

CCQ: Cross-Class Query Network for Partially Labeled Organ Segmentation

Xuyang Liu^{1*}, Bingbing Wen^{2*}, Sibe Yang^{1,3†}

¹ School of Information Science and Technology, ShanghaiTech University

² Information School, University of Washington

³ Shanghai Engineering Research Center of Intelligent Vision and Imaging
liuxy15@shanghaitech.edu.cn, bingbw@uw.edu, yangsb@shanghaitech.edu.cn

Abstract

Learning multi-organ segmentation from multiple partially-labeled datasets attracts increasing attention. It can be a promising solution for the scarcity of large-scale, fully labeled 3D medical image segmentation datasets. However, existing algorithms of multi-organ segmentation on partially-labeled datasets neglect the semantic relations and anatomical priors between different categories of organs, which is crucial for partially-labeled multi-organ segmentation. In this paper, we tackle the limitations above by proposing the Cross-Class Query Network (CCQ). CCQ consists of an image encoder, a cross-class query learning module, and an attentive refinement segmentation module. More specifically, the image encoder captures the long-range dependency of a single image via the transformer encoder. Cross-class query learning module first generates query vectors that represent semantic concepts of different categories and then utilizes these query vectors to find the class-relevant features of image representation for segmentation. The attentive refinement segmentation module with an attentive skip connection incorporates the high-resolution image details and eliminates the class-irrelevant noise. Extensive experiment results demonstrate that CCQ outperforms all the state-of-the-art models on the MOTS dataset, which consists of seven organ and tumor segmentation tasks. Code is available at <https://github.com/Yang-007/CCQ.git>

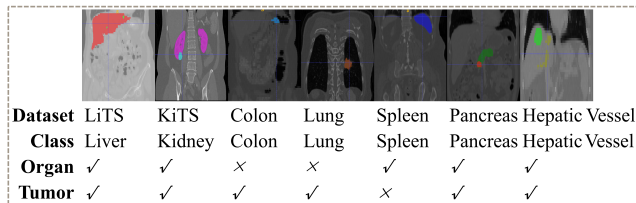
Introduction

Medical image segmentation is often an essential first step of computer-aided detection (Litjens et al. 2017). Automated segmentation of multiple abdominal organs on computed tomography (CT) is of great help to clinical applications such as surgery and radiotherapy. Algorithms of multi-class segmentation of abdominal organs (Gibson et al. 2018; Xie et al. 2021) rely on fully-labeled datasets with human annotations for several organs. However, the full annotation of medical images is extremely expensive and time-consuming since it needs to be created and checked by professional radiologists, which leads to the scarcity of large-scale, fully-labeled datasets, especially for 3D medical image segmentation of abdominal organs. Several partially-labeled datasets

*These authors contributed equally.

†Corresponding author.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



Dataset	LiTS	KiTS	Colon	Lung	Spleen	Pancreas	Hepatic Vessel
Class	Liver	Kidney	Colon	Lung	Spleen	Pancreas	Hepatic Vessel
Organ	✓	✓	×	×	✓	✓	✓
Tumor	✓	✓	✓	✓	×	✓	✓

Figure 1: Illustration of partially labeled multi-organ and tumor segmentation.

are available, *e.g.*, LiTS (Bilic et al. 2019) with liver organ and tumor, KiTS (Heller et al. 2019) with kidney organ and tumor, Colon (Simpson et al. 2019) with colon tumor, Lung (Simpson et al. 2019) with lung tumor, Spleen (Simpson et al. 2019) with spleen organ, Pancreas (Simpson et al. 2019) with pancreas organ and tumor, and Hepatic Vessel (Simpson et al. 2019) with hepatic vessel organ and tumor. Therefore, learning multi-organ segmentation from multiple partially-labeled datasets becomes a promising solution and attracts increasing attention. An illustration of partially-labeled datasets is shown in Figure 1.

