamma_{low} \text{ or } p'_v > \gamma_{high}), \tag{9}$$

where γ_{low} and γ_{high} are two thresholds for filtering out pseudo-labels without confident probabilities. The target model $f^{s \to t}$ is trained with cross-entropy loss:

$$\mathcal{L}_{seg} = \mathbf{E}[m_v \cdot |\hat{y}_v \log(f^t(x^t)_v) + (1 - \hat{y}_v) \log(1 - f^t(x^t)_v)|]. \tag{10}$$

Experiments

Dataset. Building on previous research (Chen et al. 2021; Huai et al. 2023), we opted for three datasets for fundus image segmentation: Drishti-GS (Sivaswamy et al. 2015), RIM-ONEr3 (Fumero et al. 2011), and the validation set of the REFUGE Challenge (Orlando et al. 2020). These datasets were partitioned into 50/51, 99/60, and 320/80 splits for training and testing, respectively. We designate Drishti-GS with the smallest volume as the source domain to simulate limited source data scenarios.

Implementation details and evaluation metrics. Following prior works (Chen et al. 2021; Huai et al. 2023; Wang et al. 2019; Xu et al. 2022), our segmentation network is MobileNetV2-adapted (Sandler et al. 2018) DeepLabv3+ (Chen et al. 2018). Our G consists of multi-layer ResNet (He et al. 2016), and D consists of multi-layer convolution, and F consists of multiple layers of convolution and three-branch fully connected layers. We set $\alpha = 0.2$ as the proportions of low-Q samples during the initial training phases. To ensure SAPE's fine-grained estimation, we set a sufficiently large $T = 1000$. For loss function, $\lambda_{sem} = 5$ and $\lambda_{ent} = 1$. The remaining hyperparameters follow established practices from prior work (Chen et al. 2021; Huai et al. 2023), where the confidence threshold γ is set to 0.75, and two thresholds for filtering out unconfident refined pseudo-labels are set as $\gamma_{low} = 0.4$ and $\gamma_{high} = 0.85$. Each image is pre-processed by clipping a 512 × 512 optic disc region (Wang et al. 2019). The same augmentations as in (Chen et al. 2021; Huai et al. 2023; Xu et al. 2022) are applied, including Gaussian noise, contrast adjustment, and random erasing. The Adam optimizer is adopted with learning rates of 2e-4 and 3e-4 in the geometric correspondence learning stage the target domain adaptation stage respectively. The momentum of the Adam optimizer is set to 0.9 and 0.99. The batch size is set to 8. This implementation was executed using PyTorch on an NVIDIA GeForce RTX 3090 GPU. For evaluation, we utilized widely adopted metrics such as the Dice coefficient and Average Surface Distance (ASD).

Comparision with the state-of-the-arts. Table 1 compares our method with state-of-the-art fundus segmentation approaches, including UDA methods (BEAL (Wang et al. 2019), CLR (Feng et al. 2022)) and SF-UDA techniques (SRDA (Bateson et al. 2020), TENT (Wang et al. 2020), DPL (Chen et al. 2021), FSM (Yang et al. 2022), U-D4R (Xu et al. 2022), CPR (Huai et al. 2023), CBMT (Tang et al. 2023), PLPB (Li, Zhou, and Yang 2024)). Results demonstrate that the GCC framework significantly outperforms existing methods by leveraging geometric priors from high-Q pseudo-labels and strategically injecting Gaussian perturbations to refine low-Q samples. This advancement is particularly pronounced compared to SF-UDA methods lacking explicit pseudo-label distribution consistency constraints. The performance gap validates the necessity of enforcing cross-sample anatomical consistency, especially under limited-source conditions where conventional approaches fail to mitigate structural discrepancies. Notably, GCC's explicit geometric correspondence constraints enable superior performance on clinically critical metrics, achieving state-of-the-art ASD scores even compared to UDA methods that access target domain labels during training. Qualitative comparisons in Figure 3 further confirm these advantages, with our results exhibiting the most anatomically complete segmentations and closest alignment with GT annotations.

