Published in final edited form as:
IEEE J Biomed Health Inform. 2023 September ; 27(9): 4444–4453. doi:10.1109/JBHI.2023.3285230.

Semantic-Aware Contrastive Learning for Multi-object Medical Image Segmentation

Ho Hin Lee,
Computer Science Department, Vanderbilt University, Nashville TN 37235

Yucheng Tang,
Electrical and Computer Engineering Department, Vanderbilt University, Nashville TN 37235

Qi Yang,
Computer Science Department, Vanderbilt University, Nashville TN 37235

Xin Yu,
Computer Science Department, Vanderbilt University, Nashville TN 37235

Leon Y. Cai,
Biomedical Engineering Department, Vanderbilt University, Nashville TN 37235

Lucas W. Remedios,
Computer Science Department, Vanderbilt University, Nashville TN 37235

Shunxing Bao,
Computer Science Department, Vanderbilt University, Nashville TN 37235

Bennett A. Landman [Senior Member, IEEE],
Electrical and Computer Engineering Department, Vanderbilt University, Nashville TN 37235

Yuankai Huo [Senior Member, IEEE]
Computer Science Department, Vanderbilt University, Nashville TN 37235

Abstract

Medical image segmentation, or computing voxelwise semantic masks, is a fundamental yet challenging task in medical imaging domain. To increase the ability of encoder-decoder neural networks to perform this task across large clinical cohorts, contrastive learning provides an opportunity to stabilize model initialization and enhances downstream tasks performance without ground-truth voxel-wise labels. However, multiple target objects with different semantic meanings and contrast level may exist in a single image, which poses a problem for adapting traditional contrastive learning methods from prevalent "image-level classification" to "pixel-level segmentation". In this paper, we propose a simple semantic-aware contrastive learning approach leveraging attention masks and image-wise labels to advance multi-object semantic segmentation. Briefly, we embed different semantic objects to different clusters rather than the traditional image-level embeddings. We evaluate our proposed method on a multi-organ medical image segmentation task with both in-house data and MICCAI Challenge 2015 BTCV

Corresponding author: Yuankai Huo, yuankai.huo@vanderbilt.edu.
Ho Hin Lee is the first author to the manuscript.

datasets. Compared with current state-of-the-art training strategies, our proposed pipeline yields a substantial improvement of 5.53% and 6.09% on Dice score for both medical image segmentation cohorts respectively (p-value<0.01). The performance of the proposed method is further assessed on external medical image cohort via MICCAI Challenge FLARE 2021 dataset, and achieves a substantial improvement from Dice 0.922 to 0.933 (p-value<0.01). The code is available at: https://github.com/MASILab/DCC_CL

Index Terms—

Medical image segmentation; contrastive learning; attention map; query patches

I. Introduction

CONTRASTIVE learning methods learn an augmentation invariant feature embedding, which opens a new window of developing a deep learning model with large-scale unannotated data and few annotated data [1], [2]. Traditional contrastive learning approach consists of two primary concepts: 1) the learning process pulls the target image (anchor) and a matching sample close to each other as a “positive pair” in the embedding space, and 2) the learning process pushes the anchor from non-matching samples away from each other as “negative pairs” in the embedding space. Data augmentation is used to generate the positive samples from a training sample, while the negative pairs are formed from the remaining samples of non-matching objects. Previous studies demonstrate the advantages of contrastive learning in image-level classification tasks [3]-[5]. Meanwhile, multi-organ segmentation in medical domain is a fundamental yet challenging task when limited annotated samples are available. Previous proposed self-supervised learning model is able to extract semantic oriented spatial context for initializing a multi-object segmentation deep neural network [6]. We posit that contrastive learning can also leverage the capability of sub-image-level feature encoding, to advance pixel-level segmentation tasks. However, some gaps need to be filled to achieve the latter goal, especially for multi-organ segmentation tasks in medical imaging [7]-[9]. For example, multiple semantic objects may exist in medical images (e.g., abdomen, organs, brain tissues), while each element in the convoluted/downsampled feature may correlate to multiple objects. Thus, it is difficult to align the object-wise semantics with the learned representation to enhance model interpretability in the latent space for multi-organ segmentation as the downstream task.

In this work, we propose a semantic-aware attention-guided contrastive learning (AGCL) framework to advance multi-object medical image segmentation with contrastive learning. We integrate object-corresponding attention maps as additional input channels to adapt representations into corresponding semantic embeddings (as shown in Fig. 1). To further stabilize the latent space, we propose a multi-class conditional contrastive loss that increases the arbitrary number of positive pairs within the same sub-class for contrastive learning. Instead of leveraging pixel-wise label, radiological conditions such as modality and organ semantics, are provided as image-wise multi-class label to constrain the normalized embedding. By introducing multiple semantics with our proposed contrastive learning strategy, the learned features can be classified into embeddings with multiple semantic meanings, thus enhancing the feature intrepretability in the latent space. Furthermore, such latent space is easily to provide explainability about the mdoel robustness to the clinical teams, by observing the separability between each semantic embeddings. Fig. 2 provides a visual explanation of our proposed framework. Our proposed contrastive learning strategy AGCL is evaluated with three medical imaging datasets (two public contrast-enhanced CT dataset [10], [11] and one in-house non-contrast dataset). The results demonstrate that consistent improvements are achieved on both ResNet-50 and ResNet-101 architectures [12]. Our main contributions are summarized as below:

1. We propose a semantic-aware contrastive learning framework to advance multi-object pixel-level semantic segmentation.
2. We propose a multi-conditional contrastive loss to integrate multiple radiological conditions as additional constraints for classifying representations into sub-class embeddings.
3. We demonstrate that the proposed AGCL generalizes the CT contrast phase variation in each organ and significantly boosts the segmentation performance.

II. Related Works

Contrastive Learning:

Self-supervised representation learning approaches have recently been proposed to learn useful representation from unlabeled data. Some approaches propose learning embeddings directly in lower-dimensional representation spaces instead of computing a pixel-wise predictive loss [13]. Self-distillation with a teacher-student network is further proposed to enrich the semantic correspondence using pseudo-label predictions [14], while the masked autoencoder provides an alternative to learn the spatial feature correspondence with the image reconstruction task [15]. Contrastive learning is one of such state-of-the-art methods for self-supervised learning to model the semantic-wise relationships in the latent space [1], [16]. It employs a loss function to pull latent representations closer together for positive pairs, while pushing them apart for negative pairs. Maximizing mutual information between embeddings has also been proposed as an alternative to extract the similar information between targets [17|. Adapting with memory bank and momentum contrastive approaches have been proposed to increase the batch sizes and generate more dissimilar pairs in a minibatch for contrastive learning [18]. Additionally, to constrain and stabilize the embedding spaces, class label information has been added to provide additional supervision to stabilize contrastive learning process [2].

However, most of the prior works in contrastive learning focused on improving the "image-wise classification", while relatively fewer methods have been proposed for the "pixel-wise segmentation". Pixel-wise contrastive loss is proposed to adapt the representation from the ground-truth label information [9], while dense contrastive loss is also proposed to minimize the discrepancy of image-level prediction and pixel-level prediction [19]. Furthermore, one-stage contrastive learning framework is proposed to enforce the pixel embeddings belonging to a same semantic class to be more similar than embeddings from different classes [20]–[22]. In the medical domain, the contrastive learning framework is extended to leverage the structural similarity and learn the distinctive representations of local regions without using pixel-wise ground truth labels [7] and image-wise labels [23]. Similarly, limited ground-truth labels are adapted into the pretraining step for contrastive learning and enhanced the segmentation performance [24], [25]. However, such methods typically need pixel-wise segmentation labels. Therefore, our proposed method identifies the semantic-aware regions with one-hot attentional guidance and leverages multi-class image-level labels to define region-bounded representations as an arbitrary number of positive pairs for contrastive learning, without using pixel-wise labels.

