Hierarchical Disentangled Consistency Learning for Semi-Supervised Multi-Modal Medical Image Segmentation

Feixiang Zhou[1], Zhuangzhi Gao[1], Jianyang Xie[1], Kun Li[1], Fu Wang[1], Long Chen[2], Zheheng Jiang[3], Yitian Zhao[4], Yanda Meng[5], He Zhao[1], Gregory Y.H. Lip[6], and Yalin Zheng[1](✉)

[1] Department of Eye and Vision Sciences, University of Liverpool, Liverpool, UK
yalin.zheng@liverpool.ac.uk
[2] Faculty of Medicine, Imperial College London, London, UK
[3] School of Computing and Mathematical Sciences, University of Leicester, Leicester, UK
[4] Ningbo Institute of Materials Technology and Engineering, CAS, Ningbo, China
[5] Bioengineering Program, Biological and Environmental Science and Engineering Division (BESE), King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia.
[6] Department of Cardiovascular and Metabolic Medicine, University of Liverpool, Liverpool, UK

Abstract. Semi-supervised learning (SSL) has achieved promising results in medical image segmentation by leveraging unlabeled data via consistency regularization. However, in multi-modal settings, heterogeneous modality characteristics and imbalanced reliability make effective consistency learning challenging. Existing methods typically impose prediction- or representation-level consistency on entangled features, which may suppress modality-specific cues and limit the exploitation of complementary cross-modal information in unlabeled data. To address this limitation, we propose HDCL, a hierarchical disentangled consistency learning framework to promote reliable and robust semi-supervised multi-modal medical image segmentation. Specifically, Disentangled Representation Learning (DRL) explicitly decomposes multi-modal features into shared and modality-specific representations, enabling targeted modeling of invariant semantics and modality-specific cues. We then propose Shared-Specific Dual-path Decoding (SSDD), which simultaneously decodes shared, specific, and fused representations to produce diverse modality-conditioned predictions. Furthermore, Multi-level Prediction Consistency Learning (MPCL) is proposed to enforce prediction consistency across different semantic spaces, allowing more effective utilization of unlabeled data without compromising modality-specific characteristics. Extensive experiments on two benchmarks demonstrate that our method consistently outperforms existing methods across different annotation settings. The source codes will be available at https://github.com/FeixiangZhou/HDCL .

Keywords: Semi-supervised learning · Multi-modal segmentation.

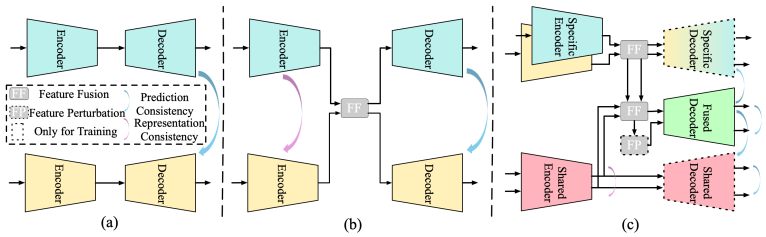

Fig. 1. Comparison of different consistency learning paradigms on unlabelled data for semi-supervised multi-modal segmentation. (a) Prediction-level consistency using a dual-encoder–decoder framework. (b) Representation- and prediction-level consistency with feature fusion. (c) Our proposed HDCL, which disentangles shared and modality-specific representations and performs hierarchical consistency learning.

1 Introduction

Multi-modal medical image segmentation [22,10,19] plays a crucial role in a wide range of clinical applications, as different imaging modalities provide complementary information for accurate anatomical and pathological delineation. By jointly leveraging heterogeneous imaging sources, multi-modal learning has demonstrated clear advantages over unimodal approaches [23] in tasks including lesion segmentation [6], disease diagnosis [20], and prognosis prediction [8]. However, most existing multi-modal segmentation methods are developed under a fully supervised learning paradigm, relying on large-scale voxel-level annotations. Despite their effectiveness, acquiring dense annotations across multiple modalities is time-consuming, costly, and requires substantial expert effort, which significantly limits the scalability and clinical applicability of these methods.