Ablation study on components. The ablation study systematically validate the necessity of each component in our framework, as shown in the Table 2. Solely applying Pseudo-label Alignment reduces performance, which confirms that rigid spatial matching corrupt correct pseudo-labels adjacent to high-confidence noisy areas. This observation is corroborated by our pseudo-label refinement ablation (Table 4), where removing Gaussian disturbance causes significant deterioration (-3.05% cup Dice). Integrating SAPE with geometric correspondence learning reverses this trend, demonstrating its role in correcting high-confidence noise through statistical distribution matching. While Filter alone achieves competitive results, its full synergy with alignment modules yields peak performance, proving our cascaded refinement philosophy: geometric correspondence learning establishes structural priors, SAPE mitigates high-confidence noise disrupt effect, and Filter eliminates outliers through confident probabilities constraints.

Ablation study on losses. Additionally, we conduct an ablation study on the loss functions of the geometric cor-

Methods	Source	Dice[%]↑			ASD[pixel]↓		
		Optic cup	Optic disc	Avg	Optic cup	Optic disc	Avg
Source: Drishti-GS; Target: REFUGE							
No adaptation	-	$42.87_{\pm 4.81}$	$58.87_{\pm 5.32}$	$50.87_{\pm 4.97}$	$51.24_{\pm 5.26}$	$52.28_{\pm 4.46}$	$51.76_{\pm 4.52}$
Upper bound	-	$87.38_{\pm 1.01}$	$95.40_{\pm 0.63}$	$91.39_{\pm 0.75}$	$4.32_{\pm 0.69}$	$3.67_{\pm 0.74}$	$3.99_{\pm 0.67}$
BEAL (Wang et al. 2019)	✓	$85.82_{\pm 1.77}$	$96.86_{\pm 0.21}$	$91.34_{\pm 0.95}$	$16.90_{\pm 3.98}$	$5.48_{\pm 0.42}$	$11.19_{\pm 2.12}$
CLR (Feng et al. 2022)	✓	$81.84_{\pm 0.97}$	$90.60_{\pm 0.62}$	$86.22_{\pm 0.68}$	$10.86_{\pm 1.07}$	$10.58_{\pm 0.93}$	$10.72_{\pm 0.87}$
DPL (Chen et al. 2021)	✗	$65.90_{\pm 5.42}$	$84.99_{\pm 4.38}$	$75.45_{\pm 4.08}$	$18.65_{\pm 4.46}$	$25.01_{\pm 8.71}$	$21.83_{\pm 5.26}$
FSM (Yang et al. 2022)	✗	$67.62_{\pm 5.38}$	$84.42_{\pm 3.85}$	$76.02_{\pm 4.19}$	$15.63_{\pm 3.98}$	$13.89_{\pm 6.84}$	$14.76_{\pm 4.97}$
CPR (Huai et al. 2023)	✗	$71.62_{\pm 4.27}$	$86.34_{\pm 2.55}$	$78.98_{\pm 2.47}$	$11.52_{\pm 2.89}$	$10.15_{\pm 2.18}$	$10.83_{\pm 1.34}$
CBMT (Tang et al. 2023)	✗	$66.92_{\pm 4.37}$	$\mathbf{90.96}_{\pm \mathbf{1.03}}$	$78.94_{\pm 2.39}$	$15.40_{\pm 2.89}$	$\mathbf{7.84}_{\pm \mathbf{1.27}}$	$11.62_{\pm 1.81}$