Medical Image Segmentation & Multi-Organ Segmentation:

Fully-supervised deep learning methods have been developed to enhance both the segmentation performance and the generalizability across different datasets [26]–[29]. However, the supervised learning strategies are limited to the quality of pixel/voxel-wise labels and the resolution of volumes [30]. Thus, hierarchical approaches and patch-wise approaches have been proposed to perform segmentation across scales and resolutions [31]-[33]. Another study further enhances the segmentation accuracy with the statistical fusion from multi-view predictions [34]. Apart from the multi-view attention, shape-aware network is proposed to consistently smoothen the label prediction by learning the signed distance function as additional constraints [35]. Furthermore, RAP-Net is proposed to leverage one-hot shape-aware mappings to provide additional localization context as additional input channel and refine the segmentation mapping hierarchically [36]. nn-UNet further enhances the generalizability with self-configuring structure to diversely predict segmentation for multi-modality imaging [37]. In terms of generic network backbone, vision transformer is introduced as the encoder network to extract attention features with large receptive field for robust segmentation [38]. On the other hand, partially-supervised, semi-supervised and self-supervised learning have also been explored to adapt unlabeled data in the medical imaging domain. Multiple single organ-labeled datasets are used to provide structural prior knowledge during the training process with multiple organ-labeled dataset to enhance multi-organ segmentation performance [39]. A quality assurance module have been proposed to adapt the segmentation quality as the supervision from unlabeled data [40]. Pretext tasks such as colorization, deformation and image rotation, have been used as pre-training features to initialize the segmentation networks [41]. Self-supervised context has also been explored by predicting the relative patch location and the degree of rotation [42], [43]. Contrastive learning has been used to extract global and local representations for domain-specific MRI images [7]. A contrastive predictive network has been used to summarize the latent vectors in a minibatch and predicts the latent representation of adjacent patches [6].

III. Method

We present our co-training approach AGCL that integrates one-hot organ attention into contrastive learning by adapting radiological context labels (modality and organ) to classify representations into sub-classes embeddings, as presented in Fig. 3.

A. Hierarchical Coarse Segmentation

The input data of the entire pipeline is a multi-contrast 3D image volume $V_i = \{X_i, Y_i\}_{i = 1, \ldots, L}$, where L is the number of all imaging samples, X is the volumetric image and Y is the corresponding multi-organ label. The corresponding outcomes of the preprocessing stage are coarse segmentation masks (attention maps) $A_i = RAP(X_i)$ from a hierarchical segmentation network $RAP(\cdot)$ [36]. We define $A \in R^{H \times W \times D \times C}$, where H and W denote as the axial dimension of the image, D denote as the number of slices and C denotes as the number of label classes. The coarse segmentation network RAP-Net consists of two hierarchical stages: 1) low-resolution whole volume segmentation and 2) organ-specific patch segmentation refinement. The low-resolution model generates a rough segmentation map and provide anatomical context as additional channel input to refine the segmentation in patch-wise setting as the second step. Both low-resolution model and patch-wise model are trained in supervised setting with 5-fold cross-validations.

B. Data Preprocessing

The goal of the data preprocessing step is to randomly sample 2D training patches for downstream contrastive learning. In our design, we first slice all the volumetric scans (image, ground truth labels and coarse segmentations) and utilize the slice-wise attention maps (1) as spatial restrictions of the organ-specific sampling process, and (2) highlight the current organ of interest to define semantic-wise embeddings for segmentation refinement. Briefly, organ-specific patches $p_i = \{x_{C,i}, y_{C,i}, s_{C,i}\}_{i = 1, \ldots, N}$ are randomly sampled within each organ class C in attention maps. The center point is randomly sampled from attention maps to crop the region of interest (ROI), N denotes as the total number of query patches, $x_{C,i}$ is the organ-corresponding image patch, $y_{C,i}$ is the binary ground-truth label patch, and $s_{C,i}$ is the coarse organ-specific attention map from A_i slice in binary setting. As the significant difference between $y_{C,i}$ and $s_{C,i}$ is the variation of segmentation quality, the trained model aims to refine the segmentation with the prior knowledge of $s_{C,i}$ as an additional input channel. As a standard process in data augmentation, random cropping, rotation (−30 to 30 degrees), scaling has been applied to augment the size of training samples.

C. Contrastive learning with Organ-Specific Attention

After generating augmented image pairs, pairwise images are then used as the inputs for contrastive learning. Specifically, a convolutional encoder network $E(\cdot)$ is used to extract high dimensional features. We further project each high-level feature mapping into 1D vector $\tilde{z}_i$ using a multi-layer perceptron network $P(\cdot)$, $\tilde{z}_i = P(E(\tilde{a}_i))$, $\tilde{z}_i \in R^{O_E}$ (pink box in Fig. 2), where O_E is the size of the output vector. Then, the standard self-supervised contrastive loss (SSCL) [1] can be defined as the following:

$$\mathscr{L}_{self} = -\sum_{k = 1}^{2N} \log \frac{\exp(\tilde{z}_k \cdot \tilde{z}_{p(k)}/\mathscr{T})}{\sum_{j \in J(k)} \exp(\tilde{z}_k \cdot \tilde{z}_j/\mathscr{T})} \qquad (1)$$

where T is a hyperparameter indicating temperature scaling to control the radius weighting on the positive pair/negative pairs. Both k and $p(k)$ represent the index of the anchor sample and the corresponding positive sample, respectively. $J(k)$ represents the number of remaining negative samples. To incorporate the attention with modality and organ semantic meanings, we extend SSCL to adapt an arbitrary number of positive pairs by introducing multi-class image-level labels into contrastive loss. Here, modalities indicate the different contrast types in CT (we have utilized both contrast-enhanced and non-contrast CT scans in the training set). As the organ-specific attention only provides one-hot voxel-wise context to preserve organ regions, the multi-class labels represent different modalities and organs for the learned representations under the attention regions. It provides flexibility to further constrain the representations into semantic-aware clusters, which is conditional to multiple image-level labels. In each batch, pairwise patches with the same organ and modality label are defined as positive pairs, while the remaining pairs are specified as negative pairs. With such positive-negative pairs definition, we further extend the contrastive loss with conditional constraints as follows:

$$\mathscr{L}_{MT} = \sum_{k=1}^{2N} \frac{-1}{|L(k)|} \sum_{l \in L(k)} \log \frac{\exp(\tilde{z}_k \cdot \tilde{z}_l / \mathscr{T})}{\sum_{j \in J(k)} \exp(\tilde{z}_k \cdot \tilde{z}_j / \mathscr{T})} \tag{2}$$

where $L(k) \equiv \{l \in J(k): m_k = m_l, o_p = o_l\}$, m and o denote as the corresponding modality and organ label, respectively. l and $\hat{z}_l$ are the index number and the projected feature representation of the corresponding positive sample with same organ and modality label. The feature vector output with 256 channels is directly used to compute the contrastive loss of modality and organ class respectively. By classifying the learned representations into multi-classes embeddings, the model is initially learned the attention-bounded representations with semantic meanings, which are hypothesized to be beneficial for downstream segmentation tasks.

