Keep Your Friends Close, and Your Enemies Farther: Distance-aware Voxel-wise Contrastive Learning for Semi-supervised Multi-organ Segmentation

Supplementary Material

The supplementary material is organized as follows:

- **§A:** Additional details on the training datasets.
- **§B:** Additional details on the data preprocessing and implementation.
- **§C:** Detailed algorithm.
- **§D:** More experimental results.
- **§E:** More qualitative visualization.
- **§F:** More experiment results on ablation study for hyper-parameters.
- **§G:** Additional details on the complete derivation of the formula.

A. Datasets

FLARE 2022. This dataset is from the MICCAI 2022 Challenge Fast and Low GPU memory Abdominal Organ Segmentation [12] and comprises 70 3D CT volumes accompanied by a corresponding ground-truth mask. In addition, it includes 2000 unlabeled 3D CT volumes. The dataset is designed to segment 13 abdominal organs: the liver (Liv), spleen (Spl), pancreas (Pan), right kidney (R.kid), left kidney (L.kid), stomach (Sto), gallbladder (Gal), esophagus (Eso), aorta (Aor), inferior vena cava (IVC), right adrenal gland (RAG), left adrenal gland (LAG), and duodenum (Duo). For dataset partitioning, we divide the labeled data into training, validation, and test sets at a ratio of 6:2:2. In our experiments, we incorporate the unlabeled data into the training set, resulting in labeled data proportions of 50% (42 labeled cases and 42 unlabeled cases) and 10% (42 labeled cases and 378 unlabeled cases) within two training sets.

AMOS. This dataset is from the Multi-Modality Abdominal Multi-Organ Segmentation Challenge 2022 [6]. The AMOS dataset is comprised of 300 CT images, which are annotated at the pixel level for 15 distinct abdominal organs, including two additional organs not found in the FLARE 2022 dataset: the bladder (Bl) and prostate/uterus (P/U). In our experiments, we divide the dataset into training, validation, and test sets using a ratio of 6:2:2. For the training set, the proportion of labeled data is set to 10% and 50%.

MMWHS. This dataset is from the Multi-Modality Whole Heart Segmentation Challenge 2017 [19]. The dataset contains 20 3D CT volumes with corresponding annotations of seven cardiac structures, *i.e.,* the myocardium of the left ventricle (MYO), left atrium blood cavity (LAC), right atrium blood cavity (RAC), left ventricle blood cavity (LVC), right ventricle blood cavity (RVC), ascending aorta (AA), and pulmonary aorta (PA). In our experiments, we divide the dataset into training, validation, and test sets at a ratio of 6:2:2. For the training set, the proportion of labeled data is set to 10% and 50%.

BTCV. This dataset is from MICCAI Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge [7] which contains 30 CT images with 13 organs annotation. In contrast to the FLARE 2022 dataset, the Duo has been replaced with the portal vein and splenic vein (P&S). In our experiments, we divide the dataset into training, validation, and test sets using a ratio of 6:2:2. For the training set, the proportion of labeled data is set to 10% and 50%.

B. Data Preprocessing and Implementation Details

Data Preprocessing. We apply hierarchical steps for data preprocessing before network training. 1) the orientation of all CT scans is standardized in the left-posterior-inferior (LPI) direction. 2) The voxel values are clipped into the range of [–325, 325] Hounsfield units (HU) to enhance the contrast of the foreground organs and suppress background interference. 3) The voxel spacing is standardized to [1.25, 1.25, 2.5]. 4) Min-max normalization is implemented using the formula $(x - x_{0.5})/(x_{99.5} - x_{0.5})$, where $x_{0.5}$ and $x_{99.5}$ represent the 0.5th and 99.5th percentiles of x, respectively.

Implementation Details. For network training, we use the stochastic gradient descent (SGD) optimizer with a weight decay of 0.0001 and momentum of 0.9 [11]. To mitigate the overfitting problem, we apply basic data augmentation, including random cropping and flipping [14]. The model uses randomly cropped patches as inputs, with a patch size of $64 \times 128 \times 128$. During network training, we adopt a polynomial learning rate policy that scales the initial learning rate by $(1 - \frac{\textit{iteration}}{\textit{max\_iteration}})^{0.9}$ [1], where *iteration* and *max_iteration* represent the current iteration and maximum number of iterations, respectively. For the MMWHS dataset, the segmentation model is trained with a batch size of 4, comprising an equal split of 2 labeled and 2 unlabeled instances, over a total of 10,000 iterations with an initial learning rate of 0.03. For the FLARE 2022, AMOS and BTCV datasets, we opt for a batch size of 8, including 4 labeled and 4 unlabeled instances, and conduct training for 20,000 iterations with an initial learning rate of 0.1. Note that, for all the evaluated methods, we make no additional modifications during the training process for fair evaluations. During inference, the final volumetric segmentation is generated using a sliding-window strategy. The stride used is $32 \times 80 \times 80$, and the sliding-window approach employs a patch size of $64 \times 160 \times 160$.

Algorithm 1: Training Procedure of Our Framework

input : Segmentation networks $\{f(\cdot;\theta_i)\}$, maximum epoch $E_{\max}$, batchsize $\mathcal{B}$, labeled training data and their ground-truth masks $\mathcal{S}_l = \{(x_i, y_i)_{i=1}^N\}$, unlabeled training data $\mathcal{S}_u = \{(x_i)_{i=1}^M\}$.

output: Trained weights of *model A* $f(\cdot;\theta_A)$ and *model B* $f(\cdot;\theta_B)$;