Existing approaches of multi-organ segmentation over partially-labeled datasets can be roughly divided into two types based on their designing principles: multiple separate models for different organs (Ledig et al. 2015; Wang et al. 2021; Kushnure and Talbar 2021; Wardhana et al. 2021) and a single model for multiple organs (Chen, Ma, and Zheng 2019; Fang and Yan 2020; Shi et al. 2021; Zhou et al. 2019; Liu, Xiao, and Zhou 2021; Dmitriev, Kaufman et al. 2019; Zhang et al. 2021). The former is not only time-consuming and memory-intensive but neglects anatomical priors (Zhou et al. 2019) which are crucial for accurate segmentation. In this paper, we focus on the latter. Specifically, it is challenging to learn a single multi-organ segmentation model based on the combination of several partially-labeled datasets due to significant differences between diverse datasets and organs. To alleviate such differences and inconsistencies, we further explore the single model approach for multiple organs and find two main directions: first is class-relevant representation learning methods (Dmitriev, Kaufman et al. 2019; Chen, Xu, and Koltun 2017) and second is non-class-relevant representation learn-

ing methods (Fang and Yan 2020; Chen, Ma, and Zheng 2019; Zhang et al. 2021). Non-class-relevant representation learning methods like TAL (Fang and Yan 2020), Multi-Head (Chen, Ma, and Zheng 2019) and DoDNet (Zhang et al. 2021) do not learn the class-relevant representation of the input image even though they utilize class information. We focus on class-relevant representation learning methods in this study.

Class-relevant representation learning methods (Dmitriev, Kaufman et al. 2019; Chen, Xu, and Koltun 2017) embed the class information into their representation of images to perform class-conditional encoding or decoding. However, their embeddings of different organs are independent, which neglects the explicit semantic relations and anatomical priors (*e.g.*, relative locations and sizes) between different classes of organs. Such relations and priors are particularly crucial for partially labeled multi-organ segmentation.

To address these issues mentioned above, we propose a Cross-Class Query network (CCQ), which focuses on embedding the class relations between multi-organ or tumors and performing the attentive segmentation over partially labeled CT abdominal datasets by generating learnable query vectors that represent semantic concepts of different categories of organs. CCQ consists of three modules: image encoder, cross-class query learning, and attentive refinement segmentation. Specifically, image encoder captures the image features with the long-range dependency via the hybrid of CNN encoder and Transformer, leading to more effective image representation. Second, cross-class query learning module first learns the cross-class semantic concepts over different organ and tumor tasks via constructing a set of learnable vectors, then generates class-relevant query vectors by incorporating class information and capturing the relations among semantic concepts via transformer self-attention. Cross-class query learning takes the class-relevant query vectors as queries of the attention module to query the image representation and obtains class-relevant representation for segmentation. Third, attentive refinement segmentation module decodes the class-relevant representation to segmentation results via attentive refinement segmentation, incorporating high-resolution image details without introducing the class-irrelevant noise.

Our contributions are summarised as follows:

- To the best of our knowledge, we are the first to model cross-class semantic concepts for multiple classes in medical image segmentation. Cross-Class Query network (CCQ) focuses on generating class-relevant query vectors by incorporating class information and capturing the relations among semantic concepts. These semantic concepts, relations, and anatomical priors contribute significantly to fully understanding and utilizing partially labeled medical image segmentation.
- We propose an attentive refinement segmentation to incorporate high-resolution image details into low-resolution. And we apply class-relevant semantic queries to generate high-resolution semantic segmentation results without introducing the class-irrelevant noise to improve segmentation accuracy.

- Extensive experiment results demonstrate that CCQ outperforms all the state-of-the-art models on the MOTs dataset consisting of seven organ and tumor segmentation tasks.