D. Co-training with Multi-Organ Segmentation

The ultimate goal of our framework is to achieve a robust patch-wise contrastive learning without using pixel-wise labels, which benefits for downstream segmentation tasks. The native two-stage strategy is to train both contrastive loss and downstream segmentation loss independently. Here, we attempt to have a co-training strategy, by training the contrastive loss and segmentation refinement tasks simultaneously. The encoder network is followed with an atrous spatial pyramid pooling (ASPP) module as the decoder network to resample the bottleneck feature with multiple effect Field Of Views (FOVs) [44]. The DeepLabV3+ is employed as the segmentation part with the shared encoder structure for contrastive learning [44]. The distinctiveness of adapting ASPP is to obtain multi-scale features during upsampling. One 1×1 convolution and three 3×3 convolution layers with different dilation rate (e.g., 6, 12,18) are leveraged. With the increased number of dilation rate, kernel stride is constrained while a larger FOV is accomplished without increasing the number of model parameters. Furthermore, image pooling is also performed in parallel to extract the global features. Features from different FOVs are finally concatenated. The channel-wise features are mixed using a 1×1 convolution layer before passing through the final layer for high-resolution prediction. The rationale of such design is to adapt the multi-view behavior and search the optimal tradeoff between the localized features (small FOV) and the global-assimilated features (large FOV). Dice loss is used to compute the predicted output with the ground truth label in binary setting for the co-training segmentation task. After computing all organ-specific patches predictions, we fuse the organ-wise patches according to the center point recorded in the data preprocessing stage. We employ majority voting [45] as the label fusion module to fuse predictions into multi-organ labels.

IV. Experiments

Datasets:

To evaluate our proposed learning approach, one in-house research cohort and two publicly available cohorts in medical imaging are used with multi-organ segmentation as the downstream task.

MICCAI 2015 Challenge Beyond The Cranial Vault (BTCV) dataset is comprised of 100 de-identified unpaired 3D contrast-enhanced CT scans with 7,968 axial slices in total. 20 scans are publicly available for the testing phase in the MICCAI 2015 BTCV challenge. All CT scans are in portal venous phase. Peak enhancement of contrast is observed in several organs, such as liver, kidney, spleen, and portal splenic vein. For each scan, 12 organ anatomical structures are well-annotated, including spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava (IVC), portal splenic vein (PSV), pancreas and right adrenal gland. Each volume consists of $47 \sim 133$ slices of 512×512 pixels, with the resolution of $([0.54 \sim 0.98] \times [0.54 \sim 0.98] \times [2.5 \sim 7.0])mm^3$.

Non-contrast clinical cohort is retrieved in de-identified form from ImageVU database of Vanderbilt University Medical Center. It consists of 56 unpaired 3D CT scans with 3,687 axial slices and expert-refined annotations for the same 12 organs as the MICCAI 2015 BTCV challenge dataset. All volumetric scans are generated without contrast enhancement procedures. Each volume consists of $49 \sim 174$ slices of 512×512 pixels, with the resolution of $([0.64 \sim 0.98] \times [0.64 \sim 0.98] \times [1.5 \sim 5.0])mm^3$.

MICCAI 2021 Challenge Fast and Low GPU memory Abdominal Organ Segmentation (FLARE) dataset leverage large scales of abdominal contrast-enhanced CT with 511 unpaired cases from 11 medical centers in multi-contrast phases (including both portal venous phase and non-contrast phase CTs). It consists of 361 3D CT scans with four organ-specific labels including spleen, kidney, liver and pancreas. Each volume consists of $43 \sim 384$ slices of 512×512 pixels, with the resolution of $([0.64 \sim 0.98] \times [0.64 \sim 0.98] \times [1.0 \sim 5.0])mm^3$.

Preprocessing:

We apply the preprocessing steps as follows: (i) applying soft tissue windowing within the range of −175 to 250 Hu and performing intensity normalization of each 3D volume, v with min-max normalization: $(v - v_1)/(v_{99} - v_1)$, where v_p denote as the p^{th} intensity percentile in v, and (ii) applying volume-wise cropping in z-axis with body part regression algorithm to extract the abdominal region only for segmentation and ensure the similar field of view between scans [48].

Network Training:

Our proposed framework AGCL is trained with unpaired samples in our scenario, while it also allows to train with paired samples. 5 -fold cross-validation is performed for both contrast-enhanced phase and non-contrast phase CT (Training: 60 volumes (contrast-enhanced) and 44 volumes (non-contrast), validation: 20 volumes (contrast-enhanced) and 6 volumes (non-contrast), and testing: 20 volumes (contrast-enhanced) and 6 volumes (non-contrast)). For training the coarse segmentation network RAP-Net, we downsample all training volumes to a resolution of $2 \times 2 \times 6$ with the dimension of $168 \times 168 \times 64$. The low-resolution volumes are leveraged to train a low-resolution segmentation model with Adam optimizer using a batch size of 1 and a learning rate of $1e - 4$. We then use the coarse segmentation output to guide and extract organ-specific patches with the dimension of $128 \times 128 \times 48$. The patch-wise segmentation refinement model is trained with Adam optimizer using a batch size of 2 and a learning rate of $1e - 4$. For contrastive learning, we perform patients-level sampling and extract 30 2D query patches of each anatomical target in each axial slices of a subject scan. Such sampling strategy ensures that all patches are fully covered the organ-specific ROIs with significant variation of anatomical morphology. More than 400k patches with dimensions 128×128 are used and shuffle to train with stochastic gradient descent (SGD) optimizer for 5 epochs with a batch size of 4 and a learning rate of 5×10^{-4}. We have evaluated the variation of the temperature parameter towards the segmentation performance and $T = 0.1$ achieves the best performances across all other temperature values. For segmentation task, the encoder's weight is frozen and only the decoder with ASPP module is trained for 10 epochs with Adam optimizer using a batch size of 4 and a learning rate of 10^{-4}. We use the validation set to choose the model with the highest mean Dice score for all semantic targets segmentation and perform inference as the quantitative representation on the testing set.

Experimental Setup:

We evaluate the segmentation performance with Dice similarity coefficient on current state-of-the-art approach in contrastive learning and segmentation task for medical imaging domain, including the testing phase of the BTCV dataset, testing cohort of the non-contrast clinical cohort and the random sampled cohort from FLARE dataset. We further perform different pre-training strategies with the multi-class image-level label using different scenarios. Apart from learning image-wise embeddings with self-supervised setting, inspired by Khosla et al. [2], we introduce patch-wise multi-label (modality & organ) classification as the pretext task via the canonical cross-entropy (CE) loss in a fully-supervised setting. The learned representations are classified into label-corresponding clusters, as in AGCL. The CE loss is defined as following:

$$\mathscr{L}_{ce} = \sum_{i = 1}^{C} y_i \log(p_i) \tag{3}$$

where y_i is the ground-truth multi-class label and p_i is the Softmax probability for the i^{th}, $i \in 1, \ldots, C$ classes. Apart from different pretext task strategies, we also perform ablation studies with the variation of hyperparameters such as temperature and the number of label used for pretraining, to investigate the optimal effect on fine-tuning segmentation task. For the encoder network, we evaluate with two common backbone architectures for segmentation in medical imaging domain: DeeplabV3+ with ResNet-50 and ResNet-101 encoder. The normalized activation of the final pooling layer with $D_E = 2048$ are used as the distinctive feature representation vector. More details of both network training and data preprocessing are demonstrated in the supplementary material.

A. Segmentation Performance—We first compare the proposed AGCL with a series of state-of-the-art approaches including 1) fully supervised approaches (training on ground-truth labeled data only), 2) a partially-supervised approach (training on one contrast phase dataset, and another with partial labels), and 3) contrastive learning approach for segmentation tasks. As shown in Table I, the contrastive learning approach demonstrates significant improvement followed by the partial-supervision and full-supervision approaches. Chaitanya et al. integrates the SSCL across global to local scale and demonstrates significant improvement across organs. Khosla et al. provides an additional single class label to address the correspondence on embeddings, which outperforms all current approaches in supervised and self-supervised contrastive learning settings. By further adding multi-class labels as conditional constraints, AGCL achieves the best performance among all state-of-the-arts with a mean Dice score of 0.926. The additional gains demonstrate that our use of supplemental imaging information allows for recognition of more positive pairs with additional label constraints. To further evaluate the generalizability of our approach, we perform external evaluations on another public multi-organ labeled datasets FLARE for multi-organ segmentation. In Table IV, AGCL demonstrates substantial improvement on all organs segmentation when comparing against the current all contrastive state-of-the-arts.