Initialize network parameters of *model A* $f(\cdot;\theta_A)$ and *model B* $f(\cdot;\theta_B)$;
for *epoch*$\in [1, E_{\max}]$ **do**
  for *batch* $\mathcal{B}$ **do**
    Get probabilities: $p_i^A \leftarrow$ segmentation head $A(x_i), p_i^B \leftarrow$ segmentation head $B(x_i)$;
    Get features: $\mathbf{r}_i^A \leftarrow$ feature head $A(x_i), \mathbf{r}_i^B \leftarrow$ feature head $B(x_i)$;
    for $x_i \in \mathcal{S}_l$ **do**
      Calculate $\mathcal{L}_{sup} \leftarrow \{p_i^A, p_i^B, y_i\}$; `// Compute supervised loss`
    end
    for $x_i \in \mathcal{S}_u$ **do**
      Get pseudo-labels: $\hat{y}_i^A \leftarrow \arg\max(p_i^A), \hat{y}_i^B \leftarrow \arg\max(p_i^B)$;
      Get reliable pseudo-labels $M_r^A(x_i)$ and $M_r^B(x_i)$;
      Get unreliable pseudo-labels $M_u^A(x_i)$ and $M_u^B(x_i)$;
      Calculate $\mathcal{L}_{cps} \leftarrow \{\hat{y}_i^A, \hat{y}_i^B, p_i^A, p_i^B\}$; `// Compute CPS loss`
    end
    Initialize $\mathcal{L}_{contra} \leftarrow 0$, $\mathcal{L}_{hqcl} \leftarrow 0$, $\mathcal{L}_{dvcl} \leftarrow 0$;
    for $x_i \in \mathcal{S}_l \cup \mathcal{S}_u$ **do**
      for $x_i \in \mathcal{S}_l \cup M_r^A(x_i) \cup M_r^B(x_i)$ **do**
        for $c \leftarrow 1$ **to** C **do**
          Get feature sets of anchor voxels $\mathcal{R}_c$;
          Random sample M anchor voxels: $\{\mathbf{r}_m\}_{m=1}^M \in \mathcal{R}_c$;
          Get positive candidates $\mathcal{P}_c$;
          Get feature sets of negative candidates $\mathcal{G}_c = \mathcal{G}_c^{Al} \bigcup \mathcal{G}_c^{Bl}$;
          Random sample N negative candidates: $\{\mathbf{r}_n\}_{n=1}^N \in \mathcal{G}_c$;
          Calculate $\mathcal{L} \leftarrow \{\{\mathbf{r}_m\}_{m=1}^M, \{\mathbf{r}_n\}_{n=1}^N, \mathcal{P}_c\}$;`// Compute HqCL loss`
          $\mathcal{L}_{hqcl} \leftarrow \mathcal{L}_{hqcl} + \mathcal{L}$;
        end
      end
      for $x_i \in M_u^A(x_i) \cup M_u^B(x_i)$ **do**
        Random sample M' query voxels: $\{x_i\}_{i=1}^{M'}$;
        Exclude potential outliers; `// QDS`
        For each query voxel x_i, calculate $\{\mathcal{O}_{i,j}\}_{j=1}^{|M_u^A(x_i) \cup M_u^B(x_i) - 1|}$;
        Get the close neighbor set $\mathcal{C}_i^K$;
        Get the distant outsider set $\mathcal{D}_i^{K'}$;
        Calculate the affinity value $\mathcal{A}_{i,j}$ of its j-th neighbor;`// NPS`
        Calculate $\mathcal{L}_{dvcl} \leftarrow \{\mathbf{r}_i, \mathcal{A}_{i,j}, \mathcal{C}_i, \mathcal{D}_i\}$; `// Compute DVCL loss`
      end
    end
    Calculate $\mathcal{L}_{contra} \leftarrow \mathcal{L}_{hqcl} + \beta\mathcal{L}_{dvcl}$;
    Update *model A* $f(\cdot;\theta_A)$ and *model B* $f(\cdot;\theta_B)$ by decreasing the stochastic gradient on $\mathcal{L}_{sup}$, $\mathcal{L}_{cps}$, and $\mathcal{L}_{contra}$; `// Update network parameters`
  end
  epoch= *epoch*+1;
end

	Methods	Mean Dice for each organ													Mean Dice	Mean Jaccard
		Liv	Spl	Sto	L.kid	R.kid	Aor	Pan	IVC	Duo	Gal	Eso	RAG	LAG		
	Fully	94.21	88.32	49.40	91.68	91.33	89.89	48.83	79.24	52.10	56.24	61.13	44.87	42.12	68.41±0.58	56.97±0.42
	50% labeled data (labeled:unlabeled=42:42)															
w/o VCL	DAN [18]	96.50	89.74	62.04	93.63	93.24	90.76	61.71	80.09	66.60	70.07	67.20	58.04	43.47	74.86±0.69	63.73±0.62
	MT [13]	96.97	89.64	66.63	93.66	92.58	91.39	68.65	82.08	60.96	69.89	71.68	57.51	62.21	77.22±0.42	65.94±0.83
	UA-MT [17]	**97.21**	88.85	71.66	94.00	93.50	92.41	70.60	82.92	64.85	76.82	72.05	60.02	60.85	78.91±0.89	68.01±1.21
	SASSnet [8]	95.25	92.03	66.48	92.47	93.79	90.03	63.61	79.94	60.14	65.57	70.83	59.98	62.78	76.38±0.61	65.20±0.60
	DTC [10]	96.47	91.33	65.94	94.46	93.57	92.52	64.88	83.77	65.58	75.80	68.53	68.87	61.35	78.70±0.79	67.64±1.05
	CPS [2]	96.77	91.23	72.63	93.38	93.70	92.35	70.34	83.30	65.48	78.81	72.37	58.34	61.47	79.24±0.56	67.99±0.56
	CLD [9]	94.27	88.54	74.88	91.48	93.33	91.51	71.51	83.21	68.15	76.68	71.03	64.57	63.98	79.47±0.27	67.95±0.28
	DHC [14]	92.54	90.81	76.96	93.39	92.18	91.84	74.65	83.25	69.44	84.60	72.91	64.52	55.88	80.23±1.07	68.90±0.26
	MagicNet [1]	96.49	88.84	80.33	90.84	93.06	91.72	69.46	81.98	67.44	82.81	**75.79**	63.40	59.98	80.16±0.33	69.08±0.12
w/ VCL	UGPCL [15]	96.00	90.63	79.75	**94.88**	94.01	90.86	73.39	82.26	66.28	81.02	71.42	61.81	65.44	80.59±0.85	69.76±0.85
	U^{2}PL [16]	96.83	90.67	77.19	94.25	**94.33**	**92.73**	74.36	**84.22**	68.37	82.99	75.57	66.83	67.27	81.97±0.31	71.69±0.31
	BaCon [5]	95.54	90.81	76.89	93.78	93.49	91.71	74.68	83.77	68.67	83.33	73.18	68.80	66.54	81.63±0.56	70.82±0.57
	CCL [4]	96.25	91.65	76.14	93.04	93.48	92.15	73.44	83.19	67.35	82.49	75.54	68.05	64.39	81.32±0.32	70.68±0.51
	Ours	96.63	**94.24**	**84.65**	93.87	93.64	91.85	**76.00**	84.03	**69.96**	**87.44**	74.75	**71.16**	**69.49**	**83.67±0.35**	**73.11±0.72**
	10% labeled data (labeled:unlabeled=42:420)															
w/o VCL	DAN [18]	95.89	84.15	67.29	92.81	91.96	91.35	63.12	79.33	66.48	77.29	67.82	50.41	48.41	75.10±0.69	63.83±0.23
	MT [13]	96.49	91.54	74.64	93.78	92.77	92.17	69.23	82.71	66.68	73.49	70.66	61.88	41.26	77.49±0.48	66.45±0.39
	UA-MT [17]	96.42	91.98	79.91	92.74	92.83	92.33	71.43	83.10	67.72	77.26	72.41	64.04	46.18	79.10±0.38	68.21±0.48
	SASSnet [8]	96.21	90.40	67.12	94.00	92.85	91.61	67.89	79.59	65.47	71.59	71.44	52.07	57.83	76.77±0.30	65.89±0.44
	DTC [10]	96.63	92.91	72.76	92.68	92.40	91.87	66.82	81.47	65.76	78.38	69.39	59.74	59.10	78.45±0.82	67.05±1.02
	CPS [2]	96.62	92.16	77.02	92.70	92.71	92.25	69.39	81.91	65.94	75.12	72.78	63.56	58.96	79.32±0.46	68.14±0.61
	CLD [9]	94.63	89.74	73.20	91.76	92.97	91.61	70.27	83.12	68.13	84.15	72.69	67.89	55.27	79.65±0.17	68.22±0.49
	DHC [14]	93.17	90.64	80.56	93.13	92.89	91.38	72.22	83.75	69.73	82.47	73.25	67.12	56.19	80.50±0.43	69.38±0.63
	MagicNet [1]	97.04	88.04	81.51	92.18	92.95	91.75	71.15	81.01	69.61	84.36	**77.07**	63.34	60.33	80.79±0.75	70.23±0.96
w/ VCL	UGPCL [15]	96.81	92.11	75.48	94.02	**94.79**	92.07	68.75	82.46	68.60	77.50	71.09	63.54	58.10	79.64±0.72	69.09±0.58
	U^{2}PL [16]	96.34	91.45	80.09	94.30	93.81	91.17	74.53	83.13	69.88	84.40	72.70	67.25	68.12	82.09±0.32	71.34±0.51
	BaCon [5]	95.88	90.35	77.17	94.35	93.98	92.10	73.32	83.75	69.67	81.15	74.87	69.44	69.55	81.97±0.33	71.04±0.40
	CCL [4]	95.77	91.44	76.81	92.71	93.54	92.02	73.59	83.58	67.17	81.19	75.45	70.15	69.99	81.80±0.11	70.91±0.32
	Ours	**97.19**	**93.37**	**84.39**	**95.22**	93.58	**92.71**	**77.03**	**84.72**	**71.98**	85.81	75.62	**70.97**	**71.38**	**84.15±0.07**	**73.62±0.29**

