Toward Fair and Accurate Cross-Domain Medical Image Segmentation: A VLM-Driven Active Domain Adaptation Paradigm

Supplementary Material

This is the supplementary material for ***Toward Fair and Accurate Cross-Domain Medical Image Segmentation: A VLM-Driven Active Domain Adaptation Paradigm***. We present the following materials:

- Sec. 1 The more details of evaluation metrics we used.
- Sec. 2 The more experiments (about different attributes) and related ablation studies.
- Sec. 3 A brief introduction to our comparative methods.
- Sec. 4 More visualizations of VLM-attribute learning.

1. Details of Evaluation Metrics

For quantifying fairness in medical image segmentation, we adopt the same metrics used in previous studies [1, 5, 7]: Dice, IoU, and the important composite metrics, ES-Dice and ES-IoU, to assess fairness alongside performance. More detailed information will be added as follows.

For Equity-Scaled metrics, we first need to compute a performance discrepancy Δ for every sensitive attribute. This discrepancy is characterized by the cumulative difference between each demographic subgroup's metric and the overall performance. It is articulated in the following formulation:

$$\mathit{\Delta} = \sum_{\mathrm{A} \in attrs} \left| M\left(\{(\hat{y}, y)\}\right) - M\left(\{(\hat{y}, y, a) \,|a = \mathrm{A}\}\right) \right|, \tag{1}$$

where $attrs$ represent demographic groups such as {Female, Male}, {Asian, Black, White}, or {Hispanic, Non-Hispanic, Unknown}, M denotes a specific metric (e.g., Dice or IoU), and $\hat{y}$ is the ground truth. A positive Δ value implies that smaller values correspond to reduced performance disparities among demographic groups relative to the overall performance, indicating improved fairness. The Equity-Scaled metric can be formulated as follows:

$$\mathrm{ESM} = \frac{M\left(\{\hat{y}, y\}\right)}{1 + \mathit{\Delta}}. \tag{2}$$

Through the above steps, we can calculate the ES-Dice and ES-IoU metrics.

2. More Experiments

More Experiments on Other Attribute. To further validate the generalizability of our method, we conducted experiments across another sensitive attribute (Gender) and performed a comprehensive comparison with state-of-the-art Domain Adaptation (DA) and Active Domain Adaptation (ADA) approaches. All ADA methods were evaluated under the same labeling quota to ensure fair comparisons. As shown in Table 1, our method significantly outperforms all DA and ADA methods across key metrics. For instance, in rim segmentation tasks, our approach achieves an ES-Dice score of 0.785 and ES-IoU of 0.661, surpassing previous methods by a notable margin. Statistical analysis (P-value < 0.05) further confirms the significant superiority of our framework in improving performance and fairness. These results highlight our method's ability to address cross-domain challenges while ensuring equitable outcomes for diverse demographic subgroups.

Detailed Ablation Studies. Our ablation study, conducted after VLM learning, encompasses four main configurations: 1) Fair-Base: This configuration adheres to the Fair Quota principle by randomly selecting samples with an equal quota across different subgroups of the same sensitive attribute, resulting in the final selection list S. 2) Fair-Attr: Building on the Fair Quota framework, this setup exclusively employs the Attribute representative selection method to choose samples from each subgroup, culminating in the selection list S. 3) Fair-Poly: Maintaining the same quota, this configuration focuses on selecting samples with Polysemy representatives within each subgroup to form the final list S. 4) Our Fair-AP: Our approach enhances the balanced allocation strategy of Fair-Base by prioritizing the selection of samples that integrate both Attribute and Polysemy considerations, thereby augmenting representation and diversity, and resulting in the final selection list S. The ablation study results for each sensitive attribute (gender, race, and ethnicity) are displayed in Table 2, Table 3, and Table 4, respectively.

Gender Attribute: In Table 2, the standalone application of Fair-Attr or Fair-Poly demonstrates potential for enhancing certain metrics, while the observed improvements exhibit variability, suggesting limited consistency in performance across all evaluation criteria. However, it is evident that our Fair-AP comprehensively considers both attribute representative and polysemy representative, ensuring superior performance and fairness.

Race Attribute: The findings from the ablation study on racial attributes indicate that if the distribution of various races is imbalanced, both Fair-Attr and Fair-Poly show improvements in some performance measures (like ES-IoU and IoU for Cup), as illustrated in Table 3. Our Fair-AP

Table 1. Cup and rim segmentation performance on the FairDomain-Segmentation benchmark using different DA and ADA methods with **Gender** as the demographic attribute. The $*$ denotes p-value < 0.05 in all paired t-test, indicating statistically significant differences.