B. Ablation Study for AGCL

Comparing with first stage training approaches: To investigate the effect of using multi-class label for contrastive learning, we perform evaluation of different pretraining approaches with/out multi-class label: 1) training with self-supervised contrastive loss (SSCL), 2) training with cross-entropy (CE) loss as classification tasks, and 3) random initialization (RI) without any contrastive learning on both ResNet-50 and ResNet-101 encoder backbone. As shown in Table II and Fig. 4, SSCL improves the segmentation performance over RI by 3.12%, which is expected because RI considers no constraint in the lower-dimensional space and relies on the decoder ability for downstream tasks. With the supervision of modality and anatomical information, the supervised image classification strategies significantly boost the segmentation performance by 5.93%. Pretraining with CE is to classify representations into corresponding embeddings related to the label given and representations in the same class is moved towards each other. Therefore, such improvement demonstrates that a good definition of the latent space in encoder can help address the corresponding representation for each semantic target and starts to achieve more favorable with the segmentation task. Eventually, AGCL surpasses CE by 1.32% in mean Dice and demonstrates the best performance across all pretraining strategies. Instead of constraining same class representations to move near only, our contrastive loss allows to push the representations out if they are not in the same class and provide a better definition on separating embeddings than pretraining with CE.

To further evaluate the segmentation using different contrastive learning approaches, the qualitative representation of the segmentation prediction with each training method is demonstrated on Fig. 5 comparing with the ground truth label. With SSCL, the boundary of the segmentation is significantly smoother than that of with RI. However, we found that additional segmentation is performed near the neighboring structures. The similar intensity range and morphological appearance may lead to the instability of representation extraction from SSCL. Pretext task with CE demonstrates a significant improvement in label quality, while the boundaries on particular organs (e.g. gall bladder) is not well preserved. With the additional constraints by AGCL, the boundary information between neighboring organs are clearly defined and the segmentation quality is comparable to the ground truth label.

Comparing with different constraints scenario in contrastive loss: We further perform evaluation in our proposed contrastive loss with class-wise label constraint: 1) modality-only label constraint, 2) organ-only label constraint, and 3) modality plus organ constraints with ResNet-50 encoder as the network backbone. As shown in Table III, the overall superior performance is achieved when applying both modality and organ constraints.

Comparing with reduced label for AGCL: In Fig. 6(a), we perform AGCL with the variation of label quantity and compare the segmentation performance by leveraging the amount of label information. We observe that the segmentation model has the best performance with fully labeled input. A significant improvement is shown with 20% labels for AGCL comparing to that with 10% labels, while an improvement to a small extent is demonstrated by using 50% label for AGCL.

Comparing with temperature variability: We experiment with the variation of temperature to investigate the optimal effect towards the segmentation performance. Fig. 6(b) demonstrates the effect of temperature on the multi-organ segmentation across all subjects in the BTCV testing dataset. We observe that low temperature achieves better performance than high temperature, as the radius of the hypersphere defined in the latent space is inversely proportional to the temperature scaling, which increases the difficulty of finding positive samples with the decrease of radius.

Comparing with single/multiple modal contrastive learning: The segmentation performance is evaluated with single modality and with multi-modality contrastive learning respectively. From Fig. 6(c), a better segmentation performance for contrast-enhanced dataset is achieved by contrastive learning with multi-modality images. Interestingly, we observe that the segmentation performance of non-contrast imaging is improved to a small extent with non-contrast modal pre-training only.

C. Discussion & Limitations—In this work, we present a co-training framework that leverages organ attention into contrastive learning and defines representations into conditional embeddings with image-level labels only. We hypothesize that the conditional embeddings defined are beneficial to the downstream segmentation task. By using organ attention as an additional input channel, we can extract meaningful representation within the organ-specific regions, instead of randomly extracting representations that may affect by the neighboring organs.

It also allows to learn and define the organ-specific context into corresponding semantic categories. Apart from using organ attention, we further leverage the multi-class labels to constrain pairwise representations into sub-class embeddings. Instead of constraining contrastive loss in pixel-wise setting, we demonstrate that constraining the latent space with multiple image-level labels is also beneficial to enhance the segmentation performance for each organ of interest. From Table I, we have shown that our proposed learning scheme outperforms the current contrastive learning state-of-the-art for multi-organ segmentation. Furthermore, Table II has shown the comparison of different pre-training strategies with/out multi-class image-level labels. It provides a better understanding about the contribution of our proposed contrastive loss in defining semantic-aware latent space for segmentation task.

Although AGCL tackles current challenges of integrating contrastive learning into multi-object segmentation, limitations still exist in the process of AGCL. One limitation is the dependency of the coarse segmentation quality. As 2D patches are extracted with the attention information in each slice, patches without corresponding organ regions may also be possible to extract due to inaccurate coarse segmentation. Incorrect label definition inputs may bring into contrastive learning process. Another limitation is performing contrastive learning in object-centric setting. We aim to innovate con- trastive learning strategy with complete volume inputs for multi-object segmentation in our future work.

V. Conclusion

Performing robust multi-object semantic segmentation using deep learning remains a persistent challenge. In this work, we propose a novel semantic-aware contrastive framework that extends self-supervised contrastive loss to adapt multiple semantic meanings into the learned features and integrates attention guidance from coarse segmentation to extract organspecific features. Our proposed method leads to a significant gain in segmentation performance on two public contrast-enhanced CT datasets and one in-house non-contrast CT dataset.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research is supported by NIH Common Fund and National Institute of Diabetes, Digestive and Kidney Diseases U54DK120058, NSF CAREER 1452485, NIH 2R01EB006136, NIH 1R01EB017230, and NIH RO1NS09529. ImageVU and RD are supported by the VICTR CTSA award (ULTR000445 from NCATS/NIH). We gratefully acknowledge the support of NVIDIA Corportaion with the donation of the Titan X Pasacl GPU usage.

References

[1]. Chen T, Kornblith S, Norouzi M, and Hinton G, “A simple framework for contrastive learning of visual representations,” in International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[2]. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, Maschinot A, Liu C, and Krishnan D, “Supervised contrastive learning,” arXiv preprint arXiv:2004.11362, 2020.

[3]. Chuang C-Y, Robinson J, Lin Y-C, Torralba A, and Jegelka S, “Debiased contrastive learning,” in Advances in Neural Information Processing Systems, Larochelle H, Ranzato M, Hadsell R, Balcan MF, and Lin H, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 8765–8775

[4]. You Y, Chen T, Sui Y, Chen T, Wang Z, and Shen Y, “Graph contrastive learning with augmentations,” in Advances in Neural Information Processing Systems, Larochelle H, Ranzato M, Hadsell R, F Balcan M, and Lin H, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 5812–5823.

[5]. Chen X, Yao L, Zhou T, Dong J, and Zhang Y, “Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images,” Pattern Recognition, vol. 113, p. 107826, 2021.

[6]. Taleb A, Loetzsch W, Danz N, Severin J, Gaertner T, Bergner B, and Lippert C, “3d self-supervised methods for medical imaging,” arXiv preprint arXiv:2006.03829, 2020.

[7]. Chaitanya K, Erdil E, Karani N, and Konukoglu E, “Contrastive learning of global and local features for medical image segmentation with limited annotations,” in Advances in Neural Information Processing Systems, Larochelle H, Ranzato M, Hadsell R, Balcan MF, and Lin H, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12546–12558.

[8]. Liu W, Ferstl D, Schulter S, Zebedin L, Fua P, and Leistner C, “Domain adaptation for semantic segmentation via patch-wise contrastive learning,” 2021.

[9]. Zhao X, Vemulapalli R, Mansfield P, Gong B, Green B, Shapira L, and Wu Y, “Contrastive learning for label-efficient semantic segmentation,” 2021.