Table 1. Quantitative results (mean Dice for each organ, mean and SD of Dice, and mean and SD of Jaccard) on two settings of **FLARE 2022 dataset**. '*w/o VCL* 'or '*w/ VCL*' indicates whether the SSL methods combined with VCL or not. **Best results are boldfaced.**

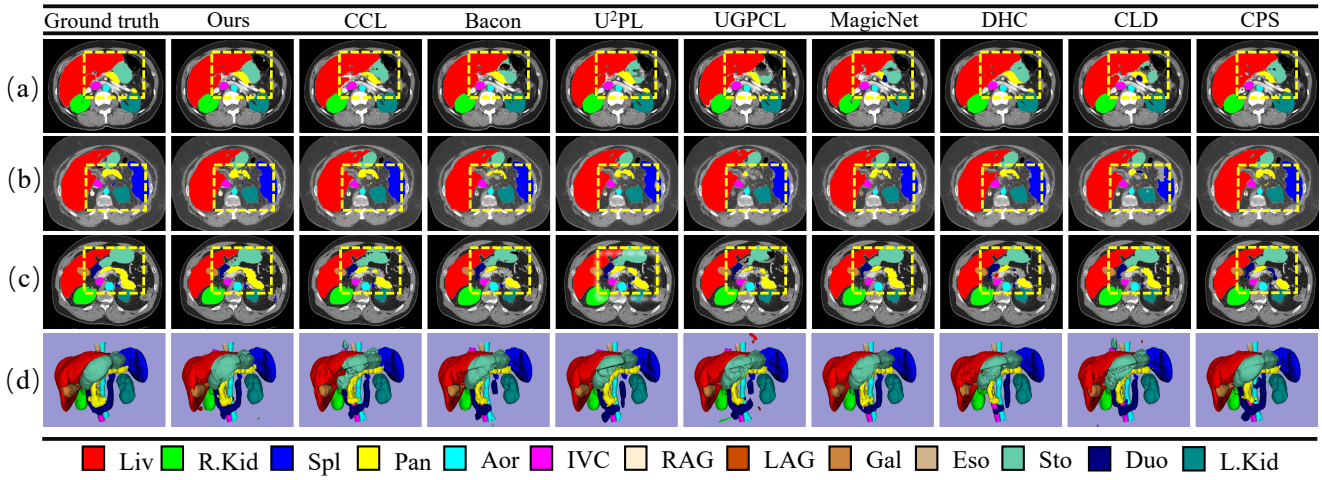

Figure 1. Visualization of the segmentation results for the **FLARE 2022 dataset**. (a-c) Segmentation results for one case of three transverse sections and (d) 3D segmentation views. As seen, better segmentation is achieved at the boundaries by our method, highlighted in yellow dashed boxes.

	Methods	Mean Dice for each organ														Mean Dice	Mean Jaccard
		Liv	Sto	Spl	L.kid	R.kid	Aor	Bla	IVC	Pan	Duo	P/U	Gal	Eso	RAG	LAG		
		10% labeled data (labeled:unlabeled=18:162)																
w/o VCL	3D U-Net [3]	85.99	40.08	82.19	79.57	80.24	75.51	11.63	59.86	26.80	20.22	22.40	15.06	37.05	32.46	9.68	45.24±0.65	39.06±0.42
	DAN [18]	86.54	50.00	83.98	86.97	85.76	85.46	54.81	67.26	48.97	43.84	52.23	33.08	40.28	28.00	18.09	57.80±0.88	47.38±0.89
	MT [13]	89.26	56.87	84.27	84.42	85.98	85.57	51.22	70.13	48.91	48.04	39.46	42.17	50.71	43.92	30.46	61.44±1.28	51.00±1.10
	UA-MT [17]	88.25	52.49	86.39	86.34	87.73	86.14	68.22	70.76	47.19	42.79	49.54	32.49	52.87	43.98	37.76	61.73±1.12	50.97±0.90
	SASSnet [8]	90.33	48.61	86.87	87.66	88.17	87.09	43.55	73.85	50.29	48.56	8.38	36.15	43.18	41.36	28.85	58.35±1.42	51.15±0.83
	DTC [10]	89.81	50.49	87.48	85.20	85.84	85.83	64.49	72.72	43.44	47.36	39.19	38.62	50.56	42.53	37.02	60.81±1.27	50.84±1.24
	CPS [2]	88.52	55.52	83.25	86.30	87.97	85.36	60.53	71.71	50.11	46.05	60.33	37.95	52.37	46.33	37.48	63.52±0.36	51.82±0.49
	CLD [9]	88.43	63.71	84.90	85.85	86.07	85.16	64.15	75.56	55.21	49.67	60.62	39.47	56.71	50.91	40.56	65.81±1.24	54.00±1.69
	DHC [14]	83.27	63.39	83.60	84.11	85.66	84.40	**74.52**	74.88	56.02	51.89	65.47	47.53	43.21	48.28	42.59	65.17±1.47	52.46±1.30
	MagicNet [1]	88.99	61.20	83.52	88.39	87.24	83.69	62.47	74.83	54.11	51.18	54.62	**56.69**	55.68	46.87	43.16	65.31±1.31	54.89±0.78
w/ VCL	UGPCL [15]	90.31	60.14	86.78	86.62	87.87	86.69	56.14	71.14	36.27	44.16	46.04	38.21	48.82	38.20	35.09	61.48±1.02	50.99±1.61
	U^{2}PL [16]	90.23	54.80	85.80	87.97	89.10	87.95	59.31	74.39	52.07	51.45	54.92	36.64	56.67	49.77	44.54	64.73±1.35	54.46±1.01
	BaCon [5]	89.84	56.57	86.36	89.41	88.78	88.35	38.96	73.62	50.98	44.82	50.47	43.36	58.67	50.81	42.81	64.40±1.50	54.61±0.72
	CCL [4]	90.58	55.51	85.13	88.01	89.11	88.33	62.34	73.72	50.73	45.13	58.34	40.11	60.09	45.33	33.20	64.00±0.16	53.33±0.70
	Ours	**91.02**	**71.31**	**89.83**	**89.94**	**89.41**	**88.96**	69.26	**78.26**	**56.34**	**57.21**	**66.11**	55.73	**64.84**	**53.66**	**45.30**	**71.36±0.56**	**59.76±0.65**
		50% labeled data (labeled:unlabeled=90:90)																
w/o VCL	3D U-Net [3]	89.25	55.60	84.23	87.40	88.58	87.32	53.49	73.71	48.56	48.21	52.68	38.43	50.27	38.48	30.30	61.29±1.74	51.62±1.35
	DAN [18]	90.49	55.91	89.63	90.08	88.74	86.71	47.44	72.09	54.98	50.33	53.04	39.13	58.34	29.57	6.49	61.39±1.16	52.06±1.45
	MT [13]	92.08	62.02	89.83	90.23	89.24	89.12	63.05	78.11	53.46	52.85	40.93	51.63	59.64	45.41	37.35	66.17±0.75	57.06±1.00
	UA-MT [17]	90.86	58.55	88.92	88.93	88.83	88.49	54.86	74.28	51.88	54.54	44.73	40.99	58.58	51.31	41.78	65.48±0.80	55.62±1.10
	SASSnet [8]	91.65	53.00	**91.54**	89.61	89.72	88.50	50.43	74.87	46.34	52.48	55.92	37.93	60.57	45.62	39.17	63.77±1.13	54.68±0.55
	DTC [10]	91.25	56.49	90.68	88.88	89.30	89.16	67.37	76.50	48.13	54.67	54.23	41.88	62.49	47.67	42.91	66.93±1.78	55.92±1.78
	CPS [2]	90.94	61.90	89.97	90.25	89.67	88.77	65.03	75.27	52.34	45.15	54.76	42.87	62.44	49.96	47.74	66.65±1.24	56.56±0.54
	CLD [9]	91.23	66.18	89.34	89.50	89.86	88.85	66.40	76.97	55.63	53.35	58.82	45.78	62.93	54.24	43.79	69.09±1.14	57.99±1.14
	DHC [14]	86.68	58.39	86.62	85.57	87.48	87.28	67.04	74.38	60.88	**56.91**	58.87	53.75	54.14	51.59	51.03	68.60±0.56	56.05±0.51
	MagicNet [1]	91.69	66.33	88.59	90.28	89.64	86.80	61.80	74.39	59.94	52.88	57.28	**58.83**	59.53	52.74	42.35	68.94±0.56	58.33±0.52
w/ VCL	UGPCL [15]	90.84	68.10	90.04	89.86	89.92	89.09	72.45	76.95	56.10	57.52	60.73	53.87	65.19	51.67	48.38	70.71±0.17	59.42±0.31
	U^{2}PL [16]	86.09	62.61	87.08	87.18	87.48	87.35	73.57	76.57	60.32	56.67	65.75	58.30	61.75	55.24	43.55	69.97±0.25	57.42±0.20
	BaCon [5]	88.51	64.40	88.78	88.58	89.22	88.27	73.17	77.69	54.44	55.58	64.46	53.14	62.29	51.34	**51.15**	70.07±0.35	57.94±0.53
	CCL [4]	88.70	67.89	88.44	88.80	89.48	88.93	**75.28**	76.96	61.22	56.86	64.47	51.07	62.65	37.07	34.13	68.80±1.65	57.28±1.02
	Ours	**92.08**	**71.71**	90.39	**90.55**	**90.70**	**89.19**	71.87	**78.77**	**61.87**	55.64	**70.24**	55.32	**66.21**	**55.61**	50.05	**72.45±0.76**	**61.15±0.62**