To mitigate the heavy reliance on large annotated datasets, SSL [2,9,16] has attracted increasing attention in medical image segmentation. SSL methods aim to exploit abundant unlabeled data by introducing additional regularization signals or pseudo-supervision during training. However, most of the existing methods are designed for single-modality scenarios. Recently, a few studies [3,4,5] have investigated semi-supervised learning for multi-modal medical image segmentation. Zhang et al. [21] and Zhu et al. [25] encouraged multi-modal interaction by minimizing consistency losses across different modalities on abundant unlabeled data. However, these methods mainly enforce prediction-level consistency on unlabeled data, as shown in Fig. 1 (a), which limits the full utilization of multi-modal information. To address this issue, Meng et al. [11] and Zhou et al. [24] proposed cross-modal collaboration strategies that perform feature fusion or alignment across modalities (Fig. 1 (b)). While these approaches enhance representation-level consistency, they still rely on entangled features and do not explicitly distinguish shared semantics from modality-specific characteristics, which may suppress modality-specific cues and hinder effective exploitation of complementary multi-modal information under semi-supervised settings.

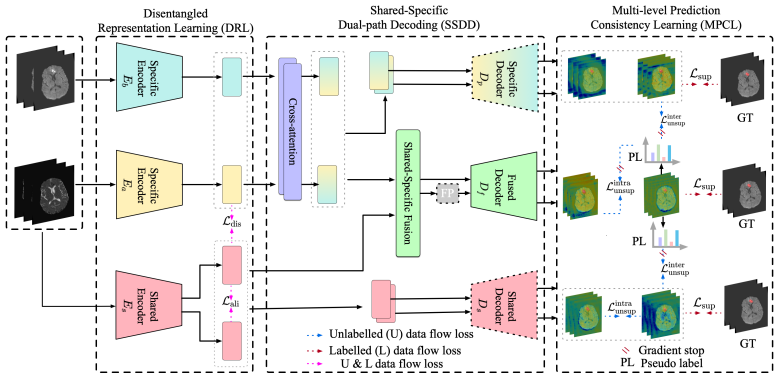

Fig. 2. Overview of the proposed HDCL. DRL disentangles multi-modal features into shared and modality-specific representations (Sec. 2.1). SSDD employs dual-path decoders to generate diverse predictions from shared, specific, and fused spaces (Sec. 2.2). MPCL enforces intra- and inter-space prediction consistency across disentangled semantic spaces (Sec. 2.3).

To overcome the above limitations, we propose HDCL, which imposes consistency constraints in a hierarchical manner. Specifically, inspired by multi-modal disentangled learning [14,17], we propose DRL to decompose multi-modal features into shared and modality-specific components and enforce consistency on the shared representations to align invariant cross-modal semantics. Then, we introduce SSDD to construct three decoders for shared, specific, and fused feature spaces, respectively, each equipped with dual inputs to promote prediction diversity. Finally, we propose MPCL to regularize unlabeled data by enforcing prediction consistency across different semantic spaces (i.e., different decoders). Compared with prior semi-supervised multi-modal segmentation methods, the key novelty of HDCL lies in hierarchical consistency learning over disentangled semantic spaces. Existing approaches mainly impose consistency on predictions or entangled fused features, whereas HDCL first separates shared and modality-specific representations, then performs targeted representation- and prediction-level consistency learning. This enables the model to exploit unlabeled data while preserving complementary modality-specific information.

The main contributions are summarized as follows: **1)** We propose a novel HDCL framework for semi-supervised multi-modal medical image segmentation, which enforces hierarchical consistency at the representation and prediction levels. **2)** We introduce DRL to separate multi-modal features into shared and modality-specific components. **3)** We develop SSDD to produce modality-conditioned predictions from shared, specific, and fused semantic spaces, promoting prediction diversity. **4)** We propose MPCL to enforce intra- and inter-space prediction consistency across different semantic spaces.