Related Work

Partially Labeled Medical Image Segmentation. Accurate segmentation of multiple organs and tumors is essential for clinical practice. Many pioneering works have been proposed for multi-organ or multi-tumor segmentation in a fully-labeled setting where manual annotations for multiple organs or tumors are available (Gibson et al. 2018; Xie et al. 2021). They are mainly based on fully convolutional semantic segmentation frameworks of natural images (Ronneberger, Fischer, and Brox 2015) and apply prior knowledge to achieve better multi-organ segmentation, *e.g.*, statistical fusion from different views (Wang et al. 2019), complementary learning of extra distance maps and contour maps (Navarro et al. 2019), and local structure exploitation via attention mechanism (Schlemper et al. 2019). However, most public datasets are partially labeled where not all organs but a few organs are labeled because fully-labeled annotation is expensive and time-consuming. Existing approaches of multi-organ segmentation over partially-labeled datasets can be roughly divided into two types based on their designing principles: multiple separate models for different organs (Ledig et al. 2015; Wang et al. 2021; Kushnure and Talbar 2021; Wardhana et al. 2021) and a single model for multiple organs (Chen, Ma, and Zheng 2019; Fang and Yan 2020; Shi et al. 2021; Zhou et al. 2019; Liu, Xiao, and Zhou 2021; Dmitriev, Kaufman et al. 2019; Zhang et al. 2021). Learning multiple separate networks for multiple partially labeled datasets (Ledig et al. 2015; Wang et al. 2021; Kushnure and Talbar 2021; Wardhana et al. 2021) is intuitive. However, such a strategy of separate learning is time-consuming and memory-intensive, more importantly, it neglects natural anatomical priors (Litjens et al. 2017; Zhou et al. 2019). In this study, we focus on building a single network for multiple partially labeled datasets.

Class-Relevant Representation Learning for Multiple Organs. Some recent works for partially labeled segmentation focus on training a single network for multiple partially labeled datasets and exploring the task via multi-task learning. There are two main directions for constructing the single model approach for multiple organs: first is class-relevant representation learning methods (Dmitriev, Kaufman et al. 2019; Chen, Xu, and Koltun 2017) and second is non-class-relevant representation learning methods (Fang and Yan 2020; Chen, Ma, and Zheng 2019; Zhang et al. 2021). Non-class-relevant representation learning methods like TAL (Fang and Yan 2020), Multi-Head (Chen, Ma, and Zheng 2019) and DoDNet (Zhang et al. 2021) don't learn the class-relevant representation of the input image even if they though utilize class information. Most works use a multi-head architecture which is applied by sharing the encoder or backbone between all organs while keeping several organ-specific decoders (Chen, Ma, and Zheng 2019) or last segmentation layers (Fang and Yan 2020; Shi