Table 2. Quantitative results (mean Dice for each organ, mean and SD of Dice, and mean and SD of Jaccard) on two settings of **AMOS dataset**. '*w/o VCL*' or '*w/ VCL*' indicates whether the SSL methods combined with VCL or not. **Best results are boldfaced.**

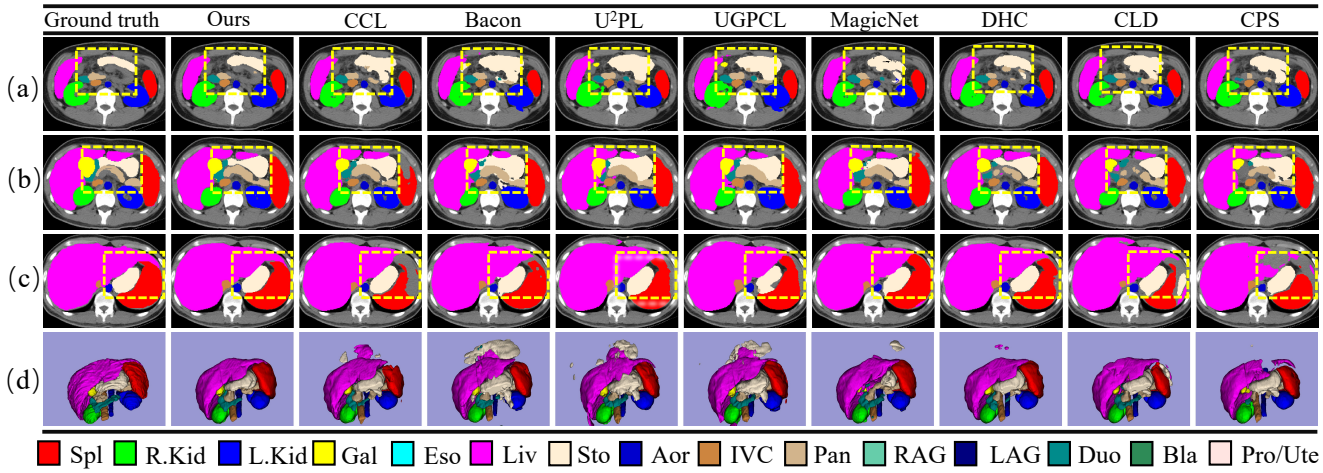

Figure 2. Visualization of the segmentation results for the **AMOS dataset**. (a-c) Segmentation results for one case of three transverse sections and (d) 3D segmentation views. As seen,better segmentation is achieved at the boundaries by our method, highlighted in yellow dashed boxes.

C. Detailed Algorithm.

Algorithm 1 provides the pseudocode for our method. By DVCL, we encourage unreliable voxels to learn more discriminative features, allowing them to fully enjoy the advantages of VCL. This further facilitates the generation of higher quality pseudo-labels.

