A Simple yet Effective Adaptive Inter-organ Contrastive Learning Framework for Unsupervised Domain Adaptation

Yiyou Sun[1] SUNYIYOU00@GMAIL.COM
Zheyao Gao[2] ZHEYAOGAO@CUHK.EDU.HK
Xiaogen Zhou[2] XIAOGENZHOU@CUHK.EDU.HK
Qi Dou[2] QIDOU@CUHK.EDU.HK
Winnie Chiu Wing Chu[1] WINNIECHU@CUHK.EDU.HK

[1] *Department of Imaging and Interventional Radiology, CU Lab of AI in Radiology, The Chinese University of Hong Kong, Hong Kong, SAR, China*

[2] *Department of Computer Science and Engineering, Institute of Medical Intelligence, The Chinese University of Hong Kong, SAR, China*

Editors: Under Review for MIDL 2026

Abstract

Recent unsupervised domain adaptation methods have shown promise in medical image segmentation dealing with cross-modal data, while their reliance on adversarial learning for global feature alignment often compromises fine-grained semantic consistency and degrades source domain performance. Although most contrastive learning (CL) methods present excellent ability for semantic representation, they extract features with a pixel-to-pixel binary thresholding strategy, which might lead to sparse feature space and bad spatial alignment. Thus, we present a CL method leveraging pseudo-label for organ-wise feature patch sampling. The proposed simple yet effective adaptive inter-organ contrastive learning framework incorporates a modality-adaptive encoder to handle multi-modal variations while maintaining shared feature representations. Furthermore, the model is trained by combined supervised consistency learning and unsupervised pseudo-label guided contrastive learning, promoting a more discriminative and compact shared latent space. Extensive experiments and ablation studies on an orbital and a cardiac dataset reveal the effectiveness of each component and significant advancement in segmentation results compared to other reference methods.

Keywords: Unsupervised Domain Adaptation, Multi-organ Segmentation

1. Introduction

Medical image segmentation plays a pivotal role in clinical practice, particularly when leveraging complementary information from multi-sequence magnetic resonance imaging (MRI) data, such as diagnosis of thyroid-associated orbitopathy (TAO) and study of presence, location, and extent of myocardial infarction (MI). However, comprehensive anatomical annotations across different imaging modalities remain scarce due to the resource-intensive nature of the labeling process. Furthermore, the integration of heterogeneous modalities presents inherent challenges due to distribution misalignment stemming from varying imaging protocols and patient movement during acquisition.

Unsupervised domain adaptation (UDA) has become a cornerstone of cross-domain medical image segmentation, transferring knowledge from a labeled source to an unlabeled target domain (Lee et al., 2021; Shin et al., 2023; Xian et al., 2023; Zhao et al., 2023). Early UDA methods emphasized either adversarial alignment in feature/output space or image-to-image translation. Adversarial alignment encourages global distribution matching but can blur semantic boundaries and underfit minority structures(Hoffman et al., 2016; Chen et al., 2017; Vu et al.). Image translation (Park et al., 2020; Han et al., 2021) may partly reduce appearance gaps but risks altering anatomy and depends on cycle or structural constraints that are hard to satisfy in practice. Pseudo-label learning (PLL) (Zhao et al., 2023; Shin et al., 2023) has become the cornerstone of modern UDA in medical segmentation. By converting model predictions into supervision, teacher–student frameworks leverage uncertainty estimation and consistency checks to improve label quality. Despite these measures, inherent noise persists; hard thresholds lose information, while class imbalance and low-confidence predictions can introduce significant bias.

Recent progress in representation learning catalyzed by contrastive learning (CL) offers a promising alternative paradigm that addresses these limitations. Unlike global distribution alignment in UDA, CL operates at pixel-to-pixel (P2P) level to learn rich feature representations by maximizing the contrast between defined similar and dissimilar samples regarding designated anchor samples in a projected embedding space. CL has demonstrated remarkable capability in matching the fine-grained feature representations across modalities in segmentation tasks (Gu et al., 2024; Wang et al.; Zhang et al., 2023). However, existing CL-based approaches for medical image segmentation still face challenges. First, conventional P2P or centroid-to-pixel (C2P) contrastive learning strategies often suffer from high memory requirements or limited feature utilization. Additionally, the binary nature of hard threshold-based sample selection in existing methods can lead to information loss and suboptimal feature space organization.

Consequently, bridging these gaps requires a unified framework that synergizes the semantic guidance of PLL with the representation power of CL, yet avoids their respective pitfalls. To circumvent the computational bottlenecks of dense pixel-wise comparisons, a strategy is required to selectively sample informative anatomical regions—specifically focusing on organ-relevant features—rather than processing redundant background noise. Furthermore, instead of discarding uncertain predictions via rigid thresholding, it is crucial to exploit the intrinsic structure of pseudo-labels to dynamically regularize the latent space. By treating pseudo-labels as semantic cues for contrastive alignment rather than just noisy supervision, we can enforce tighter class-wise clustering across domains. To address these limitations, our contributions are summarized as follows: (1) We develop an organ-wise pseudo-label guided patch sampling (PGPS) strategy in cross-modality feature alignment (CMFA) module for CL to guarantee optimal feature discrepancy and efficient feature representation. (2) We implement a pseudo-label-guided CL (PGCL) regularizer that complements the supervised consistency learning (SCL), effectively pushing apart cross-modality features from different classes while pulling together them from the same class in the latent embedding space.