2 Methodology

An overview of the proposed method is illustrated in Fig. 2. We consider a semi-supervised multi-modal segmentation setting with two modalities, denoted as modality a and b. Following [11,24], the labeled and unlabeled datasets are defined as $\mathcal{D}_l^a = \{(x_i^a, y_i)\}_{i=1}^M, \mathcal{D}_u^a = \{x_i^a\}_{i=1}^N$ and $\mathcal{D}_l^b = \{(x_i^b, y_i)\}_{i=1}^M, \mathcal{D}_u^b = \{x_i^b\}_{i=1}^N$, where x_i^b and x_i^b denote the input images of modality a and modality b with the size of $H \times W \times D$. $y_i \in \mathbb{R}^{C \times H \times W \times D}$ is the ground truth with C classes. M and N denote the number of labeled and unlabeled samples and $M \ll N$.

2.1 Disentangled Representation Learning

Inspired by multi-modal disentanglement methods [14], we adopt a shared encoder $E_s(\cdot)$ and two modality-specific encoders $E_a(\cdot)$ and $E_b(\cdot)$ to learn shared and specific representations. Given an input x^m from modality $m \in \{a, b\}$, we obtain $z_s^m = E_s(x^m)$ and $z_p^m = E_m(x^m)$, where $z_s^m, z_p^m \in \mathbb{R}^{C^{'} \times H^{'} \times W^{'} \times D^{'}}$. To align shared representations across modalities, we minimize the L_1 distance between z_s^a and z_s^b, defined as:

$$\mathcal{L}_{\text{ali}} = \mathbb{E}_{x^a, x^b \sim \mathcal{D}} \left[\left\| z_s^a - z_s^b \right\|_1 \right] \tag{1}$$

where the expectation is taken over $\mathcal{D} = \{\mathcal{D}_l^m, \mathcal{D}_u^m | m \in \{a, b\}\}$. By applying this constraint to unlabeled samples, the model learns modality-invariant semantics without ground-truth supervision, improving the utilization of unlabeled data. Meanwhile, to avoid redundancy between shared and specific representations, we enforce their disentanglement by minimizing the cosine similarity between z_s^m and z_p^m for each modality. Thus, we have:

$$\mathcal{L}_{\text{dis}} = \mathbb{E}_{x^a, x^b \sim \mathcal{D}} \left[\sum_{m \in \{a,b\}} \left| \frac{\langle \bar{z}_s^m, \bar{z}_p^m \rangle}{\| \bar{z}_s^m \|_2 \, \| \bar{z}_p^m \|_2 + \epsilon} \right| \right] \tag{2}$$

where $\bar{z}_s^m, \bar{z}_p^m \in \mathbb{R}^{C'}$ are obtained by global average pooling on z_s^m and z_p^m, respectively. $\langle \cdot, \cdot \rangle$ denotes the inner product and ϵ is a small constant added for numerical stability. The overall DRL loss is then defined as $\mathcal{L}_{\text{drl}} = \mathcal{L}_{\text{ali}} + \mathcal{L}_{\text{dis}}$. Through joint shared alignment and shared–specific disentanglement, DRL learns a structured representation space that decouples invariant and modality-specific information.

2.2 Shared-Specific Dual-path Decoding

Unlike existing methods [11,24] that decode mutually enhanced single-modality features separately, our SSDD explicitly organizes decoding in shared, specific, and fused semantic spaces. This structured multi-space design enables coordinated supervision of invariant and modality-specific information, thereby promoting prediction diversity and more complementary supervision.

