Divide, Conquer, and Aggregate: Asymmetric Experts for Class-Imbalanced Semi-Supervised Medical Image Segmentation

Supplementary Material

1. More experiments results on ACDC dataset

To assess the efficacy of our method on standard benchmarks with fewer semantic categories, we conduct experiments on the ACDC [3] dataset. This dataset serves as one of the most widely used benchmarks for developing general semi-supervised medical image segmentation (SSMIS) approaches. Originating from the Automated Cardiac Diagnosis Challenge, it contains 100 short-axis cardiac MRI scans with annotations for three classes: left ventricle, right ventricle, and myocardium. The data split follows established protocols with 70, 10, and 20 samples for training, validation, and testing, respectively.

During training, we utilize a 2D U-Net [17] as the backbone, optimized by SGD with an initial learning rate of 0.001. For preprocessing, training slices are cropped into 256×256 patches. Following prior works [1, 7, 9], we employ the Dice Score (%), Jaccard Score (%), 95th percentile Hausdorff Distance (HD95), and Average Surface Distance (ASD) as evaluation metrics. The batch size is set to 4, comprising two labeled and two unlabeled patches per iteration.

Table 1 shows the results on the ACDC dataset with 5% labeled data. General SSMIS methods like ABD and β-FFT exhibit competitive performance in average Dice and specific metrics for Right Ventricle (RV), Myocardium (Myo), and Left Ventricle (LV), likely because these general methods are typically developed and tuned on the ACDC dataset. Our method trails the best performer (β-FFT) marginally by 0.26% in average Dice but surpasses the best class-imbalance method (GA) by 2.58%. However, as shown in Table 2 with 10% labeled data, our method not only outperforms all class-imbalance methods but also surpasses the best general semi-supervised method (β-FFT) by 0.56%. Combining the results from both settings, our method proves to be highly competitive against general SSMIS approaches on datasets with fewer classes while significantly outperforming existing class-imbalance methods.

Qualitative comparisons in Figure 2 for the 10% labeled setting show that all methods struggle with incomplete segmentation of the RV. BCP, DHC, and SKCDF also exhibit incomplete segmentation for the Myo and LV. Although our method, like β-FFT, shows minor errors in the RV, it remains the closest to the Ground Truth. In the 10% labeled setting, all comparison methods generate false positives (over-segmentation) in background regions where no target organs exist. BCP, DHC, and GA misclassify background as Myo and LV; SKCDF as Myo, LV, and RV; ABD as Myo; and β-FFT as RV. Our method is the only one that produces no false positives, demonstrating superior efficacy in low-target scenarios compared to both general and imbalance-aware semi-supervised methods.

Finally, the t-SNE visualization in Figure 2 shows varying degrees of overlap between Myo and LV across all methods, which is expected due to their inherent anatomical proximity. Nevertheless, our method exhibits the best separation and minimal overlap between RV and Myo, as well as between RV and LV.

2. More experiments results on Synapse dataset

Table 3 presents the quantitative results on the Synapse [11] dataset with 10% labeled data. Our method demonstrates superior performance, surpassing the best class-imbalance method (GA) and the leading general SSMIS method (MagicNet) by margins of 4.72% and 7.79% in average Dice score, respectively. Furthermore, our approach achieves top-3 performance across all 13 organs, ranking first in 10 of them. Notably, we observe significant improvements in small organs such as the esophagus (Es), portal & splenic veins (PSV), and Left Adrenal Gland (LAG), outperforming the second-best method by 8.3%, 12.3%, and 7.5%, respectively. This highlights the efficacy of our proposed method in handling class imbalance.

As visualized in Figure 3, qualitative comparisons in the axial plane (red ellipses) reveal that BCP and ABD fail to completely segment large organs like the stomach. This further validates the limitations of transferring general SSMIS methods, developed on datasets with fewer classes, to multi-organ segmentation tasks with severe class imbalance. Meanwhile, β-FFT, GA, and SKCDF all produce erroneous segmentations in the duodenum region, and DHC results in both incomplete stomach segmentation and misclassified duodenum. In contrast, our method yields results closest to the Ground Truth, avoiding both over- and under-segmentation. In the coronal plane (blue ellipses), ABD, β-FFT, GA, and SKCDF exhibit over-segmentation in the prostate/uterus region, with ABD also showing distinct under-segmentation of the bladder. In the complex