	Methods	Mean Dice for each organ							Mean Dice	Mean Jaccard
		MYO	LAC	RAC	LVC	RVC	AA	PA		
		10% labeled data (labeled:unlabeled=2:10)								
w/o VCL	DAN [18]	89.19	69.26	80.65	68.50	84.23	83.83	44.57	74.32±0.84	61.77±1.53
	MT [13]	87.83	77.09	81.65	77.16	84.78	90.88	69.46	81.27±0.57	69.49±0.65
	UA-MT [17]	89.50	77.67	84.84	78.30	86.10	92.02	69.35	82.54±0.84	71.28±0.97
	SASSnet [8]	89.66	77.62	85.61	**80.92**	85.43	90.99	59.20	81.35±0.45	70.15±0.34
	DTC [10]	89.29	77.83	84.94	78.70	86.24	90.23	71.74	82.71±0.50	71.32±0.57
	CPS [2]	90.10	79.00	85.45	80.84	87.98	90.71	70.82	83.56±0.22	72.63±0.40
	CLD [9]	90.82	80.88	88.70	78.94	87.22	92.49	70.48	84.22±0.80	73.84±0.85
	DHC [14]	90.40	77.14	85.87	78.64	**88.91**	89.99	74.71	83.67±0.80	72.83±1.13
	MagicNet[1]	87.99	74.14	85.03	75.38	83.93	90.03	59.51	79.43±0.67	67.56±1.11
w/ VCL	UGPCL [15]	90.70	78.41	84.58	78.33	86.86	91.81	77.05	83.96±0.36	73.12±0.48
	U^2PL [16]	90.16	80.52	87.50	77.29	88.38	93.00	73.41	84.32±0.18	73.70±0.26
	BaCon [5]	90.80	79.82	85.64	80.49	87.21	91.76	73.54	84.18±0.12	73.48±0.11
	CCL [4]	90.61	80.04	87.41	80.87	87.25	92.16	68.11	83.78±0.31	73.16±0.35
	Ours	**91.27**	**82.54**	**89.01**	80.08	88.25	**93.21**	**80.43**	**86.40±0.38**	**76.56±0.58**
		50% labeled data (labeled:unlabeled=6:6)								
w/o VCL	DAN [18]	90.46	82.08	87.21	77.20	87.64	89.35	73.16	83.87±0.80	73.11±1.10
	MT [13]	90.57	84.49	88.27	80.82	88.31	92.68	77.65	86.11±0.20	76.18±0.15
	UA-MT [17]	90.94	84.59	88.47	82.27	88.36	93.12	77.30	86.43±0.07	76.71±0.19
	SASSnet [8]	88.73	79.02	85.89	79.30	85.60	91.14	71.59	83.04±0.40	71.82±0.60
	DTC [10]	91.20	84.23	88.03	80.95	88.36	90.79	77.35	85.84±0.22	75.81±0.39
	CPS [2]	91.42	84.40	88.90	82.10	88.93	92.99	76.89	86.52±0.15	76.84±0.15
	CLD [9]	91.12	85.49	90.54	79.98	88.92	91.37	80.63	86.87±0.33	77.29±0.46
	DHC [14]	90.75	84.93	89.34	81.83	88.69	93.72	79.23	86.93±0.14	77.43±0.23
	MagicNet[1]	91.27	83.58	87.75	79.07	**89.21**	90.12	66.07	83.87±0.65	73.52±0.57
w/ VCL	UGPCL [15]	91.76	85.35	89.56	80.13	88.04	93.36	76.89	86.44±0.47	76.72±0.67
	U^2PL [16]	91.07	**85.99**	90.81	81.91	88.14	93.66	75.17	86.68±0.07	77.10±0.11
	BaCon [5]	91.26	84.91	89.42	82.12	88.05	93.13	74.98	86.27±0.10	76.44±0.14
	CCL [4]	91.02	84.44	87.51	81.43	87.60	89.76	77.54	85.62±0.66	75.31±0.93
	Ours	**91.48**	85.82	**91.18**	**83.12**	88.58	**94.18**	**83.15**	**88.22±0.08**	**79.33±0.14**

Table 3. Quantitative results (mean Dice for each organ, mean and SD of Dice, and mean and SD of Jaccard) on two settings of **MMWHS dataset**. '*w/o VCL*' or '*w/ VCL*' indicates whether the SSL methods combined with VCL or not. **Best results are boldfaced.**

D. More Experimental Results.

In our experiments, the segmentation performances using two standard evaluation metrics: the Dice and the Jaccard index (referred to as Jaccard). The Dice and Jaccard values range from 0 to 1, with higher scores indicating more accurate segmentation. To reduce the randomness of network training, experiments are calculated in triplicate for all methods and the mean and standard deviation (SD) of the Dice and Jaccard values are calculated.

FLARE 2022. For the FLARE 2022 dataset, we train the semi-supervised models using training sets with labeled data proportions of 50% and 10%. The supervised 3D U-Net is trained using all 42 labeled data cases (Fully). Table.1 presents the qualitative results, including the mean Dice for each organ, as well as the mean and SD of both Dice and Jaccard. Compared to the supervised 3D U-Net, all semi-supervised methods achieve higher mean Dice and mean Jaccard by utilizing the unlabeled data. Our method significantly outperforms all the other methods, achieving a superior state-of-the-art performance.

AMOS. To further validate our method, we conduct experiments on the AMOS dataset. Table.2 presents the qualitative results, including the mean Dice for each organ, as well as the mean and SD of both Dice and Jaccard. We run the state-of-the-art semi-supervised segmentation methods and our methods on the second medical image segmentation dataset (AMOS [6]) under various label ratios (*e.g.,* 10%, 50%). Our method achieves superior segmentation performance compared to all other approaches.

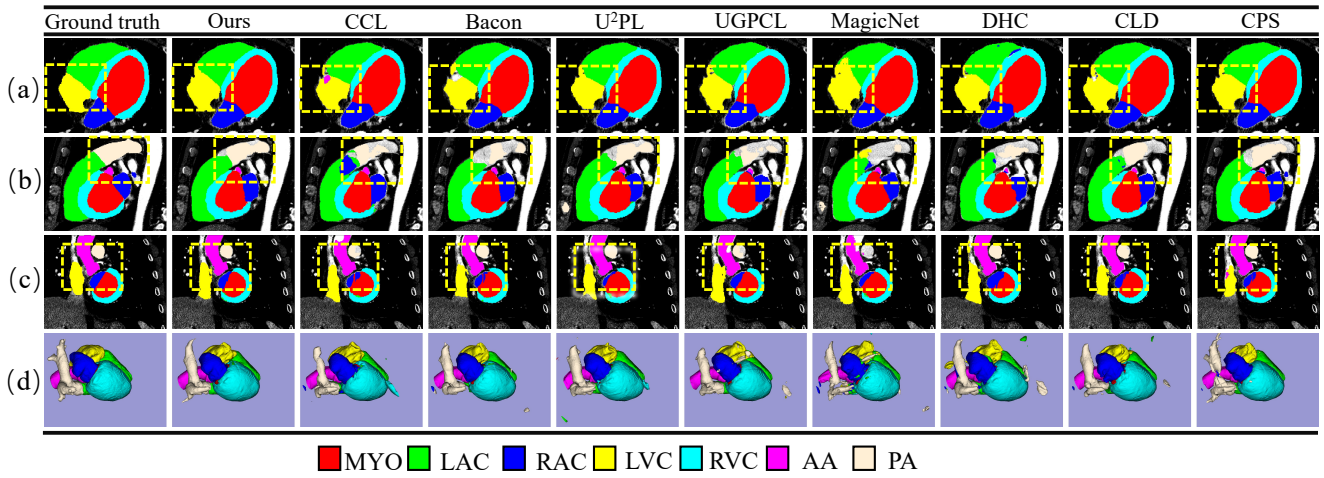

Figure 3. Visualization of the segmentation results for the **MMWHS dataset**. (a-c) Segmentation results for one case of three transverse sections and (d) 3D segmentation views. As seen,better segmentation is achieved at the boundaries by our method, highlighted in yellow dashed boxes.

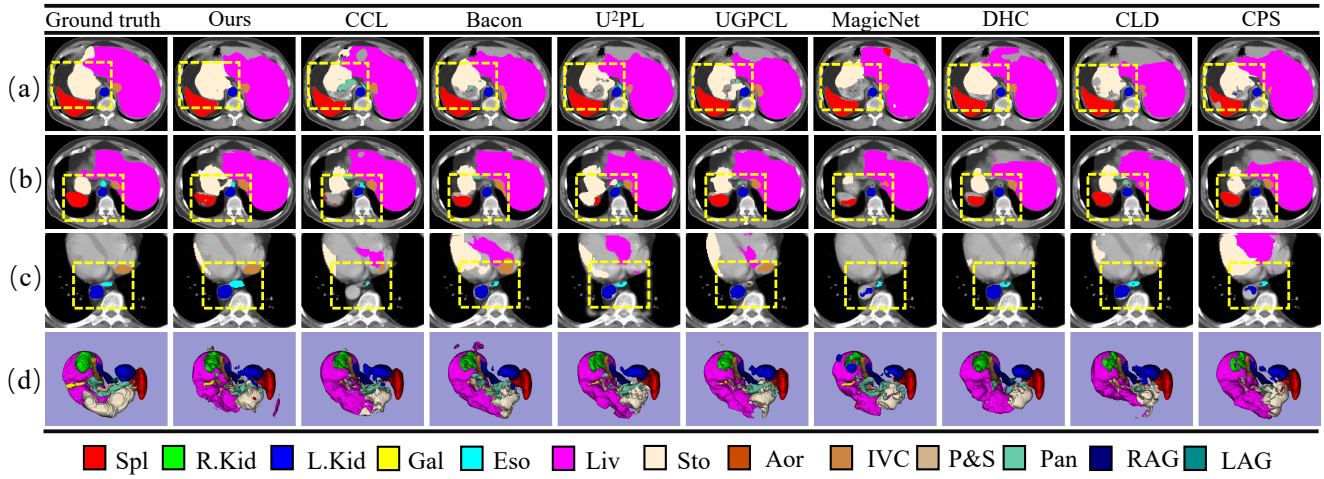

Figure 4. Visualization of the segmentation results for the **BTCV dataset**. (a-c) Segmentation results for one case of three transverse sections and (d) 3D segmentation views. As seen,better segmentation is achieved at the boundaries by our method, highlighted in yellow dashed boxes.