The shared decoding branch aims to leverage modality-invariant semantics for consistent prediction. Specifically, z_s^a and z_s^b are directly fed into a modality-shared decoder $D_s(\cdot)$, yielding the corresponding predictions $\hat{y}_s^a = D_s(z_s^a)$ and $\hat{y}_s^b = D_s(z_s^b)$. Although modality-specific representations preserve complementary information, their distributional discrepancy across modalities poses a challenge for joint decoding. To reduce the modality gap while maintaining modality-specific characteristics, we introduce a cross-attention (CA) based enhancement mechanism for the specific decoding branch. Given z_p^a and z_p^b, we apply two symmetric cross-attention modules, where each modality attends to the other:

$$\begin{aligned} \tilde{z}_p^a &= \mathrm{Softmax}\left(\frac{(W_Q z_p^a)(W_K z_p^b)^\top}{\sqrt{C'}}\right)(W_V z_p^b) \\ \tilde{z}_p^b &= \mathrm{Softmax}\left(\frac{(W_Q z_p^b)(W_K z_p^a)^\top}{\sqrt{C'}}\right)(W_V z_p^a) \end{aligned} \tag{3}$$

where W_Q, W_K, and W_V are learnable projection matrices. Through cross-attention, each modality selectively incorporates complementary cues from the other modality while preserving its own modality-specific structure. $\tilde{z}_p^a$ and $\tilde{z}_p^b$ are then fed into a modality-specific decoder $D_p(\cdot)$ to produce two predictions $\hat{y}_p^a = D_p(\tilde{z}_p^a)$ and $\hat{y}_p^b = D_p(\tilde{z}_p^b)$. To integrate complementary shared and specific information, we introduce a fused decoding branch. The shared and enhanced specific representations of each modality are combined to obtain modality-aware representations, which are further aggregated into a unified fused feature. Specifically, the fused feature is defined as:

$$z_f = \left[\mathrm{MLP}([z_s^a; \tilde{z}_p^a])\,;\, \mathrm{MLP}([z_s^b; \tilde{z}_p^b])\right] \in \mathbb{R}^{2C' \times H' \times W' \times D'} \tag{4}$$

where $[\cdot;\cdot]$ denotes concatenation and $\mathrm{MLP}(\cdot)$ is a lightweight multilayer perceptron. The resulting z_f aggregates invariant semantics and modality-specific cues, providing a compact yet expressive representation for joint decoding.

Inspired by data augmentation [18,9] in SSL, we further introduce feature perturbation (FP) in the fused space to improve robustness. Specifically, stochastic perturbations (i.e., noise or dropout) are applied to z_f to generate $\tilde{z}_f$, encouraging the decoder to produce stable predictions under feature-level variations. z_f and $\tilde{z}_f$ are subsequently fed into a fused decoder $D_f(\cdot)$ to generate two predictions $\hat{y}_f = D_f(z_f)$ and $\hat{y}_f^p = D_f(\tilde{z}_f)$.

2.3 Multi-level Prediction Consistency Learning

MPCL is designed to fully exploit the diverse predictions generated by the shared, specific, and fused decoding branches, including both intra- and inter-space prediction consistency. For labeled samples, each decoder produces two predictions, resulting in six predictions $\mathcal{P} = \left\{\hat{y}_s^a, \hat{y}_s^b, \hat{y}_p^a, \hat{y}_p^b, \hat{y}_f, \hat{y}_f^p\right\}$ in total. All predictions are supervised using cross-entropy and Dice losses [24]:

$$\mathcal{L}_{\mathrm{sup}} = \mathbb{E}_{x^a, x^b, y \sim \mathcal{D}_l}\left[\textstyle\sum_{\hat{y}_k \in \mathcal{P}} (\mathcal{L}_{\mathrm{CE}}(\hat{y}_k, y) + \mathcal{L}_{\mathrm{Dice}}(\hat{y}_k, y))\right] \tag{5}$$

where $\mathcal{D}_l = \left\{D_l^a, D_l^b\right\}$. For unlabeled data, intra-space consistency is imposed on the shared and fused spaces to promote prediction stability, while the specific space is excluded to preserve modality-specific cues. The loss is defined as:

$$\mathcal{L}_{\text{unsup}}^{\text{intra}} = \mathbb{E}_{x^a, x^b \sim D_u}\Big[\left\|\hat{y}_s^a - \hat{y}_s^b\right\|_2^2 + \mathbb{I}(\max \hat{y}_f \geq \tau)\, \mathcal{L}_{\text{Dice}}(\hat{y}_f^p, \phi(\tilde{y}_f))\Big] \tag{6}$$