2. Methods

The proposed adaptive inter-organ contrastive learning (AICL) framework processes interleaved inputs I from source and target domains m_1, m_2 within a unified batch B to

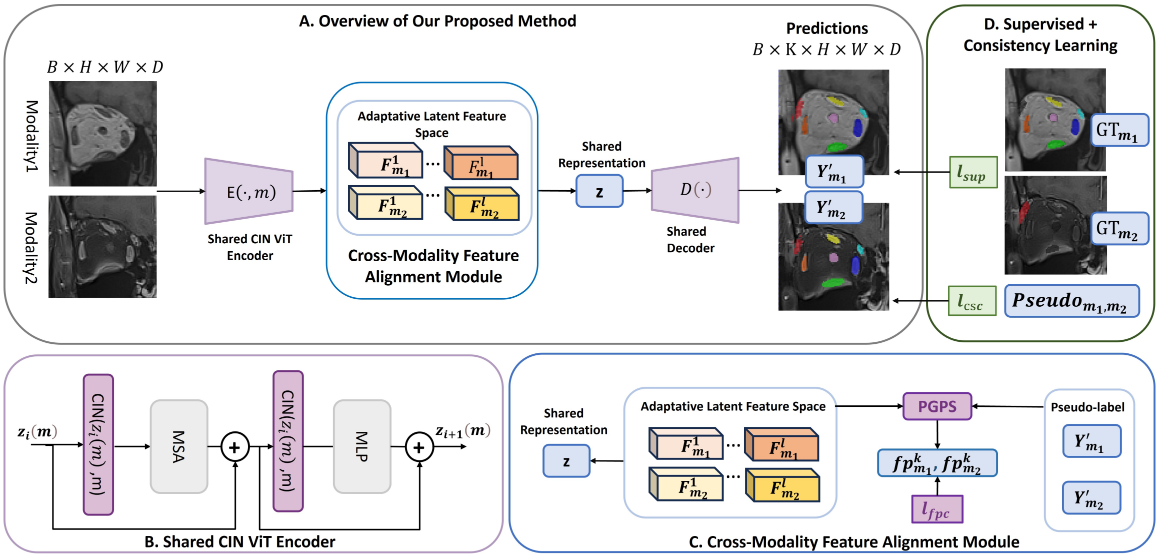
A. Overview of Our Proposed Method
$B \times H \times W \times D$
Modality1
Modality2
$E(\cdot, m)$
Shared CIN ViT Encoder
Adaptative Latent Feature Space
$F^1_{m_1}$ ... $F^l_{m_1}$
$F^1_{m_2}$ ... $F^l_{m_2}$
Cross-Modality Feature Alignment Module
Shared Representation
z
$D(\cdot)$
Shared Decoder
Predictions
$B \times K \times H \times W \times D$
Y'_{m_1}
Y'_{m_2}
D. Supervised + Consistency Learning
GT_{m_1}
GT_{m_2}
l_{sup}
l_{csc}
$Pseudo_{m_1,m_2}$
$z_i(m)$
$CIN(z_i(m), m)$
MSA
MLP
$z_{i+1}(m)$
B. Shared CIN ViT Encoder
Shared Representation
z
Adaptative Latent Feature Space
PGPS
Pseudo-label
$fp^k_{m_1}, fp^k_{m_2}$
l_{fpc}
C. Cross-Modality Feature Alignment Module

Figure 1: Flowchart of our proposed method. A. The overall pipeline takes paired multi-modal images (m_1, m_2) as input. It utilizes a shared encoder depicted in B to extract features, followed by CMFA (C) to generate a modality-invariant shared representation z, which is then decoded into segmentation predictions Y'.

generate comprehensive segmentation predictions. As illustrated in Fig. 1, our architecture first employs a shared conditional instance normalization (CIN) vision transformer (ViT) encoder to dynamically adapt feature distributions across diverse imaging modalities. Subsequently, CMFA systematically regularizes the shared latent feature space through PGCL, ensuring anatomical consistency between modalities. Finally, the framework integrates SCL to enforce consistency between predictions and corresponding available ground truth.