PLPB (Li et al. 2024)	✗	$74.46_{\pm 6.22}$	$85.76_{\pm 5.62}$	$80.11_{\pm 5.46}$	$12.21_{\pm 5.00}$	$21.21_{\pm 9.40}$	$16.71_{\pm 6.10}$
GCC (ours)	✗	$\mathbf{77.87}_{\pm \mathbf{3.26}}^{\dagger}$	$83.42_{\pm 2.74}$	$\mathbf{80.65}_{\pm \mathbf{1.46}}$	$\mathbf{8.84}_{\pm \mathbf{1.87}}^{\dagger}$	$11.55_{\pm 1.63}$	$\mathbf{10.19}_{\pm \mathbf{0.98}}^{\dagger}$
Source: Drishti-GS; Target: RIM-ONE-r3							
No adaptation	-	$61.39_{\pm 2.00}$	$78.43_{\pm 2.41}$	$69.91_{\pm 2.08}$	$33.15_{\pm 3.18}$	$33.29_{\pm 3.74}$	$33.22_{\pm 2.72}$
Upper bound	-	$82.07_{\pm 1.88}$	$95.60_{\pm 0.40}$	$88.84_{\pm 1.11}$	$8.42_{\pm 1.16}$	$5.44_{\pm 0.55}$	$6.93_{\pm 0.77}$
BEAL (Wang et al. 2019)	✓	$68.73_{\pm 2.37}$	$92.26_{\pm 1.11}$	$80.50_{\pm 1.57}$	$29.66_{\pm 6.50}$	$11.71_{\pm 2.53}$	$20.68_{\pm 4.16}$
CLR (Feng et al. 2022)	✓	$84.53_{\pm 0.67}$	$93.86_{\pm 0.40}$	$89.20_{\pm 0.33}$	$12.00_{\pm 0.97}$	$8.53_{\pm 0.60}$	$10.27_{\pm 0.44}$
SRDA (Bateson et al. 2020)	✗	$62.08_{\pm 21.48}$	$90.62_{\pm 15.48}$	$76.35_{\pm 18.48}$	$15.83_{\pm 8.51}$	$8.85_{\pm 7.93}$	$12.34_{\pm 8.22}$
TENT (Wang et al. 2020)	✗	$62.89_{\pm 19.83}$	$90.25_{\pm 10.41}$	$76.57_{\pm 15.12}$	$14.83_{\pm 8.62}$	$8.28_{\pm 6.95}$	$11.56_{\pm 7.79}$
DPL (Chen et al. 2021)	✗	$70.47_{\pm 2.68}$	$92.15_{\pm 1.41}$	$81.31_{\pm 1.37}$	$15.04_{\pm 1.73}$	$9.93_{\pm 4.02}$	$12.49_{\pm 2.33}$
FSM (Yang et al. 2022)	✗	$72.98_{\pm 4.84}$	$91.75_{\pm 2.55}$	$82.37_{\pm 3.63}$	$13.47_{\pm 3.18}$	$9.21_{\pm 2.65}$	$11.34_{\pm 2.59}$
U-D4R (Xu et al. 2022)	✗	$68.59_{\pm 13.87}$	$93.31_{\pm 5.41}$	$80.95_{\pm 9.64}$	$11.62_{\pm 5.66}$	$6.63_{\pm 6.05}$	$9.13_{\pm 5.86}$
CPR (Huai et al. 2023)	✗	$75.40_{\pm 2.65}$	$93.82_{\pm 1.42}$	$84.61_{\pm 1.75}$	$9.84_{\pm 3.10}$	$5.44_{\pm 1.97}$	$7.64_{\pm 2.38}$
CBMT (Tang et al. 2023)	✗	$70.15_{\pm 2.87}$	$89.98_{\pm 1.25}$	$80.06_{\pm 1.48}$	$12.26_{\pm 2.63}$	$9.39_{\pm 1.33}$	$10.82_{\pm 1.27}$
PLPB (Li et al. 2024)	✗	$71.70_{\pm 2.78}$	$90.22_{\pm 1.02}$	$80.96_{\pm 1.63}$	$13.70_{\pm 1.68}$	$10.56_{\pm 1.67}$	$12.13_{\pm 1.46}$
GCC (ours)	✗	$\mathbf{77.25}_{\pm \mathbf{2.00}}^{\dagger}$	$\mathbf{94.48}_{\pm \mathbf{0.72}}^{\dagger}$	$\mathbf{85.87}_{\pm \mathbf{1.35}}^{\dagger}$	$\mathbf{8.69}_{\pm \mathbf{0.82}}^{\dagger}$	$\mathbf{4.74}_{\pm \mathbf{0.57}}^{\dagger}$	$\mathbf{6.71}_{\pm \mathbf{0.67}}^{\dagger}$