		Method	Venue	Overall ES-Dice↑	Overall Dice↑	Overall ES-IoU↑	Overall IoU↑	Male Dice↑	Female Dice↑	Male IoU↑	Female IoU↑
	Cup	Baseline (Source)	-	0.885	0.888	0.806	0.808	0.886	0.889	0.807	0.810
	Cup	Baseline (Target)	-	0.688	0.700	0.535	0.555	0.693	0.711	0.557	0.574
	Rim	Baseline (Source)	-	0.854	0.861	0.753	0.762	0.864	0.856	0.767	0.755
	Rim	Baseline (Target)	-	0.485	0.495	0.336	0.342	0.486	0.507	0.334	0.353
DA	Cup	PixMatch [6]	CVPR'21	0.768	0.775	0.650	0.660	0.772	0.769	0.645	0.660
DA	Cup	DAFormer [4]	CVPR'22	0.781	0.785	0.676	0.680	0.783	0.789	0.678	0.684
DA	Cup	DAFormer-FIA [7]	ECCV'24	0.802	0.810	0.692	0.700	0.806	0.816	0.695	0.706
DA	Rim	PixMatch [6]	CVPR'21	0.660	0.673	0.519	0.523	0.669	0.688	0.519	0.528
DA	Rim	DAFormer [4]	CVPR'22	0.344	0.345	0.212	0.213	0.344	0.347	0.212	0.214
DA	Rim	DAFormer-FIA [7]	ECCV'21	0.528	0.531	0.367	0.369	0.533	0.528	0.372	0.366
ADA	Cup	Random	-	0.828	0.834	0.729	0.734	0.838	0.831	0.738	0.731
ADA	Cup	Entropy [10]	CVPR'19	0.819	0.823	0.717	0.721	0.826	0.821	0.724	0.719
ADA	Cup	MHPL [2]	CVPR'23	0.829	0.834	0.728	0.733	0.838	0.832	0.738	0.730
ADA	Cup	DML-Core [9]	NeurIPS'24	0.831	0.839	0.733	0.740	0.844	0.835	0.746	0.736
ADA	Cup	Detective [11]	CVPR'24	0.831	0.837	0.733	0.740	0.841	0.834	0.745	0.736
ADA	Cup	STDR [3]	IEEE TMI'24	0.831	0.837	0.731	0.738	0.842	0.834	0.743	0.734
ADA	Cup	Our FairAP	-	**0.839***	**0.843***	**0.742***	**0.746***	**0.845**	**0.841***	**0.749***	**0.744***
ADA	Rim	Random	-	0.778	0.780	0.652	0.654	0.779	0.782	0.652	0.655
ADA	Rim	Entropy [10]	CVPR'19	0.768	0.772	0.639	0.643	0.769	0.774	0.640	0.645
ADA	Rim	MHPL [2]	CVPR'23	0.779	0.782	0.654	0.656	0.780	0.784	0.654	0.658
ADA	Rim	DML-Core [9]	NeurIPS'24	0.782	0.783	0.658	0.658	0.782	0.784	0.658	0.658
ADA	Rim	Detective [11]	CVPR'24	0.781	0.784	0.657	0.659	0.782	0.786	0.658	0.661
ADA	Rim	STDR [3]	IEEE TMI'24	0.777	0.781	0.652	0.656	0.778	0.783	0.653	0.658
ADA	Rim	**Our FairAP**	-	**0.785***	**0.787***	**0.661***	**0.663***	**0.786***	**0.789***	**0.661***	**0.664***

Table 2. Ablation study of Optic cup and rim segmentation performance on the FairDomain-Segmentation dataset with **Gender** as the demographic attribute.

		Method	Overall ES-Dice↑	Overall Dice↑	Overall ES-IoU↑	Overall IoU↑	Male Dice↑	Female Dice↑	Male IoU↑	Female IoU↑
ADA	Cup	Fair-Base	0.832	0.836	0.734	0.737	0.838	0.834	0.739	0.735
ADA	Cup	Fair-Attr	0.831	0.836	0.732	0.738	0.841	0.833	0.743	0.735
ADA	Cup	Fair-Poly	0.834	0.841	0.735	0.743	**0.846**	0.837	**0.749**	0.739
ADA	Cup	**Ours Fair-AP**	**0.839**	**0.843**	**0.742**	**0.746**	0.845	**0.841**	**0.749**	**0.744**
ADA	Rim	Fair-Base	0.777	0.783	0.652	0.658	0.778	0.787	0.653	0.662
ADA	Rim	Fair-Attr	0.774	0.780	0.649	0.655	0.776	0.783	0.650	0.659
ADA	Rim	Fair-Poly	0.781	0.784	0.657	0.659	0.782	0.785	0.657	0.660
ADA	Rim	**Ours Fair-AP**	**0.785**	**0.787**	**0.661**	**0.663**	**0.786**	**0.789**	**0.661**	**0.664**

not only performs well in segmentation outcomes but also effectively ensures algorithmic fairness.

