Breaking Confirmation Bias: Single-Round Active Manifold Calibration for Source-Free Domain Adaptation in Segmentation-Supplementary Material

Anonymous ECCV 2026 Submission

Paper ID #10005

1 Proofs and Algorithm

In this section, we provide rigorous mathematical proofs for the theoretical bottlenecks and bounds, establishing the necessity of our Unsupervised Prototypical Alignment (UPA) and Label-Guided Manifold Calibration (LGMC).

1.1 Prototype Estimation Error Bound

To formalize the prototype estimation error, we explicitly assume the feature mapping $f(x) \in \mathbb{R}^d$ is bounded such that each coordinate $f(x)_j \in [-M, M]$.

Proposition 1 (Prototype Estimation Error Bound).
Let S_c be the subset of target samples with pseudo-label c and $n_c = |S_c|$. Define the oracle centroid

$$\mu_c^* := \mathbb{E}_{(x,y)\sim D_T}[\, f(x) \mid y = c\,],$$

and the empirical prototype

$$\hat{\mu}_c := \frac{1}{n_c} \sum_{x_i \in S_c} f(x_i).$$

Also define the misclassification mean $\mu_{err} := \mathbb{E}[\, f(x) \mid y \neq c, \hat{y} = c\,]$, the false discovery rate $\alpha_c := \mathbb{P}(y \neq c \mid \hat{y} = c)$, and the selection bias vector $\beta_c := \mathbb{E}[\, f(x) \mid y = c, \hat{y} = c\,] - \mu_c^*$.

Then, with probability at least $1 - \delta$, the estimation error satisfies

$$\|\hat{\mu}_c - \mu_c^*\|_2 \leq M \sqrt{\frac{2d \log(2d/\delta)}{n_c}} + (1 - \alpha_c)\|\beta_c\|_2 + \alpha_c \|\mu_{err} - \mu_c^*\|_2,$$

where the first term reflects sampling variability and the remainder captures structural bias.

Proof. By the triangle inequality,

$$\|\hat{\mu}_c - \mu_c^*\|_2 \leq \|\hat{\mu}_c - \mathbb{E}[\hat{\mu}_c]\|_2 + \|\mathbb{E}[\hat{\mu}_c] - \mu_c^*\|_2.$$

For the stochastic error, since each coordinate $j \in \{1, \ldots, d\}$ is bounded in $[-M, M]$ and the samples are i.i.d., we apply the scalar Hoeffding's inequality per coordinate. Using a union bound over all d dimensions, with probability at least $1 - \delta$ we have

$$\max_j |\hat{\mu}_{c,j} - \mathbb{E}[\hat{\mu}_{c,j}]| \leq M\sqrt{\frac{2\log(2d/\delta)}{n_c}}.$$

Using the norm inequality $\|v\|_2 \leq \sqrt{d}\|v\|_\infty$, this gives

$$\|\hat{\mu}_c - \mathbb{E}[\hat{\mu}_c]\|_2 \leq M\sqrt{\frac{2d\log(2d/\delta)}{n_c}}.$$

For the structural bias, applying the law of total expectation conditioned on $\hat{y} = c$ yields

$$\mathbb{E}[\hat{\mu}_c] = (1 - \alpha_c)(\mu_c^* + \beta_c) + \alpha_c\mu_{\text{err}} = \mu_c^* + (1 - \alpha_c)\beta_c + \alpha_c(\mu_{\text{err}} - \mu_c^*).$$

Thus

$$\|\mathbb{E}[\hat{\mu}_c] - \mu_c^*\|_2 \leq (1 - \alpha_c)\|\beta_c\|_2 + \alpha_c\|\mu_{\text{err}} - \mu_c^*\|_2.$$

Substituting back completes the proof. □

1.2 Query Purity Bound

Lemma 1 (Vulnerability of Query Guarantee). *Assume the active learning utility function $U(x, \mu)$ is L-Lipschitz continuous w.r.t. the prototype μ. Let*

$$\Delta := U(x_a, \mu^*) - U(x_n, \mu^*) > 0$$

be the true utility margin between a structural anchor x_a and out-of-distribution noise x_n. If the empirical prototype $\hat{\mu}$ suffers from structural bias $B_c = \|\hat{\mu} - \mu^*\|_2 > \frac{\Delta}{2L}$, then there exists a feasible adversarial configuration where the theoretical guarantee to prioritize x_a over x_n is compromised, i.e.,

$$U(x_n, \hat{\mu}) > U(x_a, \hat{\mu}).$$

Proof. By the L-Lipschitz continuity of U, for any x we have

$$|U(x, \hat{\mu}) - U(x, \mu^*)| \leq L \cdot B_c.$$

To certify correct query prioritization, the minimum possible utility of x_a must exceed the maximum possible utility of x_n:

$$U(x_a, \mu^*) - L \cdot B_c > U(x_n, \mu^*) + L \cdot B_c.$$

This inequality simplifies to the condition

$$2L \cdot B_c < U(x_a, \mu^*) - U(x_n, \mu^*) = \Delta.$$

Thus if $B_c > \frac{\Delta}{2L}$, the uncertainty intervals overlap and there exists a configuration such that

$$U(x_n, \hat{\mu}) > U(x_a, \hat{\mu}).$$

□

1.3 Generalization Bound under Distribution Mismatch

Proposition 2 (Generalization Bound via Optimal Transport). *Let $h_Q \in H$ be a hypothesis trained on a queried subset S_Q of size n drawn i.i.d. from the query distribution P_Q. Let D_T be the overall target distribution. Assume covariate shift: $P(y|x)$ is identical across D_T and P_Q. Assume the loss $\ell(y', y)$ is bounded by M_ℓ and is L_ℓ-Lipschitz w.r.t. its prediction argument, and each $h \in H$ is L_h-Lipschitz w.r.t. x. Then with probability at least $1 - \delta$,*

$$R_{D_T}(h_Q) \leq \widehat{R}_Q(h_Q) + 2J_{S_Q}(H) + 3M_\ell \sqrt{\frac{\log(2/\delta)}{2n}} + L_\ell L_h \cdot W_1(D_T^X, P_Q^X), \quad (1)$$

where $\widehat{R}_Q(h_Q)$ is the empirical risk, $J_{S_Q}(H)$ is the Rademacher complexity of H, and $W_1(D_T^X, P_Q^X)$ is the 1-Wasserstein distance between the marginal input distributions.

Proof. Decompose the target risk:

$$R_{D_T}(h_Q) \leq R_{P_Q}(h_Q) + \left| \mathbb{E}_{D_T}[\ell(h_Q(x), y)] - \mathbb{E}_{P_Q}[\ell(h_Q(x), y)] \right|.$$

By standard Rademacher complexity bounds, with probability at least $1 - \delta$,

$$R_{P_Q}(h_Q) \leq \widehat{R}_Q(h_Q) + 2J_{S_Q}(H) + 3M_\ell \sqrt{\frac{\log(2/\delta)}{2n}},$$

which is a uniform convergence result based on bounded loss and Rademacher complexity.

Under covariate shift, define

$$g(x) := \mathbb{E}_{y \sim P(y|x)}[\ell(h_Q(x), y)].$$

Because ℓ is L_ℓ-Lipschitz w.r.t. predictions and h_Q is L_h-Lipschitz w.r.t. x, $g(x)$ is $(L_\ell L_h)$-Lipschitz w.r.t. x. By the Kantorovich–Rubinstein duality for the 1-Wasserstein metric, for any $(L_\ell L_h)$-Lipschitz function g,

$$\left| \mathbb{E}_{x \sim D_T^X}[g(x)] - \mathbb{E}_{x \sim P_Q^X}[g(x)] \right| \leq L_\ell L_h \cdot W_1(D_T^X, P_Q^X).$$

Substituting these two bounds gives the claimed inequality. □

Algorithm 1 Adapt-Label-Adapt (A^2L) Framework

Input: Student θ_S, Teacher θ_T, Target pool D_t, Budget B.
Output: Adapted student θ_S.

Stage I: Unsupervised Prototypical Alignment (UPA)
for $e = 1$ **to** E_{warmup} **do**
  Predict pseudo-labels $\hat{y}$ via θ_T.
  Update pseudo-centroids: $\mu_c \leftarrow \alpha\mu_c + (1-\alpha)\frac{1}{|M^{(c)}|}\sum_{j\in M^{(c)}} z_j$
  Compute loss: $\mathcal{L}_{warmup} \leftarrow \mathcal{L}_{seg} + \beta\mathcal{L}_{sca}$
  Update student: $\theta_S \leftarrow \theta_S - \eta\nabla_{\theta_S}\mathcal{L}_{warmup}$
  Update teacher: $\theta_T \leftarrow \mathrm{EMA}(\theta_S)$
end for
Stage II: Prototypical-Aware Uncertainty Herding (PAUH)
Compute pixel uncertainty $u_{i,p}$ and image uncertainty $U(X)$.
$S^* \leftarrow \emptyset$
while $|S^*| < B$ **do**
  $X^* \leftarrow \arg\max_{X\in D_t\setminus S^*} \left[U(X)\cdot V(X)\right]$
  $S^* \leftarrow S^* \cup \{X^*\}$
end while
Get ground-truth labels Y_{S^*} for active set S^*.
Stage III: Label-Guided Manifold Calibration (LGMC)
Init anchors g_c using labeled pixels Ω_c^* in S^*.
for $e = 1$ **to** E_{final} **do**
  Update anchors: $g_c \leftarrow \mathrm{EMA}(g_c, z \mid z \in \Omega_c^*)$
  Compute loss: $\mathcal{L}_{final} \leftarrow \mathcal{L}_{sup}(S^*) + \mathcal{L}_{seg}(D_t \setminus S^*) + \beta\mathcal{L}_{sca}(D_t) + \lambda\mathcal{L}_{cal}(S^*)$
  Update student: $\theta_S \leftarrow \theta_S - \eta\nabla_{\theta_S}\mathcal{L}_{final}$
  Update teacher: $\theta_T \leftarrow \mathrm{EMA}(\theta_S)$
end for