where $\mathcal{D}_u = \left\{D_u^a, D_u^b\right\}$ and $\phi(\cdot)$ is the stop-gradient function. $\tilde{y}_f$ is the pseudo labels (PL) generated from $\hat{y}_f$, which encourages the fused decoder to produce stable predictions under feature-level perturbations. The confidence threshold τ is set as 0.95. Beyond intra-space regularization, MPCL enforces inter-space consistency across semantic spaces. Since the fused space integrates both shared and modality-specific information, it provides more reliable predictions. Accordingly, the inter-space consistency loss is defined as:

$$\mathcal{L}_{\text{unsup}}^{\text{inter}} = \mathbb{E}_{x^a, x^b \sim D_u}\left[\textstyle\sum_{m \in \{a,b\}} \mathbb{I}(\max \hat{y}_f \geq \tau)\left(\mathcal{L}_{\text{Dice}}(\hat{y}_s^m, \phi(\tilde{y}_f)) + \mathcal{L}_{\text{Dice}}(\hat{y}_p^m, \phi(\tilde{y}_f))\right)\right] \tag{7}$$

Overall, the total loss of our HDCL is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \alpha\mathcal{L}_{\text{drl}} + \beta\mathcal{L}_{\text{unsup}}$, where $\mathcal{L}_{\text{unsup}} = \mathcal{L}_{\text{unsup}}^{\text{intra}} + \mathcal{L}_{\text{unsup}}^{\text{inter}}$. α and β are hyperparameters that balance the contributions of different losses.

3 Experiments

3.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate our method on two brain lesion datasets. ISLES-2022 [7] contains 250 multi-channel MRI scans (DWI, ADC, FLAIR) for ischemic stroke lesion segmentation. We use DWI and ADC, following [1], with 200 images for training and 50 for validation and testing. BraTS-2019 [12] includes preoperative MRI scans (FLAIR, T1, T1ce, T2) from 335 glioma patients. Following [11], we use T1ce and T2 for whole tumor segmentation. As in [9], 250 samples are used for training, 25 for validation, and 60 for testing. Performance is evaluated using Dice, Jaccard, 95HD, and ASD.

Implementation Details. All experiments are conducted using PyTorch 2.1.1 with two NVIDIA H100 GPUs. We utilize the SGD optimizer with an initial learning rate of 0.01. The batch size is set to 2, and the number of iterations is set to 60k. Training images undergo 3D augmentations, such as random cropping, flipping, and rotation, to improve generalization. Patches of size $96 \times 96 \times 64$ are used for ISLES-2022, and $96 \times 96 \times 96$ for BraTS-2019. α is set to 0.01 and β represents warm-up function $\beta = 0.1\, e^{-5\left(1-\frac{t}{t_{\max}}\right)^2}$, where t denotes the current number of iterations and $t_{\max}$ represents the maximum number of iterations.

3.2 Results

Comparison with State-of-the-art Methods. We compare our proposed HDCL with several representative methods on the ISLES-2022 and BraTS-2019

Table 1. Quantitative comparison of our method with other state-of-the-art methods on the ISLES-2022 and BraTS-2019 datasets.