Ethnicity Attribute: In Table 4, a similar trend is observed: our Fair-AP model preserves fairness robustness while achieving competitive performance gains in segmentation tasks, even under imbalanced distribution scenarios.

3. Comparison Methods

We conducted a series comparisons of Fair-AP with leading DA techniques, including Pixmatch [6], DAFormer [4], and DAFormer-FIA [7], as well as ADA approaches like Entropy [10], MHPL [2], DML-Core [9], Detective [11], and STDR [3]. The details of these methods are summarized as follows:

Pixmatch [6]: Pixmatch introduces a novel unsupervised domain adaptation framework that enhances model performance on target domains by ensuring consistency in predictions under small input perturbations.

DAFormer [4]: DAFormer introduces a Transformer-based encoder-decoder architecture with three key training strategies—Rare Class Sampling, Thing-Class ImageNet Feature Distance, and learning rate warmup—to stabilize training and mitigate overfitting in unsupervised domain adaptation (UDA) for semantic segmentation.

DAFormer-FIA [7]: DAFormer-FIA introduces a novel Fair Identity Attention (FIA) module and the first fairness-

Table 3. Ablation study of Optic cup and rim segmentation performance on the FairDomain-Segmentation dataset with **Race** as the demographic attribute.

		Method	Overall ES-Dice↑	Overall Dice↑	Overall ES-IoU↑	Overall IoU↑	Asian Dice↑	Black Dice↑	White Dice↑	Asian IoU↑	Black IoU↑	White IoU↑
ADA	Cup	Fair-Base	0.816	0.836	0.714	0.736	0.815	0.837	0.838	0.712	0.742	0.738
		Fair-Attr	0.810	0.836	0.719	0.739	0.818	0.825	0.840	0.721	0.733	0.742
		Fair-Poly	0.819	0.840	0.721	0.742	0.818	0.839	0.842	0.718	0.744	0.744
		Ours Fair-AP	**0.828**	**0.843**	**0.731**	**0.747**	**0.827**	**0.843**	**0.845**	**0.730**	**0.751**	**0.748**
	Rim	Fair-Base	0.701	0.786	0.581	0.661	0.739	0.729	0.803	0.608	0.596	0.680
		Fair-Attr	0.690	0.778	0.572	0.651	0.738	0.709	0.796	0.610	0.574	0.672
		Fair-Poly	0.699	0.789	0.579	0.664	0.740	0.726	**0.807**	0.611	0.592	0.685
		Ours Fair-AP	**0.711**	**0.791**	**0.592**	**0.667**	**0.750**	**0.735**	**0.807**	**0.624**	**0.602**	**0.686**

Table 4. Ablation study of Optic cup and rim segmentation performance on the FairDomain-Segmentation dataset with **Ethnicity** as the demographic attribute.

		Method	Overall ES-Dice↑	Overall Dice↑	Overall ES-IoU↑	Overall IoU↑	Hispanic Dice↑	Non-Hispanic Dice↑	Hispanic IoU↑	Non-Hispanic IoU↑
ADA	Cup	Fair-Base	0.822	0.838	0.719	0.738	**0.854**	0.836	**0.762**	0.735
		Fair-Attr	0.822	0.833	0.717	0.732	0.845	0.832	0.751	0.730
		Fair-Poly	0.822	0.835	0.717	0.735	0.849	0.833	0.758	0.733
		Ours Fair-AP	**0.836**	**0.840**	**0.730**	**0.743**	0.844	**0.839**	0.759	**0.741**
	Rim	Fair-Base	0.778	0.786	0.656	0.661	0.794	0.784	0.667	0.659
		Fair-Attr	0.781	0.784	0.657	0.659	0.786	0.782	0.658	0.658
		Fair-Poly	0.777	0.786	0.655	0.663	**0.797**	0.784	0.672	0.661
		Ours Fair-AP	**0.784**	**0.790**	**0.663**	**0.668**	0.796	**0.788**	**0.673**	**0.666**

focused paired imaging dataset to enhance algorithmic fairness under domain shifts in medical AI, significantly improving performance across demographics in both domain adaptation (DA) and domain generalization (DG) tasks.

Entropy [10]:The AdvEnt method [10] is employed to calculate the prediction map entropy for each sample within the target domain, and those samples with the highest entropy are selected for manual annotation.

MHPL [2]: MHPL introduces a novel active source-free domain adaptation approach by identifying and exploiting ”minimum happy points” through tailored selection strategies and a neighbor focal loss, achieving significant performance gains with minimal labeling effort.