	Methods	Avg				RV				Myo				LV			
		Dice↑	Jac↑	95HD↓	ASD↓	Dice↑	Jac↑	95HD↓	ASD↓	Dice↑	Jac↑	95HD↓	ASD↓	Dice↑	Jac↑	95HD↓	ASD↓
	U-Net [17] (fully)	91.31	83.18	4.24	0.91	91.47	84.26	3.92	0.95	90.68	81.74	5.89	1.12	91.78	86.54	2.91	0.66
General	UA-MT [22]	50.21	39.72	17.63	5.37	50.38	40.51	16.42	5.61	49.71	38.42	20.18	6.38	50.54	40.23	16.29	4.12
General	CPS [6]	49.36	41.89	8.92	1.86	49.54	42.67	8.31	1.92	48.88	40.59	10.17	2.21	49.66	42.41	8.28	1.45
General	SSNet [21]	66.97	54.49	6.81	2.33	67.21	55.38	6.34	2.41	66.48	53.19	7.92	2.78	67.22	54.90	6.17	1.80
General	Co-BioNet [15]	88.05	79.61	4.22	1.18	88.31	80.49	3.92	1.23	87.52	78.31	4.81	1.39	88.32	80.03	3.93	0.92
General	BCP [1]	89.53	77.42	1.94	0.68	89.78	78.27	1.81	0.71	89.01	76.14	2.24	0.81	89.80	77.85	1.77	0.52
General	ABD [7]	89.97	81.83	1.59	0.51	90.21	82.69	1.48	0.53	89.47	80.57	1.83	0.61	90.23	82.23	1.46	0.39
General	β-FFT [9]	90.37	80.62	1.81	0.56	90.61	81.47	1.68	0.59	89.87	79.35	2.09	0.67	90.63	81.04	1.66	0.42
Imbalance	CReST [20]	45.36	36.78	37.64	14.42	45.54	37.58	35.12	15.03	44.82	35.46	42.38	17.21	45.72	37.30	35.42	10.99
Imbalance	SimiS [5]	64.72	50.97	23.73	6.89	64.94	51.83	22.14	7.18	64.17	49.68	27.31	8.14	65.05	51.40	21.74	5.35
Imbalance	Basak *et al.* [2]	53.49	41.27	42.85	16.53	53.68	42.07	39.86	17.24	52.94	39.98	48.12	19.63	53.85	41.76	40.57	12.72
Imbalance	CLD [12]	59.67	46.34	29.41	10.74	59.88	47.19	27.38	11.22	59.13	45.06	33.46	12.71	60.00	46.77	27.39	8.29
Imbalance	DHC [18]	56.72	47.89	30.68	2.09	56.91	48.72	28.64	2.18	56.18	46.61	34.78	2.49	57.07	48.34	28.62	1.60
Imbalance	GA [16]	87.53	80.46	3.96	1.09	87.78	81.33	3.68	1.14	87.01	79.17	4.52	1.29	87.80	80.88	3.68	0.84
Imbalance	SKCDF [23]	85.28	75.25	5.34	1.36	85.52	76.12	4.97	1.41	84.73	73.97	6.08	1.61	85.59	75.66	4.97	1.06
Imbalance	**DCA (Ours)**	**90.11**	**82.38**	**1.66**	**0.58**	**90.36**	**83.24**	**1.54**	**0.61**	**89.61**	**81.12**	**1.92**	**0.69**	**90.36**	**82.78**	**1.52**	**0.44**

Table 1. Quantitative comparison between our approach and SOTA methods on **5%** labeled **ACDC** dataset. 'General' or 'Imbalance' indicate whether the methods consider class-imbalance issue or not. The top-3 results are highlighted as first , second , and third .

	Methods	Avg				RV				Myo				LV			
		Dice↑	Jac↑	95HD↓	ASD↓	Dice↑	Jac↑	95HD↓	ASD↓	Dice↑	Jac↑	95HD↓	ASD↓	Dice↑	Jac↑	95HD↓	ASD↓
	U-Net [17] (fully)	91.31	83.18	4.24	0.91	91.47	84.26	3.92	0.95	90.68	81.74	5.89	1.12	91.78	86.54	2.91	0.66
General	UA-MT [22]	81.27	67.41	19.97	6.27	81.46	68.32	18.74	6.58	80.19	65.87	23.61	7.41	82.16	68.04	17.56	4.82
General	CPS [6]	86.92	74.23	5.31	1.78	87.18	75.14	4.92	1.84	86.27	72.68	6.38	2.12	87.31	74.87	4.63	1.38
General	SSNet [21]	88.41	76.42	6.21	1.37	88.67	77.31	5.78	1.42	87.83	74.96	7.21	1.59	88.73	76.99	5.64	1.10
General	Co-BioNet [15]	87.66	78.61	1.98	0.70	87.94	79.47	1.84	0.73	87.12	77.28	2.31	0.84	87.92	79.08	1.79	0.53
General	BCP [1]	90.27	82.19	3.89	1.20	90.53	83.08	3.61	1.26	89.71	80.87	4.52	1.41	90.57	82.62	3.54	0.93
General	ABD [7]	88.90	83.17	1.44	0.50	89.12	84.03	1.37	0.52	88.34	81.91	1.68	0.59	89.24	83.57	1.27	0.39
General	β-FFT [9]	90.46	82.29	2.41	0.61	90.72	83.18	2.24	0.64	89.92	81.03	2.89	0.72	90.74	82.66	2.10	0.47
Imbalance	CReST [20]	79.84	71.62	13.15	3.99	80.03	72.51	12.38	4.17	79.27	70.18	15.42	4.68	80.22	72.17	11.65	3.12
Imbalance	SimiS [5]	77.89	75.84	7.68	2.35	78.14	76.72	7.12	2.44	77.31	74.56	8.96	2.71	78.22	76.24	6.96	1.90
Imbalance	Basak *et al.* [2]	83.36	69.38	8.09	2.29	83.58	70.24	7.56	2.38	82.74	67.92	9.38	2.62	83.76	69.98	7.33	1.87
Imbalance	CLD [12]	84.97	78.84	9.43	2.54	85.21	79.71	8.82	2.63	84.38	77.47	10.84	2.89	85.32	79.34	8.63	2.10
Imbalance	DHC [18]	87.46	74.45	8.97	2.57	87.72	75.32	8.34	2.67	86.88	73.08	10.21	2.91	87.78	74.95	8.36	2.13
Imbalance	GA [16]	89.21	82.41	3.27	1.03	89.47	83.28	3.04	1.08	88.64	81.12	3.84	1.21	89.52	82.83	2.93	0.80
Imbalance	SKCDF [23]	87.74	81.01	3.94	1.10	87.98	81.88	3.67	1.15	87.18	79.74	4.48	1.29	88.06	81.41	3.67	0.86
Imbalance	**DCA (Ours)**	**91.02**	**83.67**	**1.41**	**0.54**	**91.28**	**84.53**	**1.31**	**0.57**	**90.47**	**82.41**	**1.72**	**0.64**	**91.31**	**84.07**	**1.20**	**0.41**