[10]. Landman B, Xu Z, Igelsias J, Styner M, Langerak T, and Klein A, “Miccai multi-atlas labeling beyond the cranial vault-workshop and challenge,” in Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge, 2015.

[11]. Ma J, Zhang Y, Gu S, Zhu C, Ge C, Zhang Y, An X, Wang C, Wang Q, Liu X et al., “Abdomenct-1k: Is abdominal organ segmentation a solved problem,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[12]. He K, Zhang X, Ren S, and Sun J, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[13]. Zhang R, Isola P, and Efros AA, “Split-brain autoencoders: Unsupervised learning by cross-channel prediction,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017 pp. 1058–1067.

[14]. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, and Joulin A, “Emerging properties in self-supervised vision transformers,” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9650–9660.

[15]. He K, Chen X, Xie S, Li Y, Dollár P, and Girshick R, “Masked autoencoders are scalable vision learners,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.

[16]. Zhao X, Vemulapalli R, Mansfield P, Gong B, Green B, Shapira L, and Wu Y, “Contrastive learning for label-efficient semantic segmentation,” arXiv preprint arXiv:201206985, 2020.

[17]. Bachman P, Hjelm RD, and Buchwalter W, “Learning representations by maximizing mutual information across views,” arXiv preprint arXiv: 190600910, 2019.

[18]. Misra I and L. v. d. Maaten, “Self-supervised learning of pretextinvariant representations,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6707–6717.

[19]. Wang X, Zhang R, Shen C, Kong T, and Li L, “Dense contrastive learning for self-supervised visual pre-training,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 3024–3033.

[20]. Wang W, Zhou T, Yu F, Dai J, Konukoglu E, and Van Gool L, “Exploring cross-image pixel contrast for semantic segmentation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7303–7313.

[21]. Hu H, Cui J, and Wang L, “Region-aware contrastive learning for semantic segmentation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16291–16301.

[22]. Alonso I, Sabater A, Ferstl D, Montesano L, and Murillo AC, “Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8219–8228.

[23]. Zeng D, Wu Y, Hu X, Xu X, Yuan H, Huang M, Zhuang J, Hu J, and Shi Y, “Positional contrastive learning for volumetric medical image segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021, pp 221–230.

[24]. Hu X, Zeng D, Xu X, and Shi Y, “Semi-supervised contrastive learning for label-efficient medical image segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021, pp. 481–490.

[25]. Liu Z, Zhu Z, Zheng S, Liu Y, Zhou J, and Zhao Y, “Margin preserving self-paced contrastive learning towards domain adaptation for medical image segmentation,” IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 2, pp. 638–647, 2022. [PubMed: 34990372]

[26]. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, and Ronneberger O, “3d u-net: learning dense volumetric segmentation from sparse annotation,” in International conference on medical image computing and computer-assisted intervention. Springer, 2016, pp. 424–432.

[27]. Gibson E, Giganti F, Hu Y, Bonmati E, Bandula S, Gurusamy K, Davidson B, Pereira SP, Clarkson MJ, and Barratt DC, “Automatic multi-organ segmentation on abdominal ct with dense v-networks,” IEEE transactions on medical imaging, vol. 37, no. 8, pp. 1822–1834, 2018. [PubMed: 29994628]

[28]. Zhou Z, Siddiquee MMR, Tajbakhsh N, and Liang J, “Unet++: Redesigning skip connections to exploit multiscale features in image segmentation,” IEEE transactions on medical imaging, vol. 39, no. 6, pp. 1856–1867, 2019. [PubMed: 31841402]

[29]. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen Y-W, and Wu J, “Unet 3+: A full-scale connected unet for medical image segmentation,” in ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE, 2020, pp. 1055–1059.

[30]. Gamechi ZS, Bons LR, Giordano M, Bos D, Budde RP, Kofoed KF, Pedersen JH, Roos-Hesselink JW, and de Bruijne M, “Automated 3d segmentation and diameter measurement of the thoracic aorta on non-contrast enhanced ct,” European radiology, vol. 29, no. 9, pp. 4613–4623, 2019. [PubMed: 30673817]

[31]. Roth HR, Shen C, Oda H, Sugino T, Oda M, Hayashi Y, Misawa K, and Mori K, “A multi-scale pyramid of 3d fully convolutional networks for abdominal multi-organ segmentation,” in International conference on medical image computing and computer-assisted intervention. Springer, 2018, pp. 417–425.

[32]. Zhu Z, Xia Y, Xie L, Fishman EK, and Yuille AL, “Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma,” in International conference on medical image computing and computer-assisted intervention Springer, 2019, pp. 3–12.

[33]. Tang Y, Gao R, Lee HH, Han S, Chen Y, Gao D, Nath V, Bermudez C, Savona MR, Abramson RG et al. , “High-resolution 3d abdominal segmentation with random patch network fusion,” Medical Image Analysis, vol. 69, p. 101894, 2021.

[34]. Wang Y, Zhou Y, Shen W, Park S, Fishman EK, and Yuille AL, “Abdominal multi-organ segmentation with organ-attention networks and statistical fusion,” Medical image analysis, vol. 55, pp. 88–102, 2019. [PubMed: 31035060]

[35]. Xue Y, Tang H, Qiao Z, Gong G, Yin Y, Qian Z, Huang C, Fan W and Huang X, “Shape-aware organ segmentation by predicting signed distance maps,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 12565–12572.

[36]. Lee HH, Tang Y, Bao S, Abramson RG, Huo Y, and Landman BA, “Rap-net: Coarse-to-fine multi-organ segmentation with single random anatomical prior,” in 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) IEEE, 2021, pp. 1491–1494.

[37]. Isensee F, Jaeger PF, Kohl SA, Petersen J, and Maier-Hein KH, “nnu-net: a self-configuring method for deep learning-based biomedical image segmentation,” Nature methods, vol. 18, no. 2, pp. 203–211, 2021. [PubMed: 33288961]

[38]. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth HR, and Xu D, “Unetr: Transformers for 3d medical image segmentation,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 574–584.

[39]. Zhou Y, Li Z, Bai S, Wang C, Chen X, Han M, Fishman E, and Yuille AL, “Prior-aware neural network for partially-supervised multi-organ segmentation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.

[40]. Lee HH, Tang Y, Tang O, Xu Y, Chen Y, Gao D, Han S, Gao R, Savona MR, Abramson RG et al., “Semi-supervised multi-organ segmentation through quality assurance supervision,” in Medical Imaging 2020: Image Processing, vol. 11313. International Society for Optics and Photonics, 2020, p. 113131I.

[41]. Zhou Z, Sodha V, Pang J, Gotway MB, and Liang J, “Models genesis,” Medical image analysis, vol. 67, p. 101840, 2021

[42]. Bai W, Chen C, Tarroni G, Duan J, Guitton F, Petersen SE, Guo Y, Matthews PM, and Rueckert D, “Self-supervised learning for cardiac mr image segmentation by anatomical position prediction,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 541–549.

[43]. Zhuang X, Li Y, Hu Y, Ma K, Yang Y, and Zheng Y, “Self-supervised feature learning for 3d medical images by playing a rubik’s cube,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 420–428.

[44]. Chen L-C, Papandreou G, Schroff F, and Adam H, “Rethinking atrous convolution for semantic image segmentation,” arXiv preprint arXiv:1706.05587, 2017

[45]. Zhou X, Ito T, Takayama R, Wang S, Hara T, and Fujita H, “Three-dimensional ct image segmentation by combining 2d fully convolutional network with 3d majority voting,” in Deep Learning and Data Labeling for Medical Applications. Springer, 2016, pp. 111–120.

[46]. Heinrich MP, Maier O, and Handels H, “Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities.” VISCERAL Challenge@ISBI, vol. 1390, p. 27, 2015.