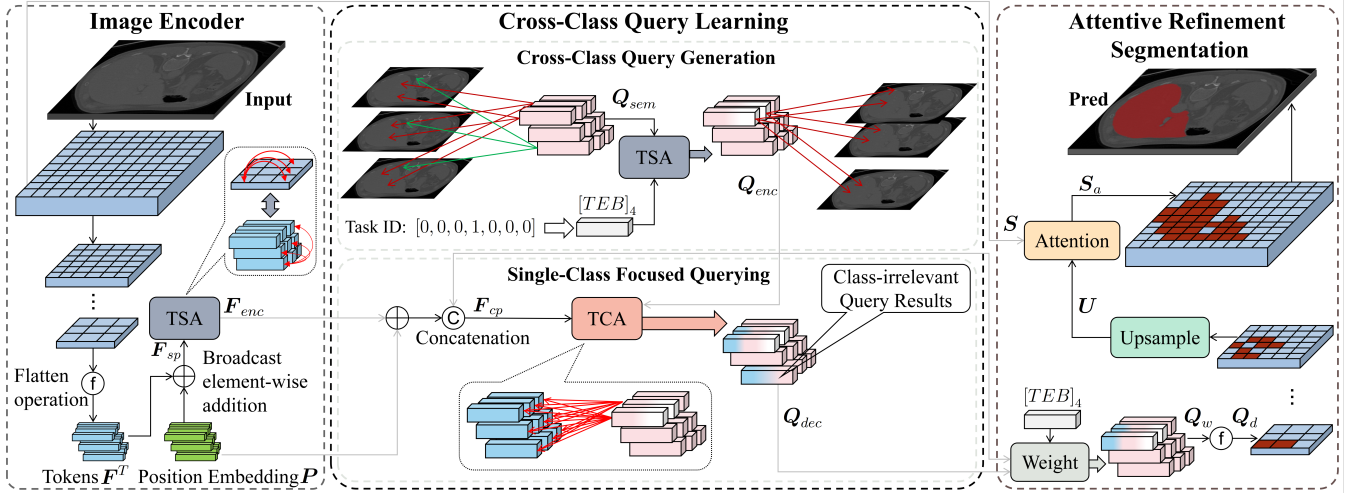


Figure 2: An overview architecture of the proposed Cross-Class Query network (CCQ). The gradient color blocks with pink and white are used to represent the class-relevant query vectors, where pink represents the semantic concepts and white represents class-relevant information.

et al. 2021). In particular, the target adaptive loss (Fang and Yan 2020) and marginal loss (Shi et al. 2021) are proposed for training each assigned task by blending the label of other tasks with the background, respectively. In addition to multi-head networks, some other approaches are also introduced. PaNN (Zhou et al. 2019) applies the prior information counted by a fully labeled dataset to help train the model on partially-labeled datasets. Liu *et al.* (Liu, Xiao, and Zhou 2021) apply the strategy of incremental learning to train a multi-organ segmentation model step by step. At each step, a new single labeled dataset is fed into the model. DoDNet (Zhang et al. 2021) generates a conditional convolution block via task information to perform dynamic segmenting.

As for class-relevant representation learning methods, we find learning the class-relevant representation of the input image is very crucial to guide the encoder or decoder in the architecture to perform class-relevant segmentation. To learn the class-relevant representation of the input image, some works (Chen, Xu, and Koltun 2017; Dmitriev, Kaufman et al. 2019) encode the class information into encoder/decoder. Cond-Enc (Chen, Xu, and Koltun 2017) incorporates the class information with the input image representation. Dmitriev *et al.* (Dmitriev, Kaufman et al. 2019) encode the class-relevant information as a part of the intermediate activation signals between the convolution layer and the non-linear layer to achieve multi-class segmentation learning from several single-class datasets.

However, they neglect the explicit relation between organs, tumors, and organs and tumors of different classes. We believe this prior relation (*e.g.*, relative locations among different organs and tumors) is crucial for multi-organ and tumor segmentation. Unlike existing methods, we generate class-relevant query vectors by incorporating class information and capturing the relations among semantic concepts.

Problem Definition

The definition of learning a single model for multi-organ and multi-tumor segmentation from partially-labeled datasets is given as follows. Given K partially-labeled datasets $\{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_K\}$, where k is the index of k -th dataset, it is defined as the task ID in this paper. The categories of the organs are not overlapped between different datasets. Therefore, each task ID indicates one class of an organ. For simplicity of demonstration, we only introduce organs here, and the tumors can be easily included in the datasets by defining new task IDs for tumors. The k -th dataset $\mathcal{D}_k = \{(X^k, Y^k)\}$ contains a set of input images $X^k = \{\mathbf{x}_i^k\}_{i=1}^{N_k}$ and corresponding segmentation masks $Y^k = \{\mathbf{y}_i^k\}_{i=1}^{N_k}$, where N_k is the number of samples in dataset \mathcal{D}_k . Each voxel in \mathbf{y}_i^k is binary where 0 represents the background and 1 represents the organ. Given a task ID \hat{k} and an input image $\hat{\mathbf{x}} \in \mathbb{R}^{D \times H \times W}$, where $H \times W$ is the size of each slice and D is the number of slice, our goal is to predict the corresponding mask $\hat{\mathbf{y}}$.