Table 2. Quantitative comparison between our approach and SOTA methods on **10%** labeled **ACDC** dataset. 'General' or 'Imbalance' indicate whether the methods consider class-imbalance issue or not. The top-3 results are highlighted as first , second , and third .

multi-organ junction indicated by the yellow ellipses, comparison methods (BCP, ABD, β-FFT, DHC, GA) erroneously misclassify liver regions as stomach, while SKCDF produces false positives for the spleen. Neither general nor imbalance-aware SSMIS methods effectively delineate boundaries in such clustered regions, whereas our method achieves the most accurate boundary distinction.

The t-SNE visualization of extracted features in Figure 4 further supports these findings. Our method produces more compact and discriminative feature clusters. For instance, the clusters for LAG and RAG are entirely distinct from other categories, whereas they exhibit varying degrees of entanglement and overlap in other methods.

3. More experiments results on AMOS dataset

Table 4 reports the results on the AMOS [10] dataset with 2% labeled data. Our method once again demonstrates dominant performance, exceeding the second-best method (GA) by 6.03% in average Dice. It achieves the best segmentation performance in 13 out of 15 categories. Remarkably, our approach strikes an excellent balance between segmenting large organs (e.g., St, Sp, LK, RK) and small organs (e.g., Es, LAG, RAG), achieving top performance in both groups.

In the qualitative results shown in Figure 5 (axial plane, red ellipses), BCP, beta-FFT, and SKCDF fail to segment the pancreas, while ABD, DHC, and GA capture only a small portion. All comparison methods fail to generate complete segmentations for the portal & splenic veins (PSV), whereas our result is relatively closest to the Ground Truth. In the coronal plane (yellow ellipses), which depicts the intersection of the pancreas, PSV, and left adrenal gland (LAG), our method successfully delineates all three organs, despite some imperfections in the PSV region. In contrast, none of the comparison methods can simultaneously segment these three structures, demonstrating our substantial lead in such challenging anatomical regions.

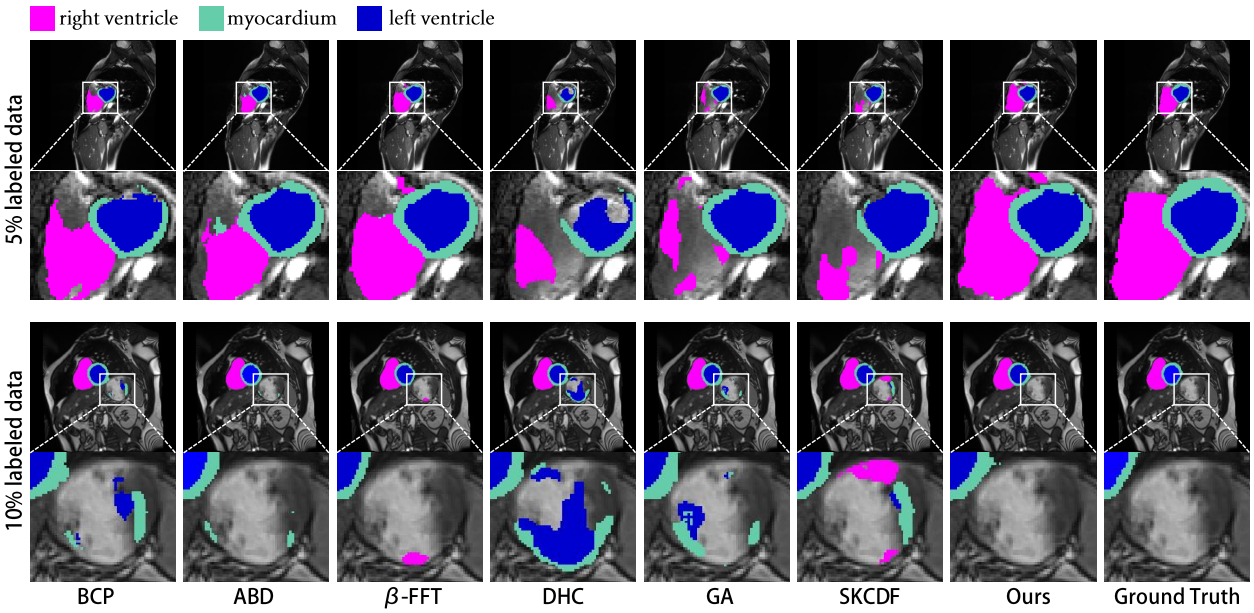
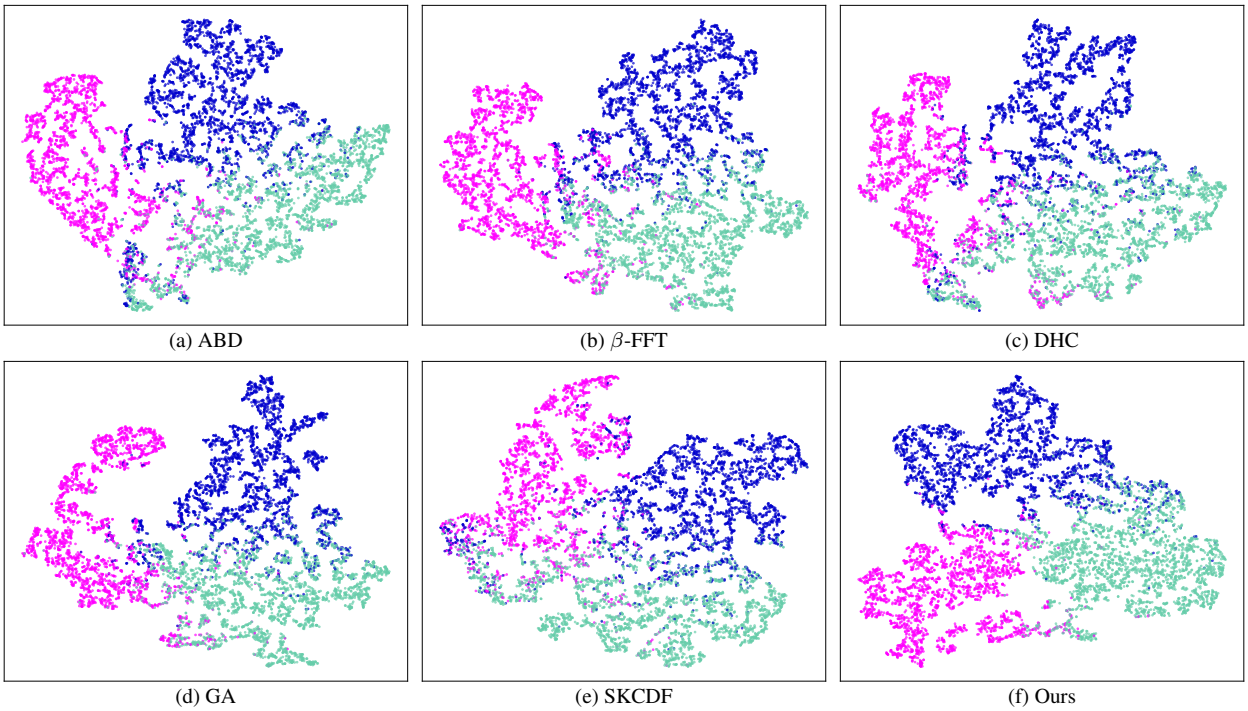