2.1. Modality-Adaptive Encoder

The shared-weight encoder $E(\cdot, m)$ integrates a CIN mechanism (Dumoulin et al., 2016; Bastico et al., 2023), parameterized by modality-specific scaling and shifting parameters γ_m and β_m for each input modality I_{m_1}, I_{m_2}. The CIN module projects inputs into a unified yet modality-adaptive latent feature space $F^l_m \in \mathbb{R}^{C \times L_h \times L_w \times L_d}$ by independently normalizing instance-specific statistics across source and target domains. In l layer of latent feature space, feature maps z across modalities m_1, m_2 have C channels and spatial dimensions of L_h, L_w, L_d as height, width, and depth. The CIN ViT encoder is defined as:

$$CIN(z, m) = \gamma_m(\frac{z - \mu(z)}{\sigma(z)}) + \beta_m, \tag{1}$$

where $\mu(z)$ and $\sigma(z)$ represent channel-wise mean and standard deviation computed per instance within each batch.

Modality-specific learnable parameters γ_m and β_m are trained to decouple sensing-specific statistics (e.g., intensity, noise, and contrast) from modality-consistent semantic content, enabling a single shared encoder to generalize across modalities. CIN provides lightweight adaptation while preserving modality fidelity, thereby facilitating efficient training without compromising cross-modality alignment.

2.2. CMFA Via PGCL

2.2.1. PGPS in Latent Feature Space

In our framework, inputs I_{m_1}, I_{m_2} pass through a shared-weight encoder $E(\cdot, m)$ and a decoder D, parameterized by θ_{CIN}, θ_D, yielding pseudo-labels $Y_m^{'} = D(E(I_m)) \in \mathbb{R}^{K \times H \times W \times D}$. For cross-modality feature alignment, the pseudo-labels are downsampled to feature-layer resolution $Y_{f_m}^{'} \in \mathbb{R}^{K \times L_h \times L_w \times L_d}$, where the organ class index is $k = 1, 2, ..., K$ and voxel indexes are $(h, w, d), h \in [0, H), w \in [0, W), d \in [0, D)$. Leveraging the compactness of anatomical structures in medical images, cross-modality feature patch embeddings fp_m^l are sampled and assigned an organ label by organ-wise masks $MS_m \in \mathbb{R}^{K \times L_h \times L_w \times L_d}$ as bounding cubes centered at the mean coordinates of valid voxels from $Y_{f_m}^{'}(k)$ for each class k and each modality m, and the formula is:

$$oh_m^k, ow_m^k, od_m^k = \overline{argwhere(Y_{f_m}^{'}(k))}, \tag{2}$$

$$MS_m(k) = \begin{cases} 1, & \text{if } \begin{array}{c} oh_m^k - PS_h \le h_m^k < oh_m^k + PS_h, \\ ow_m^k - PS_w \le w_m^k < ow_m^k + PS_w, \\ od_m^k - PS_d \le d_m^k < od_m^k + PS_d, \end{array} \\ \\ 0, & \text{Otherwise} \end{cases} \tag{3}$$

where (PS_h, PS_w, PS_d) denotes the patch size of bounding cubes, and $fp_m^l(k) = EB(F_m^l \circ MS_m(k)), fp_m^l \in \mathbb{R}^{K \times (PS_h \cdot PS_w \cdot PS_d)}$.

In our case, the patch size for sampling is a hyperparameter empirically set according to organ size, balancing local details and global context. By leveraging class-aware embeddings, PGPS enforces that patches from pseudo-label guided sampling are spatially coherent, even when original anatomical structure are misaligned. This mechanism decouples the dependency on precise anatomical correspondence and preserves the number of features exposed to the network, preparing robust and rich feature extraction for cross-modality contrastive learning.

2.2.2. PGCL

Inspired by the PatchNCE (Park et al., 2020) that enforces local feature consistency by contrasting positive or negative image patch pairs, we propose PGCL in Fig.2 to extend this concept to CMFA. PGCL leverages PGPS to define organ-wise semantic correspondence of latent feature patch embeddings $fp_{m_1}^k, fp_{m_2}^k$ when l is settled as the last layer of $E(\cdot, m)$, and we compute the feature patch contrastive (FPC) loss as:

$$l_{fpc} = -\mathbb{E}_{z \sim F} \sum_{k=1}^{K} log \left[\frac{\exp(\phi(fp_{m_1}^k \cdot fp_{m_2}^{+})/\tau)}{\exp(\phi(fp_{m_1}^k \cdot fp_{m_2}^{+})/\tau) + \sum_{n=1}^{N-1} \exp(\phi(fp_{m_1}^k \cdot fp_{m_2}^{-})/\tau)} \right], \tag{4}$$