[47]. Pawlowski N, Ktena SI, Lee MCH, Kainz B, Rueckert D, Glocker B, and Rajchl M, “Dltk: State of the art reference implementations for deep learning on medical images,” 2017.

[48]. Tang Y, Gao R, Han S, Chen Y, Gao D, Nath V, Bermudez C, Savona MR, Bao S, Lyu I et al. , “Body part regression with self-supervision,” IEEE Transactions on Medical Imaging, vol. 40, no. 5, pp. 1499–1507, 2021. [PubMed: 33560981]

[49]. Chaitanya K, Erdil E, Karani N, and Konukoglu E, “Contrastive learning of global and local features for medical image segmentation with limited annotations,” arXiv preprint arXiv:200610511, 2020.

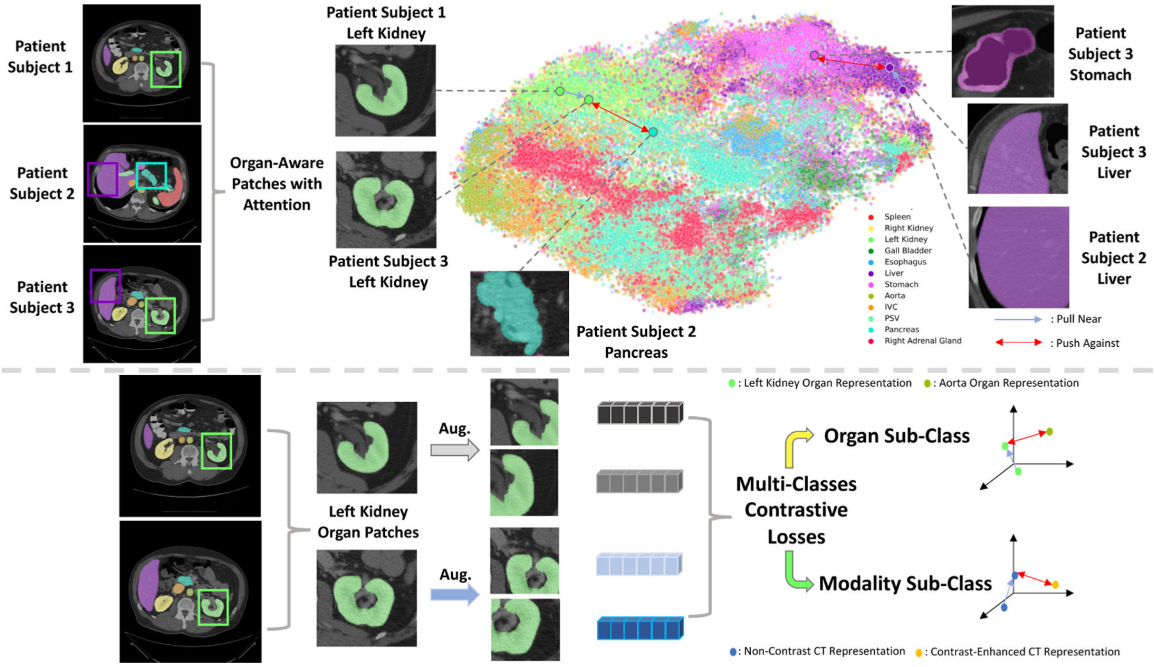

Patient Subject 1
Patient Subject 2
Patient Subject 3
Organ-Aware Patches with Attention
Patient Subject 1 Left Kidney
Patient Subject 3 Left Kidney
Patient Subject 2 Pancreas
Patient Subject 3 Stomach
Patient Subject 3 Liver
Patient Subject 2 Liver
● Spleen
● Right Kidney
● Left Kidney
● Gall Bladder
● Esophagus
● Liver
● Stomach
● Aorta
● IVC
● PSV
● Pancreas
● Right Adrenal Gland
→ : Pull Near
↔ : Push Against
● : Left Kidney Organ Representation
● : Aorta Organ Representation
Left Kidney Organ Patches
Aug.
Aug.
Organ Sub-Class
Multi-Classes Contrastive Losses
Modality Sub-Class
● : Non-Contrast CT Representation
● : Contrast-Enhanced CT Representation

Fig. 1:

With multiple organs located in a single image, organ attention maps guide their representations into corresponding embeddings and adapt contrastive learning for multi-object segmentation. Categorical information can be used for supervisory context to constrain the separation of clusters (grey arrow: pull the matching representations together, red arrow: push the non-matching representations apart).

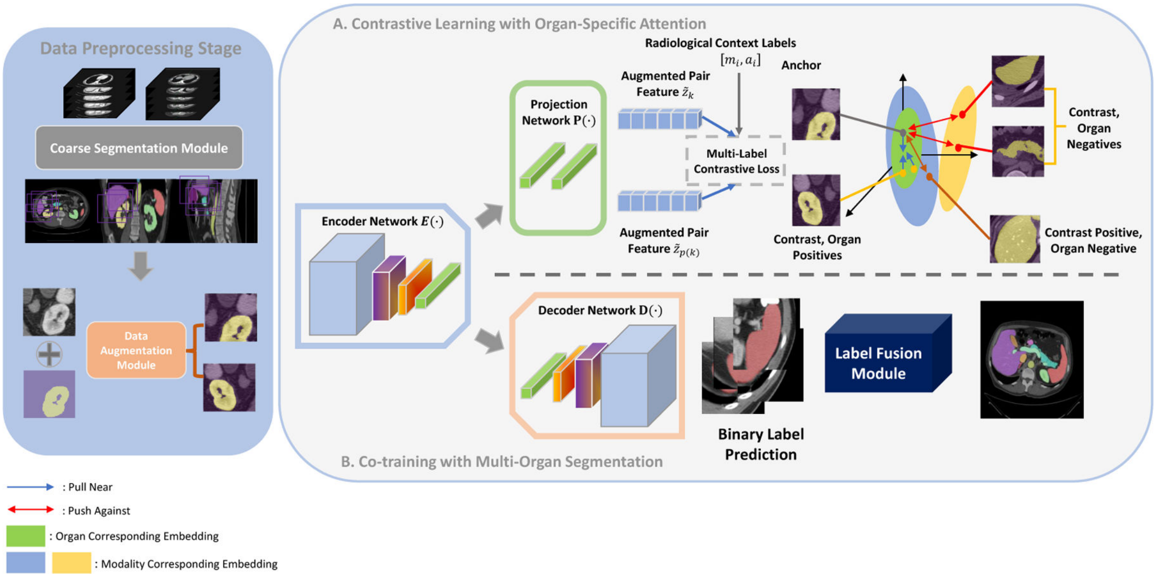

Data Preprocessing Stage
Coarse Segmentation Module
Data Augmentation Module
A. Contrastive Learning with Organ-Specific Attention
Radiological Context Labels
$[m_i, a_i]$
Anchor
Augmented Pair Feature $\tilde{z}_k$
Projection Network P(·)
Multi-Label Contrastive Loss
Contrast, Organ Negatives
Encoder Network $E(\cdot)$
Augmented Pair Feature $\tilde{z}_{p(k)}$
Contrast, Organ Positives
Contrast Positive, Organ Negative
Decoder Network D(·)
Label Fusion Module
Binary Label Prediction
B. Co-training with Multi-Organ Segmentation
: Pull Near
: Push Against
: Organ Corresponding Embedding
: Modality Corresponding Embedding

Fig. 2:

A 2D/3D segmentation pipeline (2D for natural image, 3D for medical image) is used to generate attention map for organ localization. 2D organ-corresponding query patches are randomly extracted and concatenated with the regional attention maps as an additional channel to guide embeddings of the organ targets. Data augmented pairs of the attention queries are constrained into corresponding radiological embeddings (such as organs and modalities) with additional label supervision in the proposed contrastive loss. The encoder is co-trained with decoder to generate refined segmentation with label fusion.