Figure 1. Qualitative comparison between our method and the SOTA methods on **5%** and **10%** labeled **ACDC** dataset.



(a) ABD (b) β-FFT (c) DHC

(d) GA (e) SKCDF (f) Ours

Figure 2. The t-SNE visualizations of feature representation extracted from SOTA methods and our method on **10%** labeled **ACDC** dataset. ● RV, ● Myo and ● LV.

The t-SNE visualization in Figure 6 confirms that the class discriminability of our method is significantly better than that of (a) MagicNet, (b) ABD, (c) beta-FFT, and (e) SKCDF. While (d) GA appears closest to our method, close inspection reveals that the overlap between the Right Kidney (RK) and Left Kidney (LK) clusters is noticeably smaller in our method, indicating superior feature separation.

	Methods	Avg. Dice	Avg. ASD	Average Dice of Each Class												
				Sp	RK	LK	Ga	Es	Li	St	Ao	IVC	PSV	Pa	RAG	LAG
	V-Net [14] (fully)	62.09±1.2	10.28±3.9	84.6	77.2	73.8	73.3	38.2	94.6	68.4	72.1	71.2	58.2	48.5	17.9	29.0
General	UA-MT [22]	18.07±1.2	57.64±1.8	27.1	7.1	17.0	24.4	0.0	80.6	15.6	39.3	16.7	4.4	2.7	0.0	0.0
General	URPC [13]	26.37±1.5	53.95±11.3	51.7	35.1	26.4	7.3	0.0	83.8	21.3	69.0	41.0	1.9	5.2	0.0	0.0
General	CPS [6]	21.96±1.2	55.42±4.6	37.9	31.8	19.0	31.9	0.0	65.1	15.5	44.8	29.6	4.3	5.5	0.0	0.0
General	SS-Net [21]	17.50±3.0	66.17±8.0	45.6	11.6	42.3	2.4	0.0	74.5	6.0	32.6	2.8	0.0	0.0	3.8	5.8
General	Co-BioNet [15]	40.84±2.5	11.78±2.3	59.5	68.6	52.5	6.0	30.0	91.1	41.4	72.0	48.6	13.6	9.0	10.6	28.1
General	BCP [1]	32.67±3.1	22.46±3.7	52.3	58.4	44.8	4.8	18.7	89.6	32.1	62.8	38.9	9.4	6.3	3.1	7.2
General	MagicNet [4]	54.39±2.9	19.57±6.4	73.0	83.8	82.3	13.2	0.0	90.9	63.2	78.3	69.4	47.1	35.4	23.7	46.7
General	ABD [7]	37.24±2.8	18.93±3.2	57.6	63.1	49.2	7.3	24.6	90.8	37.5	68.4	43.7	11.8	8.4	6.7	12.3