Dataset	Method	5% / 10 labeled data				10% / 20 labeled data			
		Dice (%) ↑	Jaccard (%) ↑	95HD ↓	ASD ↓	Dice (%) ↑	Jaccard (%) ↑	95HD ↓	ASD ↓
ISLES-2022	MT [13]	62.75	48.95	20.02	6.17	64.37	50.35	19.72	6.29
	SGRS-Net [15]	67.24	54.95	13.32	4.34	70.18	58.67	8.67	2.98
	MMCML [21]	65.19	51.65	17.09	5.11	68.03	54.64	16.41	4.50
	CMC [24]	68.07	55.32	11.79	2.50	71.15	58.85	9.79	2.17
	MECF [5]	71.35	58.63	10.83	2.25	74.47	62.25	7.47	2.22
	SMMS [11]	72.57	59.80	8.32	1.63	75.61	63.20	6.70	1.71
	Ours	**75.95**	**63.52**	**7.08**	**0.90**	**78.01**	**66.32**	**4.73**	**0.79**
BraTS-2019	Method	5% / 13 labeled data				10% / 25 labeled data			
	MT [13]	77.60	65.84	17.56	3.59	80.22	68.70	15.20	3.11
	SGRS-Net [15]	80.40	68.03	15.57	3.01	81.35	69.67	16.78	2.92
	MMCML [21]	73.25	60.80	18.27	7.85	77.96	66.88	17.83	6.68
	CMC [24]	79.21	67.99	14.11	2.90	80.67	69.60	14.00	2.79
	MECF [5]	80.84	70.31	14.01	2.71	82.28	71.85	13.23	2.12
	SMMS [11]	81.22	69.94	13.57	2.52	82.50	72.43	12.94	2.27
	Ours	**84.08**	**74.00**	**12.65**	**2.42**	**85.15**	**75.37**	**10.43**	**1.79**

datasets under different labeled data settings, including semi-supervised learning approaches (e.g., MT [13], SGRS-Net [15]) and semi-supervised multi-modal learning approaches (e.g., MMCML [21], CMC [24], MECF [5], and SMMS [11]).

As shown in Table. 1, under the extremely limited 5% labeled setting, HDCL achieves the best Dice of 75.95%, outperforming the strongest competing method (i.e., SMMS) by 3.38%. Meanwhile, our method achieves the lowest 95HD and ASD, indicating more accurate boundary delineation and better structural preservation of stroke lesions. When increasing the labeled ratio to 10%, HDCL consistently outperforms all compared methods across all evaluation metrics, achieving 78.01% Dice and further reducing 95HD to 4.73. These results demonstrate the robustness of our hierarchical disentangled consistency learning framework, particularly in low-annotation regimes. On BraTS-2019, a similar trend can be observed. HDCL consistently achieves the best performance under both 5% and 10% labeled settings. Overall, the improvements on two datasets validate the effectiveness of enforcing hierarchical consistency constraints within a disentangled multi-modal framework for better use complementary information.

Qualitative Analysis. Fig. 3(a) presents qualitative results under the 10% labeled setting. Compared with CMC and SMMS, our method produces lesion regions that are more complete and closer to the ground truth. In particular, HDCL better preserves small or fragmented lesions on ISLES-2022 and generates smoother and more accurate boundaries on BraTS-2019. We also visualize the learned shared and modality-specific representations, as shown in Fig. 3(b). The shared representations from different modalities are well aligned and exhibit overlapping distributions, indicating effective cross-modal semantic consistency. In contrast, modality-specific representations form clearly separated clusters, demonstrating that HDCL successfully disentangles shared and specific features.

Ablation Study. In Table 2, we conduct ablation experiments under the 10% labeled setting to evaluate the contribution of each component. Starting from the baseline, introducing DRL improves Dice from 72.71% to 74.76% on ISLES-

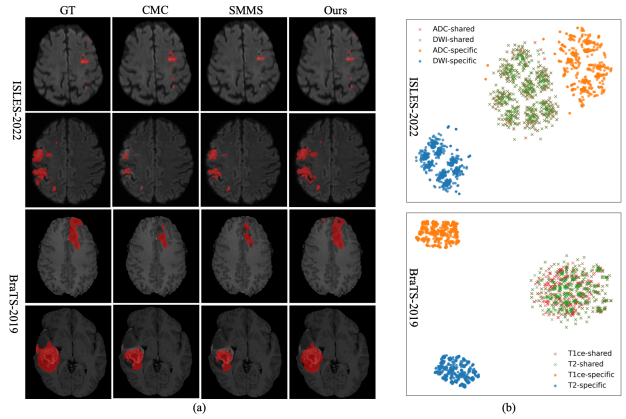

Fig. 3. (a) Segmentation results on two datasets under the 10% labeled setting. (b) t-SNE visualization of the learned shared and modality-specific representations.