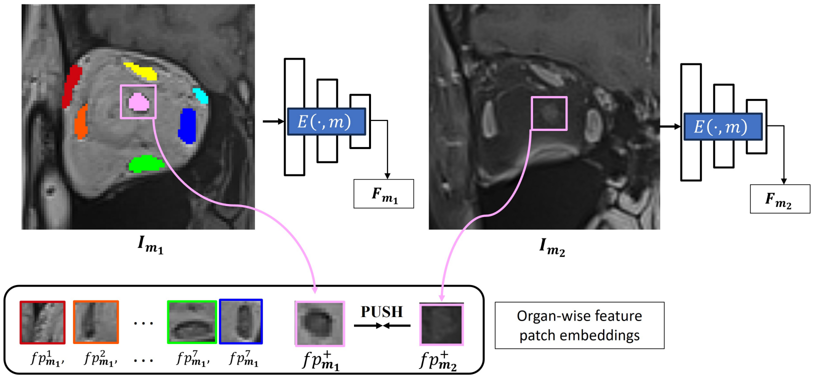
$E(\cdot, m)$
F_{m_1}
I_{m_1}
$E(\cdot, m)$
F_{m_2}
I_{m_2}
$fp^1_{m_1}$, $fp^2_{m_1}$, … $fp^7_{m_1}$, $fp^7_{m_1}$
$fp^+_{m_1}$
PUSH
$fp^+_{m_2}$
Organ-wise feature patch embeddings

Figure 2: Illustration of PGCL. Pseudo labels are downsampled into the latent feature space for organ-wise feature embeddings fp^k_m generation. The proposed method minimizes the distance ("PUSH") between positive pairs (same k), effectively facilitating cross-modality feature alignment.

where $\phi(\cdot)$ denotes cosine similarity, and τ is a temperature hyperparameter. Critically, positive pairs are feature patches sharing the same semantic class $Y^{'}_{i,m} = Y^{'}_{j,m}$, while negative pairs come from different classes $Y^{'}_{i,m} \neq Y^{'}_{j,m}$ irrespective of modality.

PGCL enforces that same-class features are close, whether from different or same modalities, while different classes are separated at the organ-wise semantic feature patch level.

2.3. Overall Learning Process

In the supervised consistency learning (SCL) phase, l_{sup} calculates the focal dice (Jadon, 2020; Lin et al., 2017) loss $l_{sup} = l_{dice} + l_{focal}$ between the predictions $Y^{'}$ and ground truth GT from annotated classes. The contrastive structural consistency loss l_{csc} calculated the InfoNCE (van den Oord DeepMind et al.) loss using class-wise predictions from different modalities $Y^{'}_{m_1}(k), Y^{'}_{m_2}(k)$ for sampling positive and negative pairs. l_{csc} is constructed as:

$$l_{csc} = -\mathbb{E}_{y^+_{m_1} \sim Y^{'}_{m_1}} log \frac{Sim(y^+_{m_1}, y^+_{m2})}{\sum_{y^k_{m_2} \in Y^{'}_{m_2}} Sim(y^+_{m_1}, y^k_{m2})}, \tag{5}$$

Overall, the total loss of our framework is:

$$l_{total} = l_{sup} + \lambda_{csc} l_{csc} + \lambda_{fpc} l_{fpc}, \tag{6}$$

where the $\lambda_{csc}, \lambda_{fpc}$ are trade-off parameters scaling the importance of each loss component.

3. Experiments

3.1. Datasets

TAO Dataset. The in-house TAO dataset comprises 3D orbital MRI scans from 100 subjects, acquired through two complementary protocols: pre-contrast T1-weighted (T1) and post-contrast T1-weighted (T1c) imaging. The dataset contains full annotations for 20 cases, including extraocular muscle (EOM) groups, optic nerves (ON), and lacrimal glands (LG), partitioned into validation (20%) and test sets (80%). The remaining 80 training cases are partially annotated with EOM and ON on T1, and LG on T1c. A standardized preprocessing pipeline composed of image registration, cropping inputs to $96 \times 96 \times 32$ patches centered in regions with dense anatomical orbital structures, and normalizing intensity distribution to range $[0, 1]$. Subsequently, random 3D rotations and axis-aligned flips constitute data augmentation.

MS-CMRSeg Dataset. The publicly available MS-CMRSeg dataset (Zhuang, 2016, 2019) encompasses 45 paired cardiac imaging data. The protocol includes balanced-steady state free precession (bSSFP) cine sequences, serving as the source modality, and late gadolinium enhancement (LGE) sequences as the target modality. We only employ expert-validated annotations delineating three cardiac structures from bSSFP: the left ventricular cavity (LV), right ventricular cavity (RV), and left ventricular myocardium (Myo) across all cases. The dataset is randomly partitioned into 35 LGE/bSSFP for model training and 10 pairs for testing. Similar data processing steps as TAO are applied, and inputs are cropped into $480 \times 480 \times 16$.

3.2. Implementation Details

All experiments were conducted using Python 3.10 and PyTorch 1.13.1 on NVIDIA A100 GPU with CUDA 11.7. We adopted SwinUNETR (Hatamizadeh et al., 2022) as the backbone architecture for TAO and UNET (Siddique et al., 2020; Tarvainen and Valpola) for MS-CMRSeg. SwinUNETR was configured with feature size of $fs = 48$, encoder layer depth of $L = 4$ and hidden size of $K = 768$. To deal with a small-scale dataset, UNET comprises five resolution levels with feature widths $[16, 64, 128, 256, 512]$. In the encoder, each level begins with residual unit for downsampling with strides of $[1, 2, 2, 2, 1]$. After iterative tuning of the hyperparameters from empirically initiated sets, we identified the optimal training protocol utilizing Adam optimizer, learning rate is set as $1e - 4$ with weight delay $1e - 5$, $\lambda_{csc} = 1$, $\lambda_{fpc} = 0.5$, patch size $ps = [16, 16, 32]$ for TAO and $ps = [96, 96, 16]$ for MS-CMRSeg. Lastly, the overlap ratio is 0.5 for sliding window inference.