General	β-FFT [9]	41.35±2.3	15.86±2.6	61.8	72.4	56.3	12.4	33.8	92.7	45.6	74.3	52.1	18.9	14.2	12.8	17.6
Imbalance	Adsh [8]	22.80±0.9	46.18±4.0	36.0	35.7	20.0	31.0	0.0	74.7	18.3	32.3	27.8	11.7	7.3	1.7	0.0
Imbalance	CReST [20]	26.56±2.9	36.17±1.0	37.3	46.5	25.2	27.1	1.7	66.3	14.2	45.2	35.8	11.2	6.8	24.2	3.8
Imbalance	SimiS [5]	25.05±3.1	43.93±2.4	42.0	38.6	27.2	19.7	0.0	74.2	16.5	51.7	35.0	13.6	5.4	0.0	1.8
Imbalance	Basak *et al.* [2]	25.30±2.2	50.02±5.7	40.9	42.3	19.2	35.2	0.0	75.7	19.2	44.7	32.8	5.0	10.4	3.5	0.0
Imbalance	CLD [12]	22.49±1.6	49.74±4.1	39.3	43.9	25.6	12.8	0.0	73.3	14.3	41.1	25.7	8.8	6.1	0.2	1.1
Imbalance	DHC [18]	31.64±0.9	21.82±1.0	45.1	47.4	33.1	36.6	7.1	71.4	17.8	58.9	34.4	16.5	9.3	21.8	12.0
Imbalance	A&D [19]	46.24±0.8	7.78±2.1	79.0	67.0	58.4	25.7	12.0	86.3	30.3	77.8	65.5	18.2	14.2	42.2	24.4
Imbalance	GA [16]	57.46±1.3	5.62±0.7	69.8	85.8	83.1	10.2	49.9	90.6	60.7	76.4	69.2	41.8	32.0	29.3	48.1
Imbalance	SKCDF [23]	48.45±0.6	7.87±3.5	73.5	66.7	64.2	24.4	27.2	90.0	30.3	79.0	68.5	33.7	18.1	37.3	17.1
Imbalance	**DCA (Ours)**	**62.18**±1.1	**3.94**±0.5	**78.3**	**89.6**	**88.7**	**36.5**	**58.2**	**92.4**	**66.1**	**84.7**	**77.9**	**59.4**	**46.8**	**42.7**	**55.6**

Table 3. Quantitative comparison between our approach and SOTA methods on **10%** labeled **Synapse** dataset. 'General' or 'Imbalance' indicate whether the methods consider class-imbalance issue or not. We report the $mean \pm std$ repeated three times. The top-3 results are highlighted as first, second, and third.

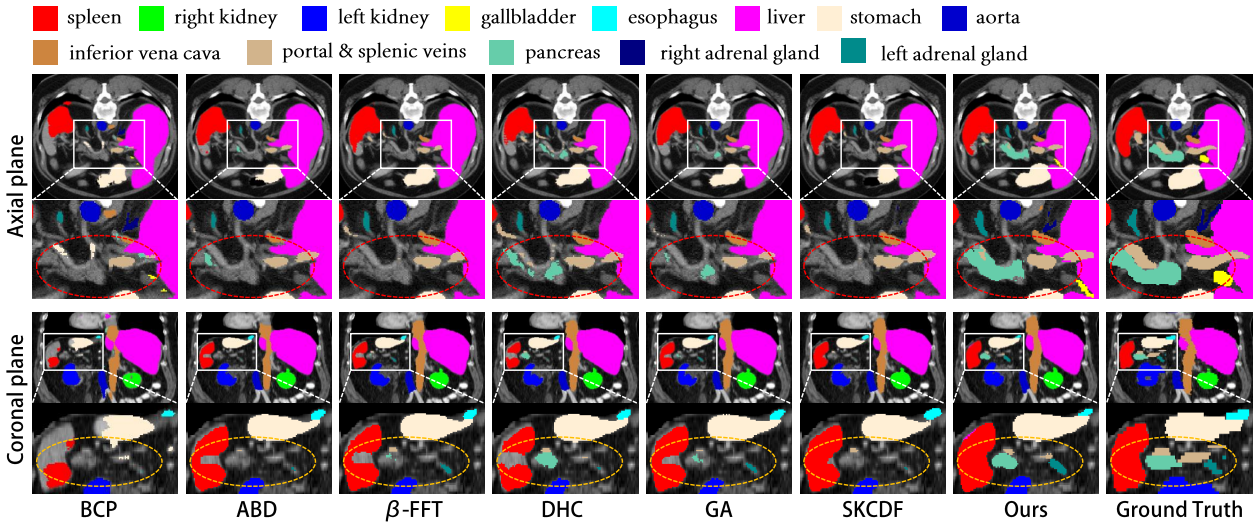


Figure 3. Qualitative comparison between our method and the SOTA methods on **10%** labeled **Synapse** dataset.