MMWHS. To further validate our method, we conduct experiments on the MMWHS dataset. Table.3 presents the qualitative results, including the mean Dice for each organ, as well as the mean and SD of both Dice and Jaccard. We run the state-of-the-art semi-supervised segmentation methods and our methods on the third medical image segmentation dataset (MMWHS [19]) under various label ratios (*e.g.,* 10%, 50%). Our method achieves superior segmentation performance compared to all other approaches.

BTCV. To further validate our method, we conduct experiments on the BTCV dataset. Table.4 presents the qualitative results, including the mean Dice for each organ, as well as the mean and SD of both Dice and Jaccard. We run the state-of-the-art semi-supervised segmentation methods and our methods on the fourth medical image segmentation dataset (BTCV [7]) under various label ratios (*e.g.,* 10%, 50%). Our method achieves superior segmentation performance compared to all other approaches.

E. More Qualitative Visualization.

We illustrate representative qualitative results of our method and different methods on the FLARE 2022 (Fig.1), AMOS (Fig.2), MMWHS (Fig.3), and BTCV (Fig.4) datasets. It is evident that DVCL produces more accurate predictions by encouraging unreliable voxels to learn more discriminative features. This helps address the issues of ambiguous organ categories that often lead to confusion in baseline models.

	Methods	Mean Dice for each organ													Mean Dice	Mean Jaccard
		Spl	R.kid	L.kid	Gal	Eso	Liv	Sto	Aor	IVC	P&S	Pan	RAG	LAG		
	10% labeled data (labeled:unlabeled=2:12)															
w/o VCL	3D U-Net [3]	0.25	31.44	30.99	0.00	0.00	47.78	3.76	41.07	9.94	3.67	0.41	0.00	0.02	13.03±0.15	8.29±0.10
	DAN [18]	59.98	82.65	77.08	4.36	0.71	80.22	8.69	46.88	60.62	30.47	11.36	18.65	8.30	37.69±1.48	29.06±1.11
	MT [13]	64.00	82.25	71.56	16.80	27.09	85.93	8.64	80.29	60.62	37.63	29.79	14.90	7.52	45.16±1.94	35.20±1.42
	UA-MT [17]	75.46	80.28	73.22	11.95	14.47	84.74	9.89	79.79	70.66	37.14	20.86	14.36	11.39	44.94±1.17	35.69±0.84
	SASSnet [8]	74.17	80.94	80.02	16.25	6.37	85.70	6.28	77.26	69.04	40.79	32.16	29.83	1.67	46.19±1.09	36.98±0.88
	DTC [10]	76.05	78.89	78.16	10.25	35.50	86.14	7.05	84.24	70.72	32.19	22.22	36.71	3.30	47.80±1.42	38.19±1.10
	CPS [2]	75.56	77.34	77.33	8.90	18.14	86.16	13.48	82.98	67.81	43.29	21.24	31.48	**15.97**	47.67±0.76	37.86±0.41
	CLD [9]	69.84	82.48	71.53	**23.56**	9.92	86.14	14.41	83.15	69.04	43.71	31.03	38.94	13.13	48.99±0.92	38.73±0.81
	DHC [14]	72.12	80.80	**80.79**	16.38	24.63	85.61	**15.29**	83.02	62.29	43.46	29.70	35.78	7.65	49.04±0.35	38.68±0.17
	MagicNet [1]	71.86	82.00	79.35	16.30	34.69	86.18	12.00	83.93	64.56	42.66	30.60	37.61	10.26	50.15±0.58	39.50±0.54
w/ VCL	UGPCL [15]	75.43	81.72	78.66	13.43	6.89	87.01	9.04	84.46	72.62	34.85	28.03	29.84	3.73	46.59±1.29	37.72±1.06
	U^{2}PL [16]	72.48	80.69	79.94	10.21	30.93	86.93	12.96	84.98	64.67	40.14	24.53	28.54	12.53	48.42±0.68	38.44±0.23
	BaCon [5]	76.02	82.64	74.78	15.67	22.33	84.35	9.73	82.58	69.35	44.20	25.56	32.21	3.26	47.90±1.50	38.20±1.12
	CCL [4]	76.20	80.37	77.95	13.63	29.25	85.58	12.22	81.36	65.88	37.94	29.18	26.14	8.08	47.98±1.43	37.84±1.08
	Ours	**76.40**	**83.16**	75.08	21.02	**44.27**	**87.32**	11.73	**85.92**	**73.46**	**44.97**	**30.61**	**42.02**	11.87	**52.91±0.22**	**42.43±0.19**
	50% labeled data (labeled:unlabeled=12:12)															
w/o VCL	3D U-Net [3]	61.26	75.08	71.21	8.08	24.06	81.37	10.69	78.87	55.11	32.56	11.04	2.66	6.96	39.92±1.17	30.70±1.23
	DAN [18]	74.07	88.80	88.56	20.19	26.10	85.69	36.81	78.79	59.77	52.95	50.12	48.31	40.16	57.72±1.21	45.94±1.15
	MT [13]	77.10	88.55	87.19	11.92	66.83	86.01	31.71	87.84	77.10	47.28	42.45	36.18	41.94	60.16±0.95	48.98±0.46
	UA-MT [17]	75.34	86.48	87.71	12.40	57.77	87.63	28.83	88.30	75.48	44.36	37.65	38.01	33.42	57.95±0.80	47.05±0.50
	SASSnet [8]	76.48	**90.28**	87.96	12.01	56.46	87.64	35.56	89.29	78.34	51.25	51.73	42.97	44.87	61.91±0.12	50.75±0.12
	DTC [10]	77.85	86.85	89.45	12.17	64.13	**88.81**	29.97	89.02	74.62	46.98	41.73	52.61	35.89	60.77±0.96	49.71±0.67
	CPS [2]	76.54	82.93	87.82	6.66	63.06	88.00	31.44	89.13	77.26	49.79	45.48	49.78	44.73	60.97±0.43	49.74±0.51
	CLD [9]	75.96	88.27	**90.20**	15.20	54.09	82.62	33.63	87.43	76.16	49.84	50.15	**58.00**	37.02	61.42±0.84	49.82±0.80
	DHC [14]	76.65	88.71	87.19	10.00	69.03	80.67	33.06	86.43	73.24	48.67	47.97	53.97	40.26	61.22±1.57	49.43±1.42
	MagicNet [1]	77.01	88.72	86.56	17.47	67.24	76.52	37.11	87.36	76.74	47.95	47.63	48.95	41.73	61.61±0.34	49.34±0.66
w/ VCL	UGPCL [15]	75.69	85.70	88.82	5.86	61.98	86.13	36.38	88.80	74.13	46.61	34.14	38.84	44.39	59.04±0.84	48.08±0.79
	U^{2}PL [16]	76.59	87.67	87.72	19.44	63.14	87.98	27.53	**89.49**	76.45	47.06	44.70	44.66	44.93	61.34±0.27	50.13±0.33
	BaCon [5]	75.91	88.17	87.04	11.15	66.55	87.77	34.52	88.14	75.76	45.18	37.69	52.99	43.18	61.08±0.04	49.71±0.10
	CCL [4]	77.14	87.77	89.53	5.81	67.04	86.75	31.91	88.54	**78.35**	43.50	36.69	53.75	38.41	60.40±0.89	49.40±0.86
	Ours	**78.44**	88.50	89.83	**25.37**	**67.75**	**89.57**	**42.23**	88.54	75.85	**54.66**	**52.69**	50.31	**45.16**	**65.30±0.78**	**53.50±0.70**