3.3. Effectiveness of Our Methods

We evaluated the effectiveness of our proposed method by measuring multi-oragn segmentation performance in the Dice Similarity Score (Dice) and Haudorff Distance 95 (HD95).

Table 1 presents quantitative comparisons of our proposed method against related unsupervised domain adaptation methods on the TAO dataset, adapting from T1 to T1c. REG compares the source domain labels transformed by a rigid matrix calculated from image registration across modalities and the target domain label. REG depicts the difficulty of resolving the domain gap between inputs. CIN-seg (Bastico et al., 2023) utilizes registered pseudo labels as supervision for cross-modal fusion, suffering an averaged 9.6% Dice decline over organs missing annotations compared to GT-supervised organs. Our method outperforms other methods by a large margin, especially in the missing label case, both in dice and HD95. DCLGAN(Han et al., 2021) leverages unsupervised contrastive learning for cross-modal image translation, enabling supervised learning on synthetic target data and

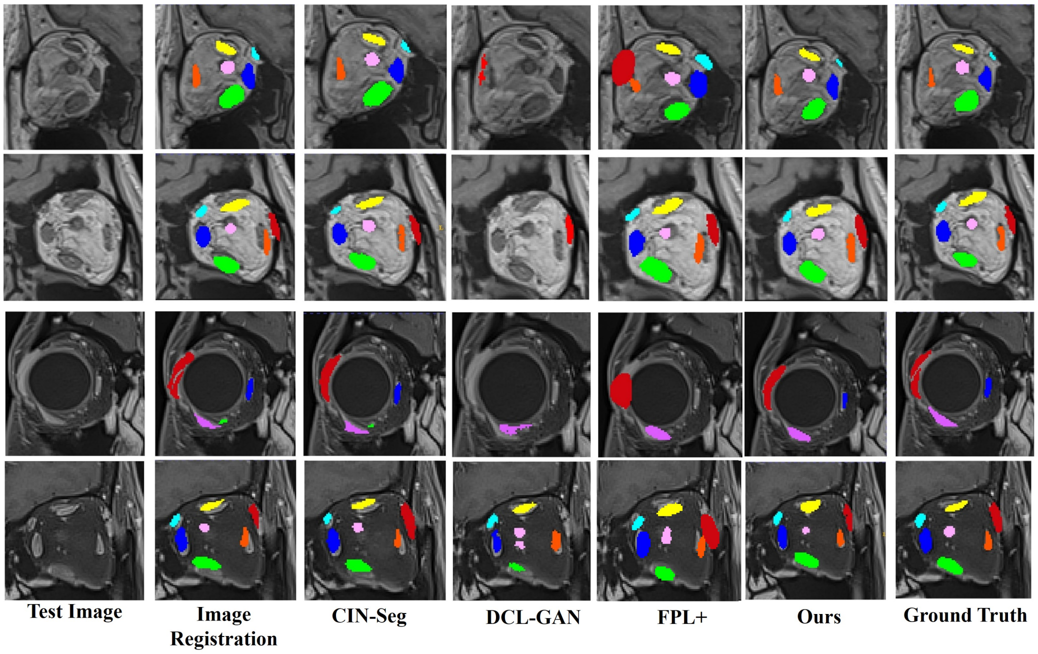

Figure 3: The qualitative comparisons of segmentation on the TAO dataset.

source domain labels. However, anatomical distortions introduced during translation degrade performance (e.g., 28.9% Dice drop for LG in T1), as synthetic images often misalign with ground truth structures. FPL+ (Wu et al., 2024) employs dual-domain pseudo-label generation with noise filtering. While effective in ideal scenarios, its reliance on heuristic thresholds amplifies error propagation under severe annotation sparsity. This framework performs better in T1c than T1 modality because it relies on synthetic T1c images with EOM and ON ground truth, and vice versa.