References

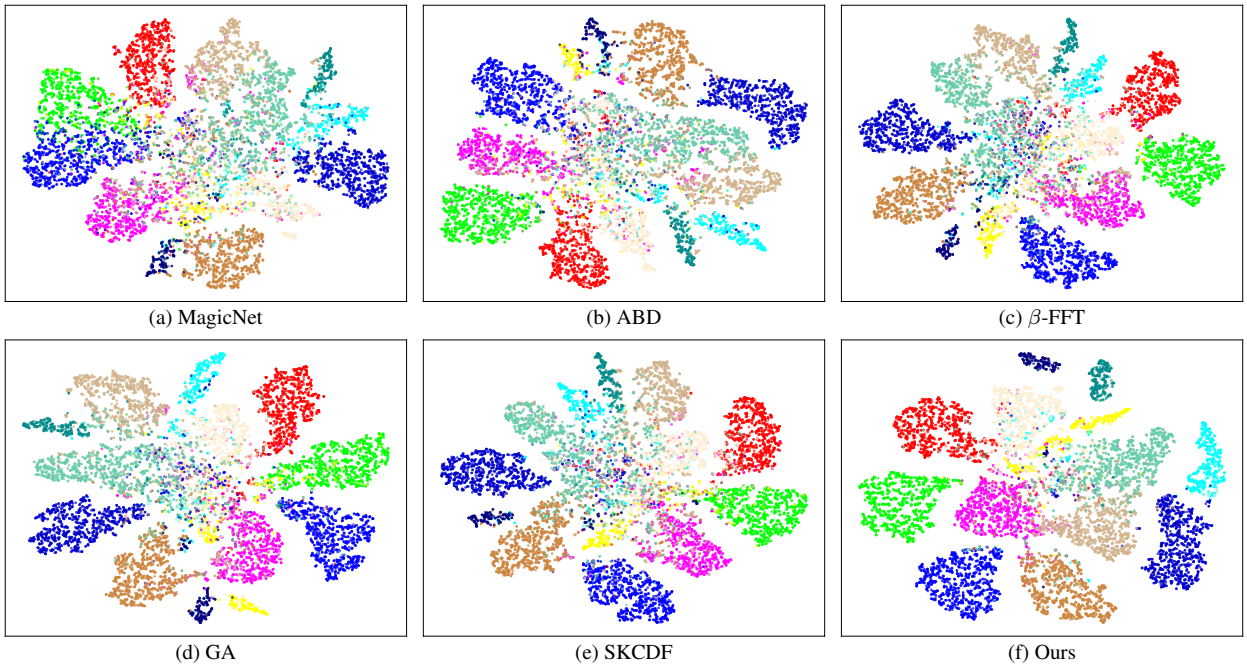
[1] Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11514–11524, 2023. 1, 2, 4, 5

[2] Hritam Basak, Sagnik Ghosal, and Ram Sarkar. Addressing class imbalance in semi-supervised image segmentation: A study on cardiac mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 224–233. Springer, 2022. 2, 4, 5

[3] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018. 1

[4] Duowen Chen, Yunhao Bai, Wei Shen, Qingli Li, Lequan Yu, and Yan Wang. Magicnet: Semi-supervised multi-organ segmentation via magic-cube partition and recovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23869–23878, 2023. 4, 5

[5] Hao Chen, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Marios Savvides, and Bhiksha Raj. An embarrassingly simple baseline for imbalanced semi-supervised learning. *arXiv preprint arXiv:2211.11086*, 2022. 2, 4, 5



(a) MagicNet (b) ABD (c) β-FFT

(d) GA (e) SKCDF (f) Ours

Figure 4. The t-SNE visualizations of feature representation extracted from SOTA methods and our method on **10%** labeled **Synapse** dataset. ● Sp, ● RK, ● LK, ● Ga, ● Es, ● Li, ● St, ● Ao, ● IVC, ● Psv, ● Pa, ● RAG and ● LAG.