Table 1: Comparison with state-of-the-arts on two settings. “No adaptation” refers to directly evaluating the source model on the target dataset. “Upper bound” refers to training the model on the target dataset with labels. Results are averaged across 10 independent runs. Statistical significance versus baselines determined by Wilcoxon signed-rank test (†: $p < 0.01$)

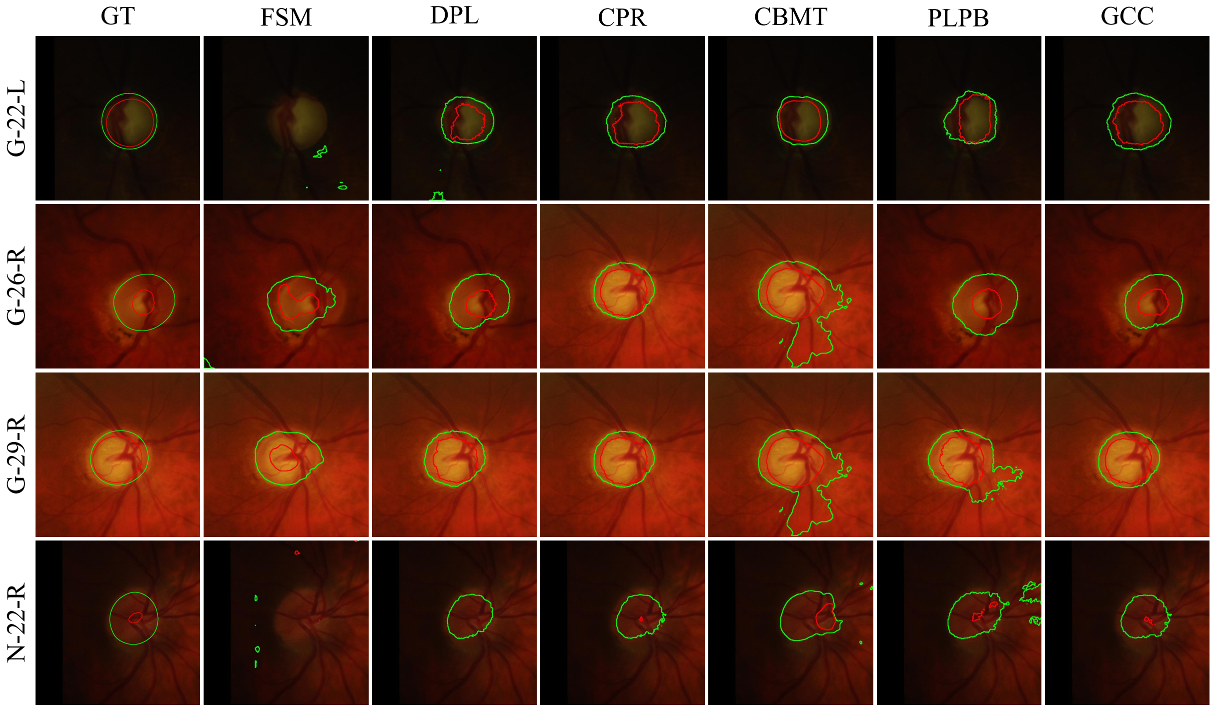


Figure 3: Comparison of the optic cup and disc segmentation results with different methods on the Drishti-GS to RIM-ONE-r3.

respondence learning module, as detailed in Table 3. The $\mathcal{L}_{sem}$ alone improves the Dice score by 0.26% and reduces the ASD by 0.77 compared to the baseline without both losses, demonstrating its effectiveness in preserving structural semantics. The $\mathcal{L}_{ent}$ alone, while not improving the Dice score, reduces the ASD by 0.86, indicating its role in refining boundary predictions. Critically, the synergistic integration of both losses yields a substantial performance gain, boosting the Dice score by 0.98% and reducing the ASD by 1.42. This underscores the complementary nature of the two losses: $\mathcal{L}_{sem}$ enforces high-level anatomical coherence while $\mathcal{L}_{ent}$ sharpens prediction confidence at ambiguous regions.

Alignment		Filter	Dice[%]		
Consistency	SAPE		Optic cup	Optic disc	Avg
✗	✗	✗	75.55	93.83	84.69
✓	✗	✗	74.71	93.40	84.01
✓	✗	✓	75.67	94.01	84.84
✓	✓	✗	76.15	93.78	84.97
✗	✗	✓	75.99	**94.67**	85.33
✓	✓	✓	**77.25**	94.48	**85.87**

Table 2: Ablation study of GCC components on the Drishti-GS to RIM-ONE-r3 adaptation (Alignment: Pseudo-label Alignment; Consistency: geometric correspondence learning; w/o SAPE: random perturbation ratio injection; Filter: pixel-level reliable pseudo-label filtering).

$\lambda_{sem}\mathcal{L}_{sem}$	$\lambda_{ent}\mathcal{L}_{ent}$	Avg.Dic	Avg.ASD
✗	✗	84.89	8.13
✓	✗	85.15	7.36
✗	✓	84.80	7.27
✓	✓	**85.87**	**6.71**

Table 3: Quantitative ablation study of geometric correspondence learning on the Drishti-GS to RIM-ONE-r3 adaptation.