Method

The overall architecture of Cross-Class Query network (CCQ) is shown in Figure 2. It contains three modules of the image encoder, cross-class query learning and attentive refinement segmentation, each of which will be introduced in an individual subsection.

Image Encoder

As shown in Figure 2, we utilize a CNN-Transformer hybrid to extract the image features and long-range dependency of the input image. Specifically, given an input image, we utilize the CNN encoder to extract its feature maps in multiple levels and denote the feature map at the last layer as $\mathbf{F} \in \mathbb{R}^{\frac{D}{16} \times \frac{H}{16} \times \frac{W}{16} \times C}$, where C is the number of channels. \mathbf{F} represents the local semantic features of the input image.

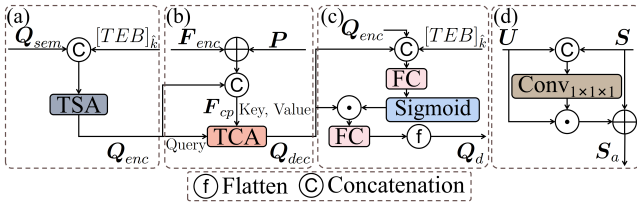


Figure 3: The detailed architectures of main modules in CCQ. (a) Cross-Class Query Generation, (b) Single-Class Focused Querying, (c) Weight and (d) Attention. \odot is the element-wise multiplication and \oplus is the element-wise addition.

We next obtain the features F_{enc} with the long-range dependency by utilizing the self-attention layer of Transformer (TSA) (Vaswani et al. 2017) to encode the local features.

We first tokenize F to $F^T \in \mathbb{R}^{N \times C}$ as the input of transformer encoder, where $N = \frac{D}{16} \times \frac{H}{16} \times \frac{W}{16}$. Then, we compute $F_{enc} \in \mathbb{R}^{N \times C}$ from the image tokens F^T as follows,

$$\begin{aligned} F_{sp} &= F^T + P, \\ F_{enc} &= \text{TSA}(F_{sp}), \end{aligned} \quad (1)$$

where $P \in \mathbb{R}^{N \times C}$ is learnable positional embedding (DeVlin et al. 2019), and TSA is the self-attention layer of Transformer (Vaswani et al. 2017).

Cross-Class Query Learning

As shown in Figure 2, Cross-Class Query Learning considers the semantic relations and anatomical priors between different tasks (*i.e.*, categories of organs), which is crucial for partially-labeled multi-organ segmentation. Cross-Class Query Learning consists of two modules, *i.e.*, Cross-Class Query Generation and Single-Class Focused Querying. (1) Cross-class query generation module first learns the cross-class semantic concepts over different organ and tumor tasks via constructing a set of learnable vectors. Then, it generates class-relevant query vectors by incorporating class information and capturing the relations among semantic concepts via transformer self-attention. (2) Single-class focused querying module adopts query vectors from the cross-class query generation module to find the class-relevant features for segmentation.

Cross-Class Query Generation. As shown in Figure 2, Cross-Class Query Generation module first learns the cross-class semantic concepts Q_{sem} that represent the diversified concepts of different classes of organs. Semantic concepts are distinguishable, high-level representations of various semantics (*e.g.*, categories, location, and shape) of organs and tumors. Each concept represents a distinct region in the semantic space. Then, it generates class-relevant query vectors Q_{enc} by incorporating class information and capturing the relations among semantic concepts via the TSA.