	Methods	Avg. Dice	Avg. ASD	Average Dice of Each Class														
				Sp	RK	LK	Ga	Es	Li	St	Ao	IVC	Pa	RAG	LAG	Du	Bl	P/U
	V-Net [14] (fully)	76.50	2.01	92.2	92.2	93.3	65.5	70.3	95.3	82.4	91.4	85.0	74.9	58.6	58.1	65.6	64.4	58.3
General	UA-MT [22]	33.96	22.43	62.5	61.7	59.8	17.5	13.8	73.4	39.4	34.6	32.4	26.5	12.1	6.5	15.3	32.4	21.7
	URPC [13]	38.39	37.58	60.8	57.7	56.5	34.6	0.0	78.4	41.4	53.3	49.6	40.4	0.0	0.0	30.1	42.5	30.6
	CPS [6]	31.78	39.23	55.9	46.9	53.1	27.7	0.0	66.4	25.2	41.8	45.2	29.4	0.1	0.0	22.1	38.7	24.2
	SS-Net [21]	17.47	59.05	37.7	20.1	26.3	9.0	3.3	57.1	25.1	28.4	28.2	0.0	0.0	0.0	0.0	26.5	0.2
	Co-BioNet [15]	42.82	31.98	68.0	55.5	54.7	40.5	32.9	75.8	41.8	56.5	50.8	27.5	0.0	20.2	19.1	52.9	46.2
	BCP [1]	37.14	36.25	58.3	48.6	47.2	35.8	15.4	71.2	36.9	50.8	44.1	23.8	18.6	12.7	16.5	31.8	38.4
	MagicNet [4]	47.29	35.14	69.4	68.4	70.3	46.7	0.0	82.7	55.0	67.3	63.3	53.8	0.0	0.0	36.9	60.2	35.4
	ABD [7]	44.57	29.86	64.8	52.7	51.9	40.1	22.6	78.5	40.3	58.9	49.6	30.1	24.3	18.9	24.7	39.6	43.2
	β-FFT [9]	49.36	24.18	70.1	60.8	59.4	44.7	29.8	83.6	46.2	65.3	56.7	36.4	31.8	26.5	33.9	46.2	49.7
Imbalance	Adsh [8]	30.30	42.48	53.9	45.1	51.2	28.5	0.0	62.1	27.0	41.4	42.7	25.0	0.0	0.0	20.3	35.8	21.6
	CReST [20]	34.13	20.15	57.9	51.5	49.1	22.7	13.2	66.2	34.4	39.4	40.4	24.6	17.2	10.2	24.4	36.5	24.4
	SimiS [5]	36.89	26.16	57.8	58.6	58.6	22.9	0.0	70.9	38.0	52.0	47.0	32.4	20.2	11.5	18.1	39.9	25.5
	Basak *et al.* [2]	29.87	35.55	50.7	47.7	44.1	21.1	0.0	61.8	27.7	38.1	40.4	21.8	9.6	9.5	14.6	36.5	24.5
	CLD [12]	36.23	27.63	55.8	55.8	59.1	23.9	0.0	69.9	38.2	50.1	44.5	32.3	18.9	9.2	18.8	42.2	24.9
	DHC [18]	38.28	20.34	62.1	59.5	57.8	25.0	20.5	66.0	38.2	51.3	47.9	26.8	26.4	7.0	17.8	43.2	24.8
	A&D [19]	34.56	21.05	44.1	55.0	49.3	22.5	18.8	57.0	36.9	44.1	47.2	33.7	16.1	12.1	32.5	29.6	19.4
	GA [16]	53.85	12.20	72.1	68.0	72.4	44.6	42.7	82.7	48.1	66.3	61.3	49.5	44.9	30.4	31.6	56.9	36.2
	SKCDF [23]	41.80	16.52	62.0	65.5	59.8	26.9	25.9	70.4	38.5	59.0	51.6	36.7	27.1	15.4	23.7	41.1	23.4
	DCA (Ours)	**59.88**	**4.97**	**72.2**	**72.1**	**73.0**	**50.5**	**49.1**	**81.4**	**57.0**	**73.4**	**66.7**	**56.9**	**48.8**	**47.1**	**44.4**	**61.9**	**43.8**

Table 4. Quantitative comparison between our approach and SOTA methods on **2%** labeled **AMOS** dataset. 'General' or 'Imbalance' indicate whether the methods consider class-imbalance issue or not. The top-3 results are highlighted as first, second, and third.

[6] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2613–2622, 2021. 2, 4, 5

[7] Hanyang Chi, Jian Pang, Bingfeng Zhang, and Weifeng Liu. Adaptive bidirectional displacement for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4070–4080, 2024. 1, 2, 4, 5

[8] Lan-Zhe Guo and Yu-Feng Li. Class-imbalanced semi-supervised learning with adaptive thresholding. In *International conference on machine learning*, pages 8082–8094. PMLR, 2022. 4, 5

[9] Ming Hu, Jianfu Yin, Zhuangzhuang Ma, Jianheng Ma, Feiyu Zhu, Bingbing Wu, Ya Wen, Meng Wu, Cong Hu, Bingliang Hu, et al. beta-fft: Nonlinear interpolation and differentiated training strategies for semi-supervised