3.4. Ablation Study

3.4.1. TAO

We conducted a systematic ablation study on the TAO dataset, focusing on the bidirectional domain adaptation between T1 and T1c modalities, to evaluate the contributions of the proposed core components, CMFA and SCL. As summarized in Tab. 2, these components are denoted as l_{fpc}, and l_{csc}, respectively. Our analysis began with a baseline model devoid of components, trained solely on cross-modal inputs with standard supervised dicefocal loss, achieving a modest average Dice of 69.05%, and a relatively high average HD95 of 9.57 mm. A closer inspection reveals significant performance bottlenecks in segmenting small, irregular structures and low-contrast boundaries of IOM, showing a particularly poor Dice score of 47.38%. This underscores the challenge of learning robust representations for fine-grained orbital structures under domain shift. The integration of both components achieves superior performance, validating the synergistic effect of our dual-consistency framework. The full model reaches an average Dice of 76.38% (a 7.33% improvement over the baseline)

Methods	LR	IR	MR	SR	SOM	ON	IOM	LG	**Avg.**
T1 Dice[%] ↑									
REG	–	–	–	–	–	–	–	61.31	61.31
CIN-seg	78.37	88.54	88.59	78.69	80.51	83.20	67.31	65.75	79.08
DCLGAN	–	–	–	–	–	–	–	55.40	55.40
FPL+	55.12	71.64	70.13	66.07	56.95	49.69	16.20	60.50	55.79
Ours	**82.60**	**90.31**	**89.74**	**80.76**	**82.89**	**85.57**	**72.08**	**68.24**	**81.52**
T1 HD95[mm] ↓									
REG	–	–	–	–	–	–	–	9.51	9.51
CIN-seg	6.13	3.41	3.48	8.70	9.54	7.16	6.47	**7.47**	6.54
DCLGAN	–	–	–	–	–	–	–	9.80	9.80
FPL+	8.31	10.30	10.20	9.06	8.12	8.31	12.25	10.05	9.57
Ours	**5.53**	**3.83**	**3.11**	**4.59**	**3.58**	**4.22**	**5.25**	7.71	**4.58**
T1c Dice[%] ↑									
REG	58.26	73.68	69.54	57.94	56.98	59.36	44.15	-	59.98
CIN-seg	70.50	78.21	79.40	68.38	70.06	68.68	51.06	77.12	70.43
DCLGAN	47.80	63.48	70.18	59.03	65.65	49.23	35.07	–	55.78
FPL+	50.16	69.78	77.94	67.15	73.48	58.88	38.45	71.51	63.42
Ours	**80.67**	**81.90**	**85.63**	**74.97**	**81.84**	**74.48**	**62.79**	**78.32**	**77.39**
T1c HD95[mm] ↓									
REG	6.83	5.05	7.02	5.38	7.00	5.23	9.25	-	6.53
CIN-seg	6.49	5.64	5.29	8.23	10.47	4.44	10.29	8.42	7.41
DCLGAN	20.13	13.32	9.93	10.50	10.03	20.94	14.12	–	14.14
FPL+	5.71	4.93	5.20	5.57	3.48	5.51	10.53	9.27	6.28
Ours	**5.98**	**4.68**	**4.17**	**5.05**	**2.88**	**4.34**	**8.97**	7.71	**5.47**

Table 1: Comparison of different methods for the segmentation of TAO-affected organs on the T1 and T1c modality.

and drastically reduces the HD95 to 5.47 mm. Most remarkably, the segmentation of the challenging IOM, SOM improves by over 15%, 11% compared to the baseline. These results demonstrate that combining global alignment l_{csc} with local feature refinement l_{fpc} effectively mitigates domain discrepancies, ensuring anatomically plausible segmentation even for complex orbital structures.

To intuitively verify the impact of the PGCL in CMFA module on feature representation learning, we visualized the distribution of feature embeddings fp_m^k using t-SNE. Figure 3.4.1 compares the feature spaces before and after applying PGCL. Feature distribution without PGCL exhibits loose distribution with low intra-class compactness. Specifically, IOM from T1 and ON from T1c show vague boundaries and potential overlap in the central region. This lack of distinct separability explains the baseline model's struggle with class discrepancy. In contrast, Fig. 3.4.1B demonstrates that introducing PGCL significantly regularizes the latent space. Enhanced semantic separability validates that our method effectively mitigates the domain shift problem by aligning feature distributions.

l_{fpc}	l_{csc}	**T1 → T1c**							**T1c → T1**	Avg
		LR	IR	MR	SR	SOM	ON	IOM	LG	
						Dice[%] ↑				
		72.15	76.82	81.47	69.61	70.46	68.72	47.38	65.75	69.05
	✓	79.55	81.89	81.75	72.41	77.27	69.28	44.22	67.15	71.78
✓		80.20	80.32	84.10	74.88	76.74	70.84	53.87	66.27	73.40
✓	✓	**80.67**	81.90	**85.64**	**74.97**	**81.84**	**74.48**	**62.79**	**68.24**	**76.38**
						HD95[mm] ↓				
		8.31	10.30	10.20	9.06	8.12	8.31	12.25	10.05	9.57
	✓	7.79	7.09	7.20	5.90	7.10	5.29	9.68	7.69	7.21
✓		7.42	8.27	5.45	5.96	5.09	5.14	9.13	9.87	7.04
✓	✓	**5.98**	**4.68**	**4.17**	**5.05**	**2.88**	**4.34**	**8.97**	**7.71**	**5.47**

Table 2: Ablation results of Dice and HD95 on unlabeled organs (LR, IR, MR, SR, SOM, ON, IOM in T1c, and LG in T1) from TAO dataset.