Methods	Dice[%]	
	Optic cup	Optic disc
Initial pseudo-label	70.97	93.62
Refined pseudo-label	**72.34**	**93.72**
Refined pseudo-label (w/o perturb)	69.29	93.00

Table 4: Comparison of pseudo-label quality of the training set with different methods on the Drishti-GS to RIM-ONE-r3 adaptation.

Analysis of high-confidence noise. Figure 4 reveals a systematic prediction bias in pseudo-label generation: For both optic cup and disc segmentation at the 0.75 confidence threshold, false positives substantially outnumber false negatives (376k vs. 39k for cups; 321k vs. 92k for discs), indicating strong foreground over-prediction in background regions. This systematic bias motivates our SAPE design focused exclusively on background noise mitigation through precision-controlled perturbation intensity.

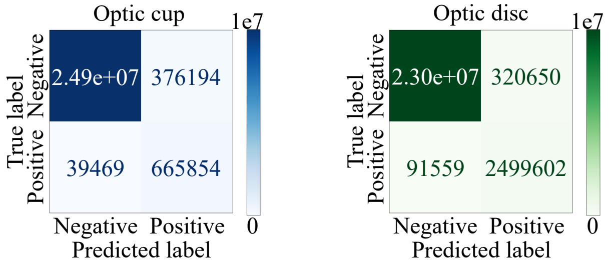
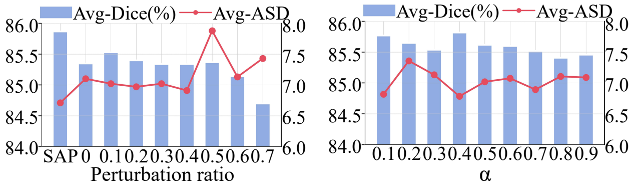


Figure 4: Confusion matrix analysis of pseudo-labels produced by (Huai et al. 2023) on the Drishti-GS to RIM-ONE-r3 adaptation.

SAPE vs. fixed perturbation ratios. To validate SAPE's capability for sample-adaptive estimation of optimal perturbation ratios, we compare it against fixed-ratio perturbations. Experimental results demonstrate SAPE's consistent performance superiority in domain adaptation (Figure 5(a)), confirming its effectiveness in enabling sample-specific perturbation estimation.



(a) Effectiveness of SAPE.

(b) Sensitivity of α.

Figure 5: Analysis of SAPE effectiveness and hyperparameter sensitivity on the Drishti-GS to RIM-ONE-r3 adaptation.

Analysis of hyperparameter sensitivity. We conduct a comprehensive sensitivity analysis of the hyperparameter α, with results presented in Figure 5(b). The constrained performance fluctuations demonstrate significant model robustness to α variations, indicating minimal dependence on precise hyperparameter tuning. This stability is particularly advantageous in clinical deployment scenarios where exhaustive parameter optimization is often impractical.

Conclusion

This work proposes GCC framework for SF-UDA in fundus image segmentation. GCC identify low-Q pseudo-labeled samples and refine them by leveraging the geometric knowledge inherent in high-Q pseudo-labels. Additionally, GCC controls Gaussian perturbation injection to mitigate the impact of high-confidence noise. Finally, the refined pseudo-labels are denoised with consideration of structural and semantic consistency, used for adaptation of model. Experimental results demonstrate that GCC outperforms other SF-UDA methods. Our approach provides a novel perspective on pseudo-label refinement in SF-UDA.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62376174, Grant 62025601, and Grant U24A20341.

References

Bateson, M.; Kervadec, H.; Dolz, J.; Lombaert, H.; and Ayed, I. B. 2022. Source-free domain adaptation for image segmentation. *Medical Image Analysis*, 82: 102617.

Bateson, M.; Kervadec, H.; Dolz, J.; Lombaert, H.; and Ben Ayed, I. 2020. Source-relaxed domain adaptation for image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, 490–499. Springer.

Chen, C.; Liu, Q.; Jin, Y.; Dou, Q.; and Heng, P.-A. 2021. Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, 225–235. Springer.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Feng, W.; Wang, L.; Ju, L.; Zhao, X.; Wang, X.; Shi, X.; and Ge, Z. 2022. Unsupervised domain adaptive fundus image segmentation with category-level regularization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 497–506. Springer.

Fumero, F.; Alayón, S.; Sanchez, J. L.; Sigut, J.; and Gonzalez-Hernandez, M. 2011. RIM-ONE: An open retinal image database for optic nerve evaluation. In *2011 24th international symposium on computer-based medical systems (CBMS)*, 1–6. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 630–645. Springer.