Table 2. Ablation study on the key components using 10% labeled data.

Settings			ISLES-2022 (20 labeled data)				BraTS-2019 (25 labeled data)			
DRL	SSDD	MPCL	Dice ↑	IoU ↑	95HD ↓	ASD ↓	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
			72.71	60.16	9.15	1.93	79.00	67.53	14.33	2.78
✓			74.76	62.46	7.73	1.77	81.84	70.50	13.73	2.53
✓	✓		76.71	64.74	6.28	1.25	83.23	72.83	12.34	1.91
✓	✓	✓	**78.01**	**66.32**	**4.73**	**0.79**	**85.15**	**75.37**	**10.43**	**1.79**

Table 3. Ablation study of intra- and inter-space prediction consistency losses.

Settings	ISLES-2022 (20 labeled data)				BraTS-2019 (25 labeled data)			
	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓	Dice ↑	IoU ↑	95HD ↓	ASD ↓
$\mathcal{L}_{\text{unsup}}^{\text{intra}}$	77.23	65.40	5.09	1.07	84.58	74.94	11.63	2.06
$\mathcal{L}_{\text{unsup}}^{\text{inter}}$	76.91	65.01	6.16	1.23	83.81	73.47	12.08	2.23
$\mathcal{L}_{\text{unsup}}$	**78.01**	**66.32**	**4.73**	**0.79**	**85.15**	**75.37**	**10.43**	**1.79**

2022 and also brings consistent gains on BraTS-2019, indicating that disentangling shared and modality-specific representations enhances feature learning. Adding SSDD further boosts performance, increasing Dice to 76.71% and reducing boundary errors. This can be attributed to the dual-path decoding strategy, which provides complementary prediction signals from disentangled semantic spaces. Finally, incorporating MPCL achieves the best results across all metrics. These results show that each component contributes positively and that their combination leads to consistent performance improvements. We further evaluate the effects of intra- and inter-space prediction consistency losses in Table 3. Using only intra-space consistency yields strong performance, while inter-space consistency alone performs slightly worse. Combining both achieves the best results across all metrics, demonstrating their complementary effects.

4 Conclusions

In this paper, we propose a novel HDCL framework that performs hierarchical representation- and prediction-level consistency leaning across disentangled semantic spaces for semi-supervised multi-modal medical image segmentation. We introduced DRL to disentangle shared and modality-specific representations, followed by SSDD to produce predictions from different semantic spaces, thereby enhancing prediction diversity. MPCL is further proposed to enforce intra- and inter-space prediction consistency, leading to improved segmentation reliability. Extensive experiments on two datasets demonstrate that our method achieves state-of-the-art performance under different labeled data settings. Although HDCL is evaluated on two-modality brain lesion segmentation tasks, its shared-specific representation design is modality-agnostic and can potentially be extended to more modalities by introducing additional modality-specific encoders and aggregating their shared and specific representations in the fused space. In future work, we will further investigate its generalization to missing-modality scenarios, larger modality combinations, and more diverse clinical applications.

Acknowledgments. This study was funded by TARGET, a project funded by the EU HORIZON EUROPE framework programme for research and innovation under the Grant Agreement No. 101136244.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Assefa, M., Naseer, M., Ganapathi, I.I., Ali, S.S., Seghier, M.L., Werghi, N.: Dycon: Dynamic uncertainty-aware consistency and contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30850–30860 (2025)
2. Basak, H., Yin, Z.: Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19786–19797 (2023)
3. Chartsias, A., Papanastasiou, G., Wang, C., Semple, S., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A.: Disentangle, align and fuse for multimodal and semi-supervised image segmentation. IEEE transactions on medical imaging **40**(3), 781–792 (2020)
4. Chen, X., Zhou, H.Y., Liu, F., Guo, J., Wang, L., Yu, Y.: Mass: Modality-collaborative semi-supervised segmentation by exploiting cross-modal consistency from unpaired ct and mri images. Medical Image Analysis **80**, 102506 (2022)
5. Chung, T.D., Lam, B.T., Nguyen, T.H., Nguyen, T., Vu, N.L.V., Cao, H.L., Huynh, P.K., Xu, M.: Modality-specific enhancement and complementary fusion for semi-supervised multi-modal brain tumor segmentation. arXiv preprint arXiv:2512.09801 (2025)
6. Fang, L., Wang, X.: Brain tumor segmentation based on the dual-path network of multi-modal mri images. Pattern Recognition **124**, 108434 (2022)