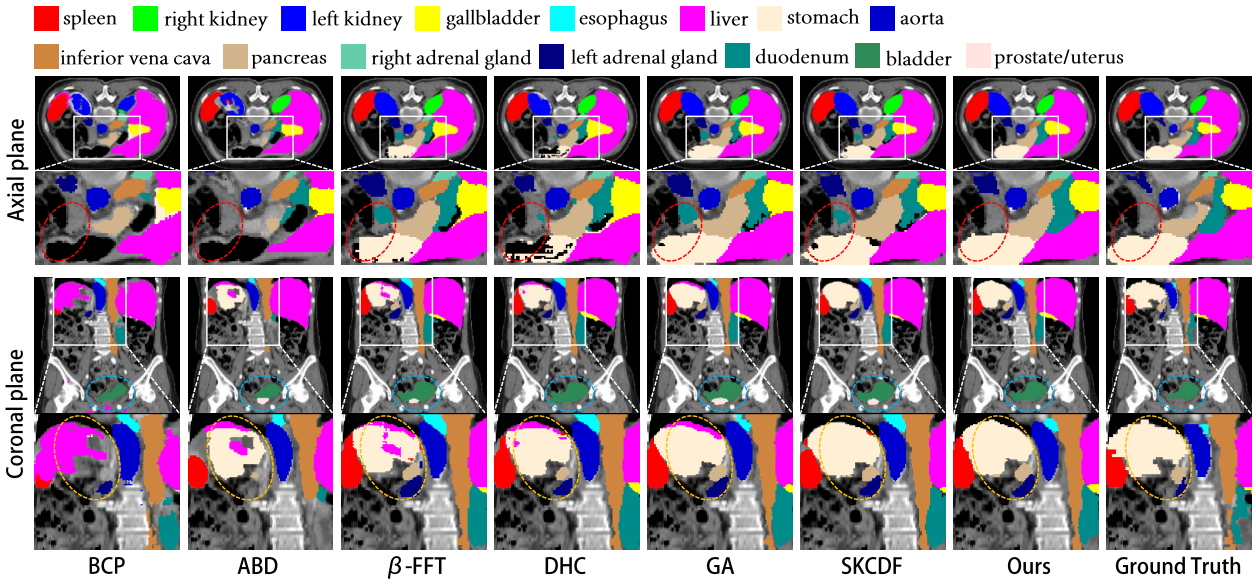
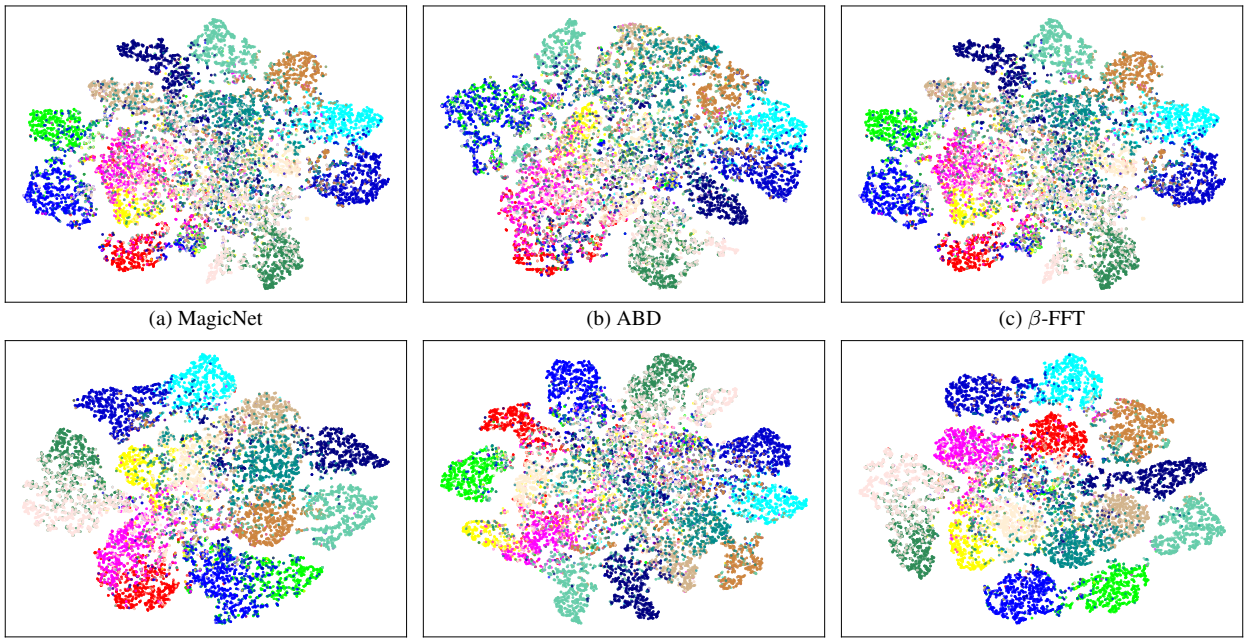


Figure 5. Qualitative comparison between our method and the SOTA methods on **2%** labeled **AMOS** dataset.



(a) MagicNet (b) ABD (c) β-FFT

(d) GA (e) SKCDF (f) Ours

Figure 6. The t-SNE visualizations of feature representation extracted from SOTA methods and our method on **5%** labeled **AMOS** dataset. ● Sp, ● RK, ● LK, ● Ga, ● Es, ● Li, ● St, ● Ao, ● IVC, ● Pa, ● RAG, ● LAG, ● Du, ● Bl and ● P/U.

medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30839–30849, 2025. 1, 2, 4, 5

[10] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022. 2

[11] Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, page 12. Munich, Germany, 2015. 1

[12] Yiqun Lin, Huifeng Yao, Zezhong Li, Guoyan Zheng, and Xiaomeng Li. Calibrating label distribution for class-imbalanced barely-supervised knee segmentation.

In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–118. Springer, 2022. 2, 4, 5

[13] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 318–329. Springer, 2021. 4, 5

[14] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 4, 5

[15] Himashi Peiris, Munawar Hayat, Zhaolin Chen, Gary Egan, and Mehrtash Harandi. Uncertainty-guided dual-views for semi-supervised volumetric medical image segmentation. *Nature Machine Intelligence*, 5(7):724–738, 2023. 2, 4, 5

[16] Wenbo Qi, Jiafei Wu, and SC Chan. Gradient-aware for class-imbalanced semi-supervised medical image segmentation. In *European Conference on Computer Vision*, pages 473–490. Springer, 2024. 2, 4, 5

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 2

[18] Haonan Wang and Xiaomeng Li. Dhc: Dual-debiased heterogeneous co-training framework for class-imbalanced semi-supervised medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 582–591. Springer, 2023. 2, 4, 5

[19] Haonan Wang and Xiaomeng Li. Towards generic semi-supervised framework for volumetric medical image segmentation. *Advances in Neural Information Processing Systems*, 36:1833–1848, 2023. 4, 5

[20] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10857–10866, 2021. 2, 4, 5

[21] Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 34–43. Springer, 2022. 2, 4, 5

[22] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 605–613. Springer, 2019. 2, 4, 5

[23] Zheng Zhang, Guanchun Yin, Bo Zhang, Wu Liu, Xiuzhuang Zhou, and Wendong Wang. A semantic knowledge complementarity based decoupling framework for semi-supervised class-imbalanced medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25940–25949, 2025. 2, 4, 5