Inspired by the fixed small set of learned object queries of DETR (Zhu et al. 2020), we first construct a set of learnable semantic concepts $Q_{sem} \in \mathbb{R}^{N_q \times C}$, where N_q is the number

of these semantic concepts. Here, semantic concepts are supposed to learn the semantic representations for all the organs and are the same for all samples, which can be considered as a semantic codebook. After obtaining semantic concepts Q_{sem} , we also get the learnable task token $[TEB]_{\hat{k}} \in \mathbb{R}^{1 \times C}$ corresponding to task ID \hat{k} . Q_{sem} and $[TEB]_{\hat{k}}$ are learnable vectors, they are randomly initialized at the beginning of training and optimized via gradient descent and backpropagation during training. Next, we capture the relations between semantic concepts Q_{sem} and the learnable task token $[TEB]_{\hat{k}}$ via the TSA as follows,

$$Q_{enc} = \text{TSA}([TEB]_{\hat{k}}; Q_{sem}), \quad (2)$$

where $;$ is concatenation operation. The encoded query vectors $Q_{enc} \in \mathbb{R}^{N_q \times C}$ are not only class-relevant but also be encoded with the relations between classes. For example, given a specific category of an organ (*e.g.*, Liver), the class-relevant query vectors Q_{enc} are the semantic concepts relevant to the specific category, which are generated by capturing the relations between the category and semantic concepts Q_{sem} . The detailed architecture is shown in Figure 3(a).

Single-Class Focused Querying. For a given image, the Single-Class Focused Querying module finds the class-relevant features of each class. As shown in Figure 2, after obtaining the class-relevant query vectors Q_{enc} , we perform querying process in the image representation to find the region of class-relevant features for segmentation. Specifically, we adopt the cross-attention layer of Transformer (TCA) (Zhu et al. 2020) to compute the class-relevant image features $Q_{dec} \in \mathbb{R}^{N_q \times C}$. The class-relevant query vectors Q_{enc} serves as the query, and F_{cp} works as the key and value in the cross-attention mechanism. The computation process is formulated as follows,

$$\begin{aligned} F_{cp} &= [Q_{enc}; F_{enc} + P], \\ Q_{dec} &= \text{TCA}(Q_{enc}, F_{cp}), \end{aligned} \quad (3)$$

where $;$ is concatenation operation, and the F_{cp} encodes the query vectors, the image features, and position information of image features. The detailed architecture is shown in Figure 3(b).

After performing the class-relevant query from the class-relevant query vectors to the image representations, the visual cues corresponding to the query are obtained. Note that, to further capture the relations between tasks, we additionally concatenate the query vectors to image representations to form the queried vectors in Eq. 3, which guides the cross-attention to capture the intra-relation among class-relevant query vectors.

Attentive Refinement Segmentation

We perform attentive refinement approach to predict the segmentation results by incorporating the class-relevant query vectors Q_{enc} , class-relevant features Q_{dec} , task ID \hat{k} and the feature maps via the attentive skip connection.

First, to better utilize the task information, we weight class-relevant features via the learnable task token $[TEB]_{\hat{k}}$

of the task \hat{k} and query vectors \mathbf{Q}_{enc} , and the computation of weighted query vectors $\mathbf{Q}_w \in \mathbb{R}^{N_q \times C}$ is as follows,

$$\mathbf{Q}_w = \sigma(\text{FC}([\mathbf{Q}_{dec}; \mathbf{Q}_{enc}; [\text{TEB}]_{\hat{k}}])) \odot \mathbf{Q}_{dec}, \quad (4)$$

where FC is the linear projection, σ is the sigmoid activation function, $[\cdot]$ is concatenation operation and \odot is the element-wise multiplication. The detailed computation of \mathbf{Q}_w is shown in Figure 3(c). The $\mathbf{Q}_w \in \mathbb{R}^{N_q \times C}$ is then linearly projected and reshaped to $\mathbf{Q}_d \in \mathbb{R}^{\frac{D}{16} \times \frac{H}{16} \times \frac{W}{16} \times C}$ to match the input of the CNN decoder.