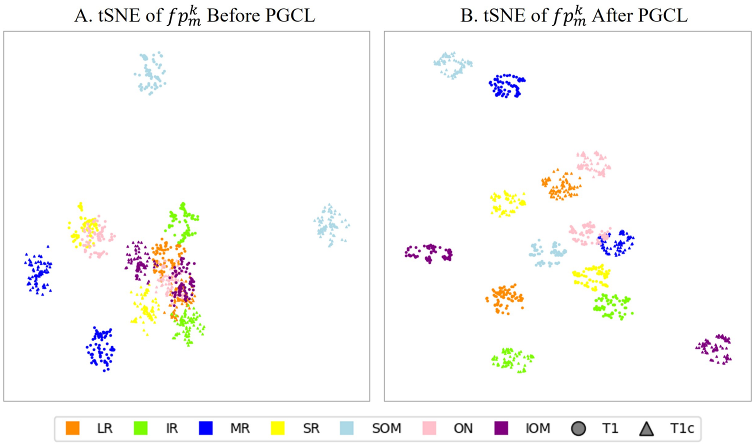

Figure 4: Left and right t-SNE visualize feature embeddings from TAO dataset without and with PGCL applied. Different markers as o, △ indicate features from T1 and T1c modality, respectively, with corresponding colors of organ groups marked in the legend

3.4.2. MS-CMRSeg

To further verify the generalizability of our proposed framework, we conducted an extensive ablation study on the organs of LV, RV, and myo in LGE from MS-CMRSeg dataset

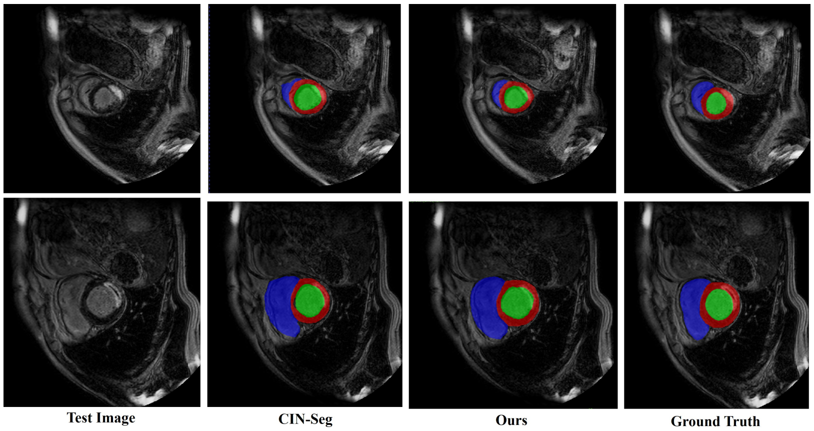

Figure 5: The qualitative results of segmentation on the MS-CMRSeg dataset.

shown Tab. 3. Consistent with the findings on TAO, the baseline model exhibited a similar suboptimal average Dice score of 76.27% and a high average HD95 of 35.08 mm. This performance degradation is particularly evident in myo and RV, which are notoriously difficult to segment in LGE images due to heterogeneous tissue contrast and potential pathology-mimicking artifacts. However, the application of our proposed CMFA and SCL modules yielded comparable performance gains. Individual items result in a substantial improvement, illustrating that l_{csc} works better for larger organs and l_{fpc} is particularly effective at preserving fine-grained structural details in the feature space. Robust and most significant performance is achieved by the integration of both components. In visual comparisons (see Fig. 3), our framework learns more accurate shapes and locations.

l_{fpc}	l_{csc}	Dice[%] ↑				HD95[mm] ↓			
		myo	LV	RV	Avg	myo	LV	RV	Avg
		68.97	86.32	73.52	76.27	45.36	10.60	49.27	35.08
	✓	71.77	87.55	77.09	78.80	10.22	9.45	27.73	15.80
✓		72.85	86.29	77.27	78.80	**9.69**	26.00	20.92	18.87
✓	✓	**74.97**	**88.54**	**78.71**	**80.74**	15.98	**9.36**	**13.80**	**13.04**

Table 3: Ablation results of Dice and HD95 on unlabeled organs (myo, LV, RV in LGE) from MS-CMRSeg dataset.

4. Conclusion

We propose a simple yet effective ACMCL method that combines SCL with CMFA normalized by CIN, to enhance alignment of cross-modality fine-grained semantic features. This strategy is advantageous for accommodating multi-modal data, by simply feeding interleaved inputs into the same batch. Our model performs comparably to or better than prevailing models in multi-organ segmentation from partly labeled multi-modal MRI.

Acknowledgments

This work was supported by the General Research Funding from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No. CUHK 14200721].