Hou, Y.; and Zheng, L. 2021. Visualizing adapted knowledge in domain transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13824–13833.

Huai, Z.; Ding, X.; Li, Y.; and Li, X. 2023. Context-aware pseudo-label refinement for source-free domain adaptive fundus image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 618–628. Springer.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Li, L.; Zhou, Y.; and Yang, G. 2024. Robust source-free domain adaptation for fundus image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 7840–7849.

Li, W.; Goodchild, M. F.; and Church, R. 2013. An efficient measure of compactness for two-dimensional shapes and its application in regionalization problems. *International Journal of Geographical Information Science*, 27(6): 1227–1250.

Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; and Tan, M. 2022. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, 16888–16905. PMLR.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*.

Orlando, J. I.; Fu, H.; Breda, J. B.; Van Keer, K.; Bathula, D. R.; Diaz-Pinto, A.; Fang, R.; Heng, P.-A.; Kim, J.; Lee, J.; et al. 2020. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59: 101570.

Pang, T.; Zhao, S.; Han, J.; Zhang, S.; Guo, L.; and Liu, T. 2022. Gumbel-softmax based neural architecture search for hierarchical brain networks decomposition. *Medical image analysis*, 82: 102570.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Sivaswamy, J.; Krishnadas, S.; Chakravarty, A.; Joshi, G.; Tabish, A. S.; et al. 2015. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. *JSM Biomedical Imaging Data Papers*, 2(1): 1004.

Tajbakhsh, N.; Jeyaseelan, L.; Li, Q.; Chiang, J. N.; Wu, Z.; and Ding, X. 2020. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical image analysis*, 63: 101693.

Tang, L.; Li, K.; He, C.; Zhang, Y.; and Li, X. 2023. Source-free domain adaptive fundus image segmentation with class-balanced mean teacher. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 684–694. Springer.

Tian, L.; Zhou, L.; Zhang, H.; Wang, Z.; and Ye, M. 2023. Robust self-supervised learning for source-free domain adaptation. *Signal, Image and Video Processing*, 17(5): 2405–2413.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.

Wang, S.; Yu, L.; Li, K.; Yang, X.; Fu, C.-W.; and Heng, P.-A. 2019. Boundary and entropy-driven adversarial learning for fundus image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, 102–110. Springer.

Xu, Z.; Lu, D.; Wang, Y.; Luo, J.; Wei, D.; Zheng, Y.; and Tong, R. K.-y. 2022. Denoising for relaxing: unsupervised domain adaptive fundus image segmentation without source data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 214–224. Springer.

Yan, F.; Yang, G.; Liu, A.; and Chen, X. 2025. RIOG: Rectify-to-match Gradient for Source-Free Domain Adaptive Medical Image Segmentation. *IEEE Sensors Journal*.

Yang, C.; Guo, X.; Chen, Z.; and Yuan, Y. 2022. Source free domain adaptation for medical image segmentation with fourier style mining. *Medical Image Analysis*, 79: 102457.

Zeng, H.; Zou, K.; Chen, Z.; Zheng, R.; and Fu, H. 2024. Reliable source approximation: Source-free unsupervised domain adaptation for vestibular schwannoma MRI segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 622–632. Springer.

Zhang, G.; Qi, X.; Yan, B.; and Wang, G. 2024. IPLC: iterative pseudo label correction guided by SAM for source-free domain adaptation in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 351–360. Springer.

Zhao, Z.; Zhou, F.; Xu, K.; Zeng, Z.; Guan, C.; and Zhou, S. K. 2022. LE-UDA: Label-efficient unsupervised domain adaptation for medical image segmentation. *IEEE transactions on medical imaging*, 42(3): 633–646.

Zheng, K.; Xia, H.; Xia, S.; Shao, M.; and Ding, Z. 2025. Supportive Negatives Spectral Augmentation for Source-Free Cross-Domain Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10573–10581.

Zhou, L.; Ye, M.; and Xiao, S. 2022. Domain adaptation based on source category prototypes. *Neural Computing and Applications*, 34(23): 21191–21203.

Zhou, W.; Ji, J.; Cui, W.; Yi, Y.; and Chen, Y. 2024. Diffusion-driven Dual-flow Source-Free Domain Adaptation for Medical Image Segmentation. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 4082–4087. IEEE.