Next, we utilize the skip connections from CNN encoder to help the CNN decoder to restore the complete spatial resolution of the \mathbf{Q}_d (Drozdzal et al. 2016). The input of the CNN decoder of CCQ is class-relevant, where skip connections from class-irrelevant CNN encoder mislead the decoding segmentation. To eliminate the class-irrelevant noise, we modify the skip connection by performing the extra attention between the feature map \mathbf{U} in each layer of CNN decoder and its corresponding skip connection \mathbf{S} . Specifically, we compute the attentive feature \mathbf{S}_a as follows,

$$\mathbf{S}_a = \sigma(\text{Conv}_{1 \times 1 \times 1}([\mathbf{U}; \mathbf{S}])) \odot \mathbf{S} + \mathbf{S}, \quad (5)$$

where $[\cdot]$ is concatenation operation, σ is the sigmoid activate function and $\text{Conv}_{1 \times 1 \times 1}$ is a CNN block with kernel size of $1 \times 1 \times 1$. The detailed architecture is shown in Figure 3(d).

Finally, we add every attentive skip connection with its corresponding feature map in CNN decoder for segmentation.

Loss Function

Following previous methods (Zhang et al. 2021; Chen, Xu, and Koltun 2017; Dmitriev, Kaufman et al. 2019), we use Dice loss and binary cross-entropy (BCE) loss for training. Specifically, given the prediction \mathbf{y} , ground truth mask $\hat{\mathbf{y}}$ and voxel number V , our loss function is formulated as follows.

$$L = 1 - \frac{2 \sum_{i=1}^V y_i \hat{y}_i}{\sum_{i=1}^V (y_i + \hat{y}_i + \epsilon)} - \sum_{i=1}^V (\hat{y}_i \log y_i + (1 - \hat{y}_i) \log(1 - y_i)), \quad (6)$$

where $y_i \in \mathbf{y}$ and $\hat{y}_i \in \hat{\mathbf{y}}$ are each voxel in them.

Experiment

Experiment Setup

Dataset. We evaluate the proposed CCQ on the large-scale, partially-labeled MOTS (Zhang et al. 2021) dataset, which consists of seven 3D medical image segmentation datasets of abdominal organs and tumors, including LiTS (Bilic et al. 2019), KiTS (Heller et al. 2019), Colon (Simpson et al. 2019), Lung (Simpson et al. 2019), Spleen (Simpson et al. 2019), Pancreas (Simpson et al. 2019) and Hepatic Vessel (Simpson et al. 2019) from Medical Segmentation Decathlon (MSD) (Simpson et al. 2019). MOTS contains total 1155 3D abdominal CT scans. For a fair comparison, we follow previous methods (Zhang et al.

2021; Chen, Xu, and Koltun 2017; Dmitriev, Kaufman et al. 2019), to split the dataset, including the same 920 scans for training, 235 for testing. All the scans are re-sampled to $1.5 \times 0.8 \times 0.8 \text{mm}^3$. The CT intensity values are linearly normalized to $[-1, 1]$ with the window of $[-325, 325]$.

Metrics. Dice similarity coefficient (Dice) and Hausdorff distance (HD) are used as evaluation metrics to evaluate the models. The former evaluates the coincidence degree between the prediction and the ground-truth, while the latter is the maximum value of the shortest distance from the point in prediction to the point in the ground-truth, which detects the quality of prediction boundaries.

Training and Inference. The optimizer of stochastic gradient descent (SGD) with a momentum of 0.99 is used to optimize the network. The learning rate is set to 0.01 with 0.9 decay. We randomly obtain the sub-volume with a size of $64 \times 192 \times 192$ of every input image in the training stage and use the sliding window with the same size in the prediction stage. All models are trained in a workstation with 4 Tesla V100 GPUs.