References

Matteo Bastico, David Ryckelynck, Laurent Corté, Yannick Tillier, and Etienne Decencière. A Simple and Robust Framework for Cross-Modality Medical Image Segmentation applied to Vision Transformers. *Proceedings - 2023 IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2023*, pages 4130–4140, 10 2023. doi: 10.1109/ICCVW60793.2023.00446.

Yuhua Chen, Wen Li, and Luc Van Gool. ROAD: Reality Oriented Adaptation for Semantic Segmentation of Urban Scenes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 11 2017. ISSN 10636919. doi: 10.1109/CVPR.2018.00823.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A Learned Representation For Artistic Style. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 10 2016.

Mingxuan Gu, Mareike Thies, Siyuan Mei, Fabian Wagner, Mingcheng Fan, Yipeng Sun, Zhaoya Pan, Sulaiman Vesal, Ronak Kosti, Dennis Possart, Jonas Utz, and Andreas Maier. Unsupervised Domain Adaptation Using Soft-Labeled Contrastive Learning with Reversed Monte Carlo Method for Cardiac Image Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 15009 LNCS:681–691, 2024. ISSN 16113349. doi: 10.1007/978-3-031-72114-4{_}65/TABLES/4.

Junlin Han, Mehrdad Shoeiby, Lars Petersson, and Mohammad Ali Armin. Dual Contrastive Learning for Unsupervised Image-to-Image Translation, 2021.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12962 LNCS:272–284, 1 2022. ISSN 16113349. doi: 10.1007/978-3-031-08999-2{_}22.

Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation. 12 2016.

Shruti Jadon. A survey of loss functions for semantic segmentation. *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2020*, 10 2020. doi: 10.1109/CIBCB48159.2020.9277638.

Suhyeon Lee, Junhyuk Hyun, Hongje Seong, and Euntai Kim. Unsupervised Domain Adaptation for Semantic Segmentation by Content Transfer. *Proceedings of the AAAI*

Conference on Artificial Intelligence, 35(9):8306–8315, 5 2021. ISSN 2374-3468. doi: 10.1609/AAAI.V35I9.17010.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection, 2017.

Taesung Park, Alexei A. Efros, Richard Zhang, and Jun Yan Zhu. Contrastive Learning for Unpaired Image-to-Image Translation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12354 LNCS:319–345, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58545-7{_}19/FIGURES/15.

Hyungseob Shin, Hyeongyu Kim, Sewon Kim, Yohan Jun, Taejoon Eo, and Dosik Hwang. SDC-UDA: Volumetric Unsupervised Domain Adaptation Framework for Slice-Direction Continuous Cross-Modality Medical Image Segmentation, 2023.

Nahian Siddique, Paheding Sidike, Colin Elkin, and Vijay Devabhaktuni. U-Net and its variants for medical image segmentation: theory and applications. *IEEE Access*, 9:82031–82057, 11 2020. doi: 10.1109/ACCESS.2021.3086020.

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

Aaron van den Oord DeepMind, Yazhe Li DeepMind, and Oriol Vinyals DeepMind. Representation Learning with Contrastive Predictive Coding.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation.

Xiaoyang Wang, Bingfeng Zhang, Limin Yu, and Jimin Xiao. Hunting Sparsity: Density-Guided Contrastive Learning for Semi-Supervised Semantic Segmentation.

Jianghao Wu, Dong Guo, Guotai Wang, Qiang Yue, Huijun Yu, Kang Li, and Shaoting Zhang. FPL+: Filtered Pseudo Label-based Unsupervised Cross-Modality Adaptation for 3D Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 2024. ISSN 1558254X. doi: 10.1109/TMI.2024.3387415.

Junlin Xian, Xiang Li, Dandan Tu, Senhua Zhu, Changzheng Zhang, Xiaowu Liu, Xin Li, and Xin Yang. Unsupervised Cross-Modality Adaptation via Dual Structural-Oriented Guidance for 3D Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 42(6):1774–1785, 6 2023. ISSN 1558254X. doi: 10.1109/TMI.2023.3238114.

Shuo Zhang, Jiaojiao Zhang, Biao Tian, Thomas Lukasiewicz, and Zhenghua Xu. Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 83:102656, 1 2023. ISSN 1361-8415. doi: 10.1016/J.MEDIA.2022.102656.

Ziyuan Zhao, Fangcheng Zhou, Kaixin Xu, Zeng Zeng, Cuntai Guan, and S. Kevin Zhou. LE-UDA: Label-Efficient Unsupervised Domain Adaptation for Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 42(3):633–646, 3 2023. ISSN 1558254X. doi: 10.1109/TMI.2022.3214766.

Xiahai Zhuang. Multivariate mixture model for cardiac segmentation from multi-sequence MRI. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9901 LNCS:581–588, 2016. ISSN 16113349. doi: 10.1007/978-3-319-46723-8{_}67/TABLES/1.

Xiahai Zhuang. Multivariate Mixture Model for Myocardial Segmentation Combining Multi-Source Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12): 2933–2946, 12 2019. ISSN 19393539. doi: 10.1109/TPAMI.2018.2869576.