7. Hernandez Petzsche, M.R., De La Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.L., Kofler, F., Ezhov, I., et al.: Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific data **9**(1), 762 (2022)
8. Hou, J., Zhang, R., Xie, Y., Li, C., Qin, W.: Multimodal deep learning for cancer prognosis prediction with clinical information prompts integration. npj Digital Medicine (2025)
9. Lei, T., Yang, Z., Wang, X., Wang, Y., Wang, X., Sun, F., Nandi, A.K.: Adaptive learning of high-value regions for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21450–21459 (2025)
10. Liang, J., Dai, L., Sheng, X., Chen, X., Yao, C., Tao, G., Leng, Q., Cai, H., Zhong, X.: Hwa-unetr: Hierarchical window aggregate unetr for 3d multimodal gastric lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 273–282. Springer (2025)
11. Meng, D., Li, S., Wu, H., Wang, G., Yan, X.: Semi-supervised multi-modal medical image segmentation for complex situations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 487–496. Springer (2025)
12. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
13. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
14. Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Multi-modal learning with missing modality via shared-specific feature modelling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15878–15887 (2023)
15. Wang, T., Zhang, X., Chen, Y., Zhou, Y., Zhao, L., Tan, T., Tong, T.: Synergy-guided regional supervision of pseudo labels for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 530–540. Springer (2025)
16. Wang, Y., Song, K., Liu, Y., Ma, S., Yan, Y., Carneiro, G.: Leveraging labelled data knowledge: A cooperative rectification learning network for semi-supervised 3d medical image segmentation. Medical Image Analysis **101**, 103461 (2025)
17. Wang, Y., Albrecht, C.M., Braham, N.A.A., Liu, C., Xiong, Z., Zhu, X.X.: Decoupling common and unique representations for multimodal self-supervised learning. In: European Conference on Computer Vision. pp. 286–303. Springer (2024)
18. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7236–7246 (2023)
19. Yu, F., Cao, J., Liu, L., Jiang, M.: Superlightnet: Lightweight parameter aggregation network for multimodal brain tumor segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5197–5206 (2025)
20. Yu, J., Ge, R., Wang, Z., Yang, C., Lin, C., Fu, X., Liu, J., Elazab, A., Wang, C.: Small lesions-aware bidirectional multimodal multiscale fusion network for lung disease classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 589–598. Springer (2025)

21. Zhang, S., Zhang, J., Tian, B., Lukasiewicz, T., Xu, Z.: Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation. Medical Image Analysis **83**, 102656 (2023)
22. Zhang, Y., Yang, J., Tian, J., Shi, Z., Zhong, C., Zhang, Y., He, Z.: Modality-aware mutual learning for multi-modal medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 589–599. Springer (2021)
23. Zhou, F., Gao, Z., Zhao, H., Xie, J., Meng, Y., Zhao, Y., Lip, G.Y., Zheng, Y.: Glcp: Global-to-local connectivity preservation for tubular structure segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 237–247. Springer (2025)
24. Zhou, X., Sun, Y., Deng, M., Chu, W.C.W., Dou, Q.: Robust semi-supervised multimodal medical image segmentation via cross modality collaboration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 57–67. Springer (2024)
25. Zhu, L., Yang, K., Zhang, M., Chan, L.L., Ng, T.K., Ooi, B.C.: Semi-supervised unpaired multi-modal learning for label-efficient medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 394–404. Springer (2021)