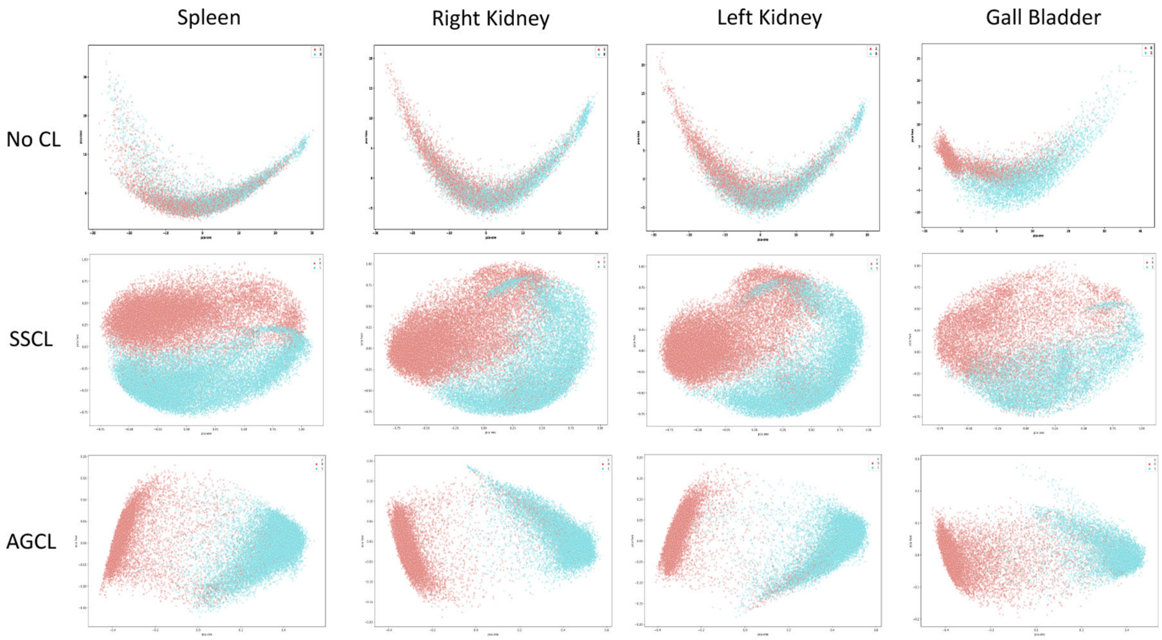


Fig. 3:

The latent distributions of four randomly selected organs using principal component analysis (PCA) are plotted with their corresponding modality (Blue: contrast-enhanced phase CT, red: non-contrast phase CT). The first two components are plotted as a visualization. With AGCL, the organ representation can be well separated into specific modal clusters.

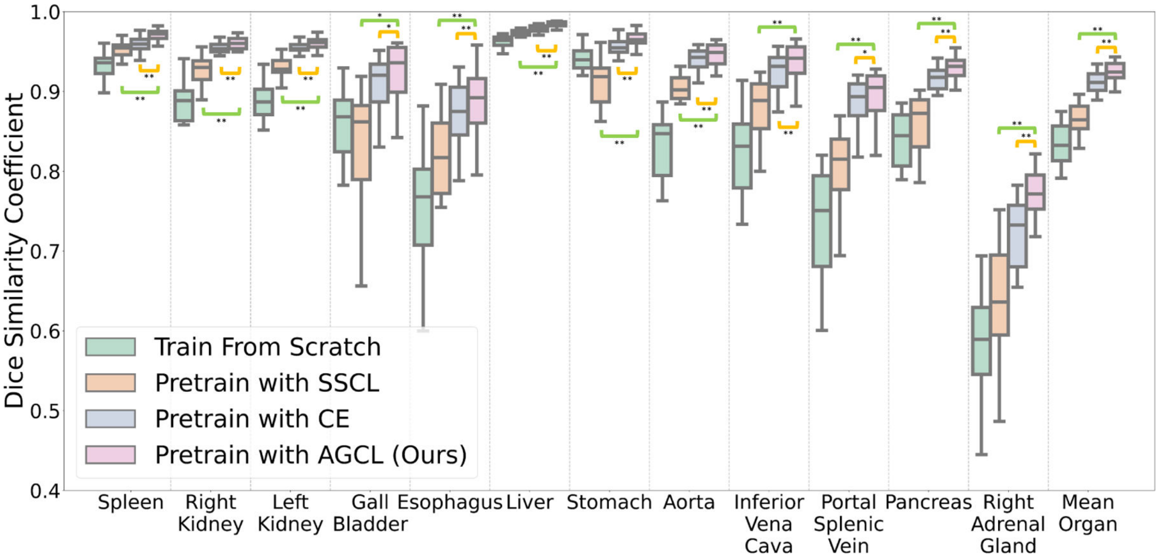


Fig. 4:

Comparison of different supervised / self-supervised pre-training strategies using multi-class labels for multi-organ segmentation. AGCL outperforms the current state-of-the-art pre-training methods with SSCL and classification pre-training with CE across all organs. (*: $p < 0.05$, **: $p < 0.01$, with Wilcoxon signed-rank test)

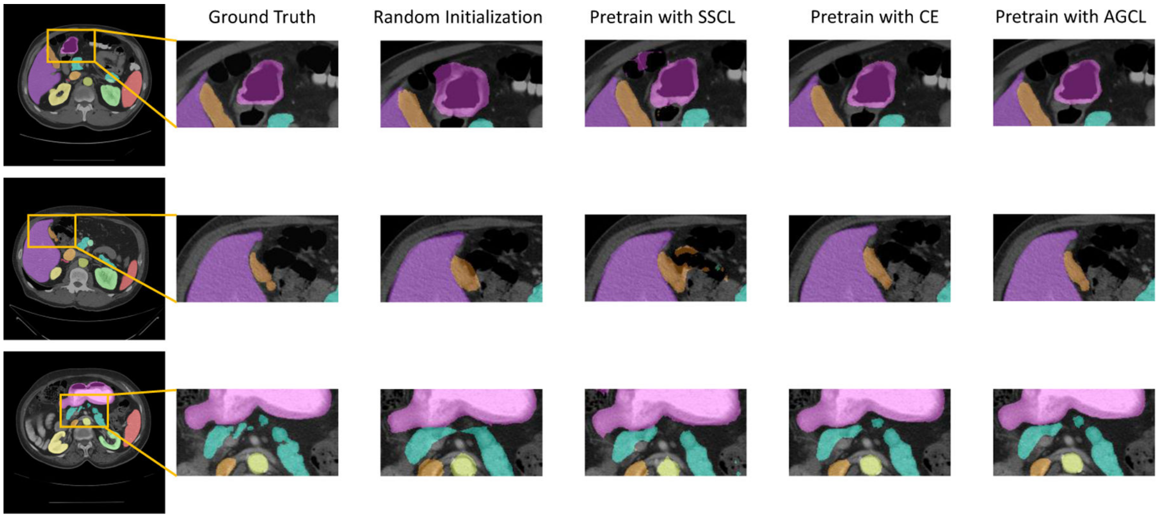
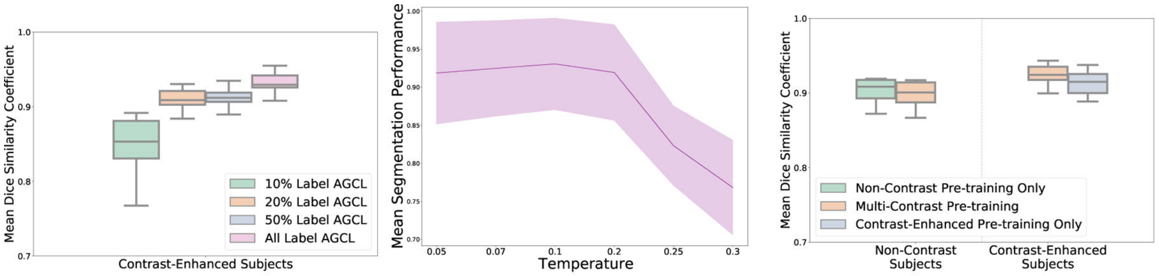


Fig. 5:

Qualitative representations of different pretraining strategies are demonstrated with ResNet-50 encoder backbone. Incremental improvement on segmentation quality is shown and AGCL demonstrates smooth boundaries and accurate morphological information between neighboring organs.