1.4 Algorithm

Our proposed A^2L framework is summarized in Algorithm 1.

2 More Visual Results

We provide further visual comparisons with other methods, as illustrated in Figure 1 and Figure 2, which display the qualitative segmentation results on the RIM-ONE-v3 and DRISHTI-GS datasets, respectively. As can be observed, our method consistently delineates smooth and highly accurate anatomical boundaries for both the optic disc and cup. It not only effectively eliminates local prediction noise but also reconstructs morphological structures that are highly consistent with the Ground Truth, continuously maintaining the best visual performance.

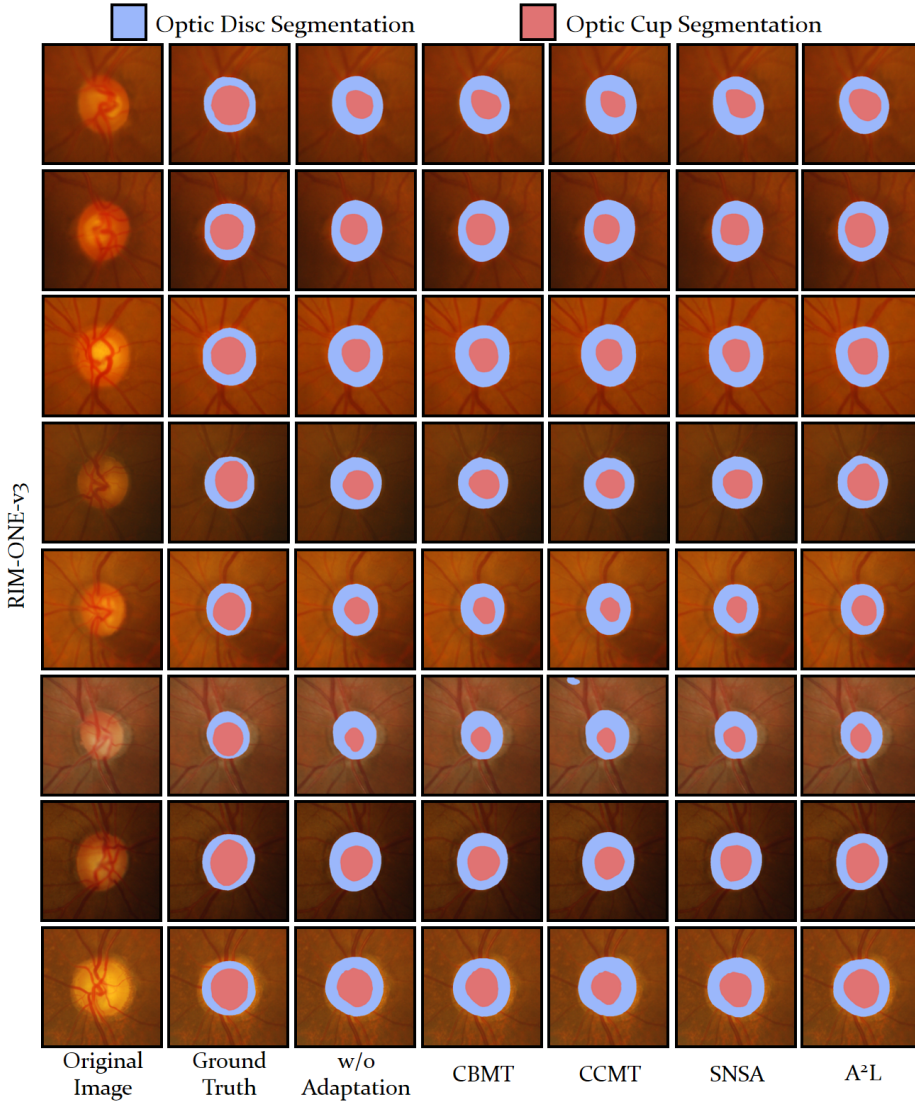

Fig. 1: Quantitative results of different methods on the RIM-ONE-v3 dataset.

Fig. 2: Quantitative results of different methods on the DRISHTI-GS datasets.