DML-Core [9]:DML-Core introduces a slice-based active learning method integrating deep metric learning with Coreset, significantly reducing annotation costs for 3D medical segmentation while achieving high performance under low annotation budgets.

Detective [11]:Detective introduces a dynamic domain adaptation model with evidential uncertainty valuation and contextual diversity enhancement, effectively selecting informative target samples by jointly evaluating domain shifts and prediction confidence.

STDR [3]: STDR improves gross tumor volume segmentation for nasopharyngeal carcinoma by employing a dual reference strategy to select and annotate representative target-domain samples, enabling effective source-free domain adaptation while ensuring data privacy and minimizing the annotation effort.

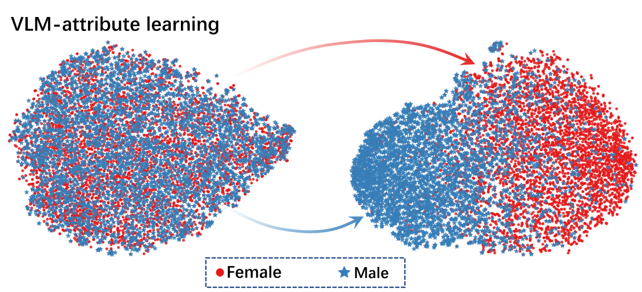

Figure 1. Visualization of gender feature distribution before and after feature alignment via VLM-attribute learning.

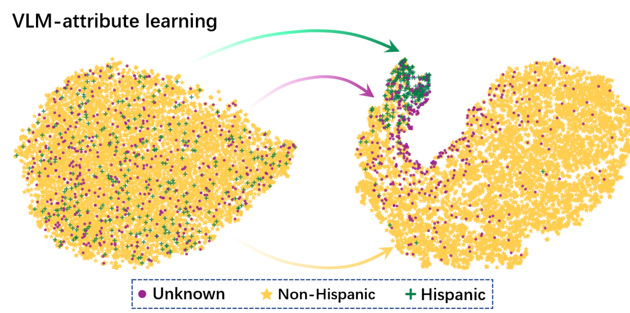

Figure 2. Visualization of ethnic feature distribution before and after feature alignment via VLM-attribute learning.

4. More Visualizations

The t-SNE visualizations [8] of sample distributions before and after feature alignment via VLM-attribute learning for gender and ethnic attributes are depicted in Fig. 1 and Fig. 2. Our VLM-attribute learning results in a more concentrated distribution of features across different subgroups, thus improving the distinguishability of features associated with various attributes. In Fig. 1, we note that male and female samples were effectively distinguished after VLM-attribute learning. We also notice a balanced representation of male and female samples, which aids in the subsequent active selection strategy. Conversely, in Fig. 2, the ethnic proportions are uneven, and the distribution of Unknown samples is irregular, causing significant degradation in the ablation study results when either representativeness or diversity is considered in isolation. Our proposed Fair-AP method achieves success in both segmentation performance and fairness, even when subgroup proportions are imbalanced and sample distributions are irregular.

References

[1] Fairseg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. In *ICLR*. 1

[2] Mhpl: Minimum happy points learning for active source free domain adaptation. In *CVPR*, 2023. 2, 3

[3] Dual-reference source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals. *IEEE Transactions on Medical Imaging*, 2024. 2, 3

[4] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9924–9935, 2022. 2

[5] Yan Luo, Yu Tian, Min Shi, Louis R Pasquale, Lucy Q Shen, Nazlee Zebardast, Tobias Elze, and Mengyu Wang. Harvard glaucoma fairness: a retinal nerve disease dataset for fairness learning and fair identity normalization. *IEEE Transactions on Medical Imaging*, 2024. 1

[6] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12435–12445, 2021. 2

[7] Yu Tian, Congcong Wen, Min Shi, Muhammad Muneeb Afzal, Hao Huang, Muhammad Osama Khan, Yan Luo, Yi Fang, and Mengyu Wang. Fairdomain: Achieving fairness in cross-domain medical image segmentation and classification. In *European Conference on Computer Vision*, pages 251–271. Springer, 2024. 1, 2

[8] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 3

[9] Arvind Vepa, Zukang Yang, Andrew Choi, Jungseock Joo, Fabien Scalzo, and Yizhou Sun. Integrating deep metric learning with coreset for active learning in 3d segmentation. *Advances in Neural Information Processing Systems*, 37:71643–71671, 2024. 2, 3

[10] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2517–2526, 2019. 2, 3

[11] Wenqiao Zhang, Zheqi Lv, Hao Zhou, Jia-Wei Liu, Juncheng Li, Mengze Li, Yunfei Li, Dongping Zhang, Yueting Zhuang, and Siliang Tang. Revisiting the domain shift and sample uncertainty in multi-source active domain transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16751–16761, 2024. 2, 3