(a) (b) (c)

Fig. 6:

a) The segmentation performance gradually improved with the additional quantities of image-level labels for AGCL. **b)** Ablation studies of temperature scaling the distance between positive/negative pairs demonstrates that the segmentation performance is best optimized when $T = 0.1$. **c)** Performance trade-off is demonstrated between non-contrast and contrastenhanced CTs with multi-modal training.

TABLE I:

Comparison of the fully-supervised, unsupervised, semi-supervised and partially supervised state-of-the-art methods on the 2015 MICCAI BTCV challenge leaderboard. (We show 8 main organs Dice scores due to limited space, *: fullysupervised approach, ○: unsupervised approach, Δ: partially supervised approach, ◊: contrastive learning approach.)

Method	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Aorta	IVC	Average Dice	Mean Surface Distance	Hausdorff Distance
Heinrich et al. ○ [46]	0.920	0.894	0.915	0.604	0.692	0.948	0.857	0.828	0.790	2.262	25.504
Cicek et al. * [26]	0.906	0.857	0.899	0.644	0.684	0.937	0.886	0.808	0.784	2.339	15.928
Roth et al. * [31]	0.935	0.887	0.944	0.780	0.712	0.953	0.880	0.804	0.816	2.018	17.982
Pawlowski et al. * [47]	0.939	0.895	0.915	0.711	0.743	0.962	0.891	0.826	0.815	1.861	62.872
Zhu et al. * [32]	0.935	0.886	0.944	0.764	0.714	0.942	0.879	0.803	0.814	1.692	18.201
Lee et al. * [36]	0.959	0.920	0.945	0.768	0.783	0.962	0.910	0.847	0.842	1.501	16.433
Isensee et al. * [37]	0.956	0.923	0.940	0.760	0.764	0.965	0.905	0.850	0.839	1.523	16.201
Hat. et al. ∗ [38]	0.959	0.912	0.940	0.724	0.746	0.968	0.905	0.840	0.836	1.602	16.355
Zhou et al. Δ [39]	0.968	0.920	0.953	0.729	0.790	0.974	0.925	0.847	0.850	1.450	18.468
Chaitanya et al. ◊ [7]	0.956	0.935	0.946	0.920	0.854	0.970	0.915	0.893	0.874	1.236	15.281
Alonso et al. ◊ [22]	0.954	0.933	0.932	0.903	0.858	0.973	0.918	0.904	0.890	1.291	15.032
Wang et al. ◊ [19]	0.963	0.939	0.900	0.815	0.838	0.976	0.922	0.907	0.882	1.303	14.759
Khosla et al. ◊ [2]	0.959	0.939	0.947	**0.932**	0.867	0.978	0.922	0.911	0.907	0.978	14.136
Wang et al. ◊ [20]	0.966	0.942	0.955	0.886	0.860	0.975	0.930	0.908	0.913	0.991	13.785
Ours (SSCL w/o Attention)	0.940	0.909	0.918	0.740	0.790	0.962	0.890	0.854	0.838	1.967	17.773
Ours (SSCL w/ GT Attention)	0.947	0.910	0.928	0.805	0.815	0.967	0.890	0.861	0.858	2.013	18.101
Ours (SSCL Pretraining)	0.953	0.922	0.930	0.830	0.822	0.972	0.899	0.874	0.863	1.899	17.073
Ours (SSCL Co-training)	0.957	0.930	0.935	0.843	0.827	0.974	0.895	0.869	0.869	1.803	16.714
Ours (AGCL w/o Attention)	0.950	0.933	0.935	0.803	0.805	0.967	0.903	0.862	0.852	1.923	17.215
Ours (AGCL w/ GT Attention)	0.962	0.937	0.940	0.857	0.846	0.973	0.915	0.890	0.891	1.312	14.852
Ours (AGCL Pretraining)	0.971	0.955	**0.963**	0.910	0.886	0.984	**0.941**	**0.932**	0.923	0.932	**13.024**
Ours (AGCL Co-training)	**0.975**	**0.958**	0.960	0.914	**0.890**	**0.985**	0.937	0.920	**0.926**	**0.927**	13.102

TABLE II:

Ablation studies of segmentation performance in various network backbones of the BTCV testing cohort.

Encoder	Pretrain	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stomach	Aorta	IVC	PSV	Pancreas	R.A	Mean
ResNet50	×	0.932	0.877	0.887	0.860	0.761	0.962	0.941	0.832	0.815	0.735	0.833	0.587	0.840
ResNet50	SSCL	0.953	0.922	0.930	0.842	0.822	0.972	0.907	0.899	0.874	0.800	0.854	0.625	0.868
ResNet50	CE	0.959	0.948	0.957	0.890	0.868	0.978	0.956	0.935	0.919	0.884	0.903	0.725	0.905
ResNet50	AGCL	**0.971**	**0.955**	**0.963**	**0.910**	**0.886**	**0.984**	**0.965**	**0.941**	**0.932**	**0.893**	**0.917**	**0.769**	**0.923**
ResNet101	×	0.939	0.870	0.880	0.859	0.745	0.960	0.915	0.840	0.800	0.736	0.825	0.567	0.834
ResNet101	SSCL	0.950	0.928	0.935	0.805	0.792	0.969	0.900	0.905	0.877	0.800	0.846	0.602	0.868
ResNet101	CE	0.960	0.933	0.945	0.887	0.822	0.975	0.952	0.920	0.901	0.834	0.877	0.670	0.891
ResNet101	AGCL	**0.965**	**0.948**	**0.954**	**0.901**	**0.875**	**0.981**	**0.962**	**0.930**	**0.917**	**0.876**	**0.902**	**0.748**	**0.917**

TABLE III:

Ablation studies of segmentation performance with adapting different constraints in contrastive loss.

Label Constraints	**Spleen**	**R.Kid**	**L.Kid**	**Gall.**	**Eso.**	**Liver**	**Stomach**	**Aorta**	**IVC**	**PSV**	**Pancreas**	**R.A**	**Mean**
Modality	0.965	0.948	0.958	0.890	0.865	0.975	0.950	0.930	0.920	0.879	0.895	0.732	0.910
Organ	0.968	0.947	0.959	0.893	0.872	0.979	0.957	0.935	**0.925**	0.886	0.903	0.744	0.914
Modality + Organ	**0.975**	**0.958**	**0.960**	**0.914**	**0.890**	**0.985**	**0.970**	**0.937**	0.920	**0.901**	**0.925**	**0.772**	**0.926**

TABLE IV:

Comparison of the current contrastive state-of-the-art methods on FLARE dataset.

Method	**Spleen**	**Kidney**	**Liver**	**Pancreas**	**Average Dice**
Lee et al. [36]	0.956	0.903	0.954	0.730	0.885
Chai. et al. [49]	0.961	0.923	0.956	0.787	0.908
Wang et al. [19]	0.966	0.918	0.964	0.800	0.912
Khosla et al. [2]	0.963	0.918	0.966	0.830	0.919
Wang et al. [20]	0.968	0.940	0.964	0.811	0.922
Ours (SSCL)	0.960	0.910	0.960	0.756	0.896
Ours (AGCL)	**0.975**	**0.952**	**0.971**	**0.835**	**0.933**