Table 4. Quantitative results (mean Dice for each organ, mean and SD of Dice, and mean and SD of Jaccard) on two settings of **BTCV dataset**. '*w/o VCL*' or '*w/ VCL*' indicates whether the SSL methods combined with VCL or not. **Best results are boldfaced.**

K Neighbors	*K'* Outsiders	Mean Dice	Mean Jaccard
5	5	82.38±0.27	71.07±0.25
5	10	82.92±0.45	71.81±0.34
10	10	83.24±0.17	72.24±0.17
10	15	**84.15±0.07**	**73.62±0.29**
15	15	83.81±0.12	73.12±0.09

Table 5. Ablation study on the number of neighbors and outsiders.

F. More Ablation Study for Hyper-parameters.

To validate the robustness of our method, we conduct ablation studies on the FLARE 2022 dataset by varying several hyperparameters, including the number of neighbors and outsiders, variance factor α, temperature coefficient τ, number of negative candidates N, anchor voxels M, Scalar β, contrastive loss weight λ_c, and dimensionality of voxel-level features $\mathcal{V}$.

Number of Neighbors and Outsiders. As shown in Table.5, the optimal values are *K*=10 and *K'*=15, which yield the best segmentation results. When *K* is small, the query has fewer neighbors. Although these neighbors have high similarity, they may contain limited semantic information. Conversely, when *K* is large, the number of neighbors increases, potentially introducing misclassified samples with high similarity but belonging to different categories.

Variance Factor α. In the entropy-based selection module, α is the variance factor influences the proportion of reliable and unreliable voxels. As shown in Fig.5(a), our method trained with $\alpha = 0.5$ obtains the best segmentation results.

Temperature Coefficient σ. In the calculation of $\mathcal{L}_{hqcl}$, σ plays a significant role in adjusting the emphasis on challenging samples. To evaluate the effect of different σ values, ablation experiments are performed using $\sigma \in \{0.05, 0.1, 0.5, 1, 5\}$. The results (Fig.5(b)) indicate that $\tau = 1$ achieves optimal segmentation performance.

Number of Negative Candidates N and anchor voxels M. N and M are used to calculate $\mathcal{L}_{hqcl}$. As shown in Fig.5(c), the $\mathcal{L}_{dvcl}$ trained with $N = 50$ and $M = 256$ achieves the best segmentation results.

Scalar β. Scalar β is used to balance $\mathcal{L}_{hqcl}$ and $\mathcal{L}_{dvcl}$. As

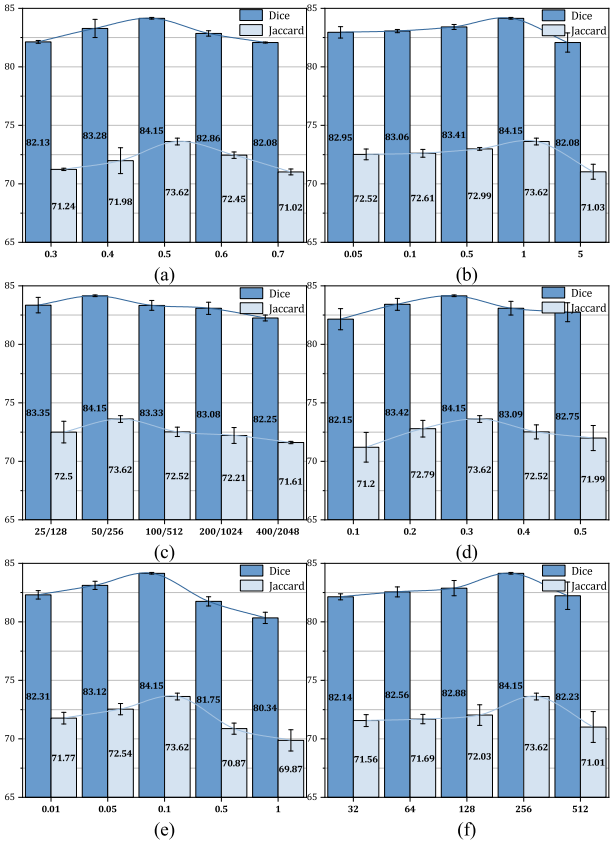

Figure 5. Quantitative comparisons of different hyper-parameters. (a-f) present mean Dice and mean Jaccard generated by the our method trained with different α, τ, M/N, β, λ_c, and $\mathcal{V}$, respectively.

shown in Fig.5(d), $\beta = 0.3$ obtains the highest mean Dice and mean Jaccard.

Contrastive Loss Weight λ_c. λ_c determines the contribution of contrastive loss to the total loss. As shown in Fig.5(e), $\lambda_c = 0.1$ obtains the best segmentation results.

Dimensionality of Voxel-level Features $\mathcal{V}$. $\mathcal{V}$ determines the dimensionality of voxel-level features generated by the feature head. Experiments are conducted using $\mathcal{V} \in \{32, 64, 128, 256, 512\}$. The results (Fig.5(f)) demonstrate that $\mathcal{V} = 256$ achieves the optimal segmentation performance.

G. The Complete Derivation of The Formula .

Proof of Eq.19 in our manuscript:

$$\begin{aligned}\psi(\mathcal{C}_m^K, \mathcal{D}_m^{K'}) &= -\log \frac{\mathcal{N}\left(\mathcal{C}_m^K \mid \mathbf{r}_m, \theta_{\mathcal{S}}\right)}{\mathcal{N}\left(\mathcal{D}_m^{K'} \mid \mathbf{r}_m, \theta_{\mathcal{S}}\right)} \\ &= -\log \frac{\prod_{\mathbf{r}_n \in \mathcal{C}_m^K} \mathcal{O}_{m,n}(\mathcal{A}_{m,n})}{\prod_{\mathbf{r}_k \in \mathcal{D}_m^{K'}} \mathcal{O}_{m,k}} \\ &= -\log \prod_{\mathbf{r}_n \in \mathcal{C}_m^K} \mathcal{O}_{m,n}(\mathcal{A}_{m,n}) + \log \prod_{\mathbf{r}_k \in \mathcal{D}_m^{K'}} \mathcal{O}_{m,k}\end{aligned}$$

Since the logarithm of a product is equal to the sum of the logarithms, we have

$$\begin{aligned}\psi(\mathcal{C}_m^K, \mathcal{D}_m^{K'}) &= -\log \prod_{\mathbf{r}_n \in \mathcal{C}_m^K} \mathcal{O}_{m,n}(\mathcal{A}_{m,n}) + \log \prod_{\mathbf{r}_k \in \mathcal{D}_m^{K'}} \mathcal{O}_{m,k} \\ &= -\sum_{\mathbf{r}_n \in \mathcal{C}_m^K} \log(\mathcal{O}_{m,n}(\mathcal{A}_{m,n})) + \sum_{\mathbf{r}_k \in \mathcal{D}_m^{K'}} \log(\mathcal{O}_{m,k}) \\ &= -\sum_{\mathbf{r}_n \in \mathcal{C}_m^K} \log\left(\frac{e^{\mathcal{A}_{m,n} p_m^T p_n}}{\sum_{\mathbf{r}_q \in U_{\mathcal{R}}} e^{p_m^T p_q}}\right) + \sum_{\mathbf{r}_k \in \mathcal{D}_m^{K'}} \log\left(\frac{e^{p_m^T p_k}}{\sum_{\mathbf{r}_q \in U_{\mathcal{R}}} e^{p_m^T p_q}}\right) \\ &= -\sum_{\mathbf{r}_n \in \mathcal{C}_m^K} [\mathcal{A}_{m,n} p_m^T p_n - \log(\sum_{\mathbf{r}_q \in U_{\mathcal{R}}} e^{p_m^T p_q})] + \sum_{\mathbf{r}_k \in \mathcal{D}_m^{K'}} [p_m^T p_k - \log(\sum_{\mathbf{r}_q \in U_{\mathcal{R}}} e^{p_m^T p_q})] \\ &= -\sum_{\mathbf{r}_n \in \mathcal{C}_m^K} \mathcal{A}_{m,n} p_m^T p_n + \sum_{\mathbf{r}_k \in \mathcal{D}_m^{K'}} p_m^T p_k + (K - K') \log(\sum_{\mathbf{r}_q \in U_{\mathcal{R}}} e^{p_m^T p_q})\end{aligned} \tag{1}$$

Consider the last term of Eq.1 row 5:

$$\begin{aligned}&(K - K') \log(\sum_{\mathbf{r}_q \in U_{\mathcal{R}}} e^{p_m^T p_q}) \\ &= (K - K') \log(|U_R| \cdot \frac{\sum_{\mathbf{r}_q \in U_{\mathcal{R}}} e^{p_m^T p_q}}{|U_R|}) \\ &= (K - K')(\log|U_R| + \log(\frac{\sum_{\mathbf{r}_q \in U_{\mathcal{R}}} e^{p_m^T p_q}}{|U_R|}))\end{aligned}$$

Using *Jensen Inequality* $\log(\mathbb{E}[X]) \geq \mathbb{E}[\log(X)]$, we have

$$\begin{aligned}&(K - K')(\log|U_R| + \log(\frac{\sum_{\mathbf{r}_q \in U_{\mathcal{R}}} e^{p_m^T p_q}}{|U_R|})) \\ &\leq (K - K')(\log|U_{\mathcal{R}}| + \sum_{\mathbf{r}_q \in U_{\mathcal{R}}} \frac{p_m^T p_q}{|U_{\mathcal{R}}|})\end{aligned}$$

The expression $p_m^T p_q$ represents the inner product between the probability vector of $\mathbf{r}_m$ and the probability vectors of all $\mathbf{r}_n$ in $U_{\mathcal{R}}$. Since $\mathbf{r}_m$ and $\mathbf{r}_n$ are not identical, it follows that $p_m^T p_q < 1$, and thus $\sum_{\mathbf{r}_q \in U_{\mathcal{R}}} \frac{p_m^T p_q}{|U_{\mathcal{R}}|} < 1$. We have

$$(K - K')(\log|U_{\mathcal{R}}| + \sum_{\mathbf{r}_q \in U_{\mathcal{R}}} \frac{p_m^T p_q}{|U_{\mathcal{R}}|}) < (K - K')(\log|U_{\mathcal{R}}| + 1) \tag{2}$$

Substituting Eq.2 into the last term of Eq.1 row 5 gives

$$\begin{aligned}&\psi(\mathcal{C}_m^K, \mathcal{D}_m^{K'}) \\ &= -\sum_{\mathbf{r}_n \in \mathcal{C}_m^K} \mathcal{A}_{m,n} p_m^T p_n + \sum_{\mathbf{r}_k \in \mathcal{D}_m^{K'}} p_m^T p_k + (K - K') \log(\sum_{\mathbf{r}_q \in U_{\mathcal{R}}} e^{p_m^T p_q}) \\ &\leq -\sum_{\mathbf{r}_n \in \mathcal{C}_m^K} \mathcal{A}_{m,n} p_m^T p_n + \sum_{\mathbf{r}_k \in \mathcal{D}_m^{K'}} p_m^T p_k + (K - K')(\log|U_{\mathcal{R}}| + \sum_{\mathbf{r}_q \in U_{\mathcal{R}}} \frac{p_m^T p_q}{|U_{\mathcal{R}}|}) \\ &< -\sum_{\mathbf{r}_n \in \mathcal{C}_m^K} \mathcal{A}_{m,n} p_m^T p_n + \sum_{\mathbf{r}_k \in \mathcal{D}_m^{K'}} p_m^T p_k + (K - K')(\log|U_{\mathcal{R}}| + 1) \\ &= \overline{\psi}(\mathcal{C}_m^K, \mathcal{D}_m^{K'})\end{aligned}$$

References

[1] 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. 1, 3, 4, 5, 7

[2] 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. 3, 4, 5, 7

[3] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 4, 7

[4] Qinyi Deng, Yong Guo, Zhibang Yang, Haolin Pan, and Jian Chen. Boosting semi-supervised learning with contrastive complementary labeling. *Neural Networks*, 170:417–426, 2024. 3, 4, 5, 7

[5] Qianhan Feng, Lujing Xie, Shijie Fang, and Tong Lin. Bacon: Boosting imbalanced semi-supervised learning via balanced feature-level contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11970–11978, 2024. 3, 4, 5, 7

[6] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022. 1, 5

[7] Bennett Landman, Zhoubing Xu, J 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, 2015. 1, 6

[8] Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 552–561. Springer, 2020. 3, 4, 5, 7

[9] 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. 3, 4, 5, 7

[10] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8801–8809, 2021. 3, 4, 5, 7

[11] 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 *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 318–329. Springer, 2021. 1

[12] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, Fan Zhang, Wentao Liu, YuanKe Pan, Shoujin Huang, Jiacheng Wang, Mingze Sun, Weixin Xu, Dengqiang Jia, Jae Won Choi, Natália Alves, Bram de Wilde, Gregor Koehler, Yajun Wu, Manuel Wiesenfarth, Qiongjie Zhu, Guoqiang Dong, Jian He, the FLARE Challenge Consortium, and Bo Wang. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*, 2023. 1

[13] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3, 4, 5, 7

[14] 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. 1, 3, 4, 5, 7

[15] Tao Wang, Jianglin Lu, Zhihui Lai, Jiajun Wen, and Heng Kong. Uncertainty-guided pixel contrastive learning for semi-supervised medical image segmentation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 1444–1450, 2022. 3, 4, 5, 7

[16] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4248–4257, 2022. 3, 4, 5, 7

[17] 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 *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 605–613. Springer, 2019. 3, 4, 5, 7

[18] Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 408–416. Springer, 2017. 3, 4, 5, 7

[19] Xiahai Zhuang and Juan Shen. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. *Medical image analysis*, 31:77–87, 2016. 1, 6