Comparison with State-of-the-Arts

We compare the proposed model with the state-of-the-art models (SOTAs). These methods are divided into two main types: class-relevant representation learning methods and non-class-relevant representation learning methods. The non-class-relevant representation learning methods include the individual networks respectively trained on seven partially-labeled datasets (*i.e.*, Multi-Nets), two multi-head networks (*i.e.*, TAL (Fang and Yan 2020), Multi-Head (Chen, Ma, and Zheng 2019)), and dynamic head network (*i.e.*, DoDNet (Zhang et al. 2021)). The class-relevant representation learning method includes Cond-Enc (Chen, Xu, and Koltun 2017), Cond-Dec (Dmitriev, Kaufman et al. 2019) and our CCQ. For a fair comparison, we strictly follow the experimental setting of SOTAs and use the same backbone (*i.e.*, 3D UNet) as theirs. The comparison results are presented in Table 1.

Overall Performance. Our CCQ improves the SOTAs by +1.73% and -7.59 in terms of the average Dice and average HD, respectively, compared with the existing best class-relevant representation learning method (*i.e.*, Cond-Enc (Chen, Xu, and Koltun 2017)). Our CCQ also outperforms the existing best non-class-relevant representation learning method (*i.e.*, DoDNet (Zhang et al. 2021)) by +0.77% and -2.7 in terms of the average Dice and average HD, respectively. The results demonstrate the effectiveness of CCQ for the partially-labeled multi-organ and tumor segmentation. More specifically, compared with Cond-Enc (Chen, Xu, and Koltun 2017), CCQ achieves the top one in 18 of 24 indicators (*i.e.*, Dice and HD on every task, average Dice and HD on all tumor tasks, organ tasks, and tumor and organ tasks), showing the consistent improvement of our method over different organs and tumors.

Challenging Tumor Segmentation. Thanks to the semantic relations modeling between categories of the organs or tumors, our CCQ improves significantly for the

Methods	Average		Lung		Spleen		Hepatic Vessel				Colon	
	Dice	HD	Dice	HD	Dice	HD	Dice		HD		Dice	HD
			Tumor	Tumor	Organ	Organ	Organ	Tumor	Organ	Tumor	Tumor	Tumor
Non-Class-Rep:												
Multi-Nets	71.67	28.95	54.51	53.68	93.76	2.65	63.04	72.19	13.73	50.70	34.33	103.91
TAL	73.35	23.56	61.85	39.92	93.01	3.10	61.90	72.68	13.86	43.57	48.08	66.42
Multi-Head	74.55	26.22	64.75	34.22	94.01	3.86	59.49	69.64	19.28	79.66	50.89	59.00
DoDNet	75.64	19.50	71.25	10.37	93.91	3.67	62.42	73.39	13.49	53.56	51.55	58.89
Class-Rep:												
Cond-Enc	74.68	24.39	60.29	58.02	93.51	4.32	62.17	73.17	13.61	43.32	51.43	44.18
Cond-Dec	72.71	29.63	57.68	53.27	90.14	6.52	61.29	72.46	14.05	65.57	51.80	63.67
Our CCQ	76.41	16.80	70.51	11.04	94.57	2.71	62.50	76.94	13.86	20.76	54.77	57.88
Methods	Liver				Pancreas				Kidney			
	Dice		HD		Dice		HD		Dice		HD	
	Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ	Tumor
Non-Class-Rep:												
Multi-Nets	96.61	61.65	4.25	41.16	82.53	58.36	9.23	26.13	96.52	74.89	1.79	11.19
TAL	96.18	60.82	5.99	38.87	81.35	59.15	9.02	21.07	95.95	75.87	1.98	15.36
Multi-Head	96.75	64.08	3.67	45.68	83.49	61.22	6.40	18.66	96.60	79.16	4.69	13.28
DoDNet	96.87	65.47	3.35	36.75	82.64	60.45	7.88	15.51	96.52	77.59	2.11	8.91
Class-Rep:												
Cond-Enc	96.68	65.26	6.21	47.61	82.53	61.20	8.09	31.53	96.82	78.41	1.32	10.10
Cond-Dec	95.27	63.86	5.49	36.04	77.24	55.69	17.60	48.47	95.07	79.27	7.21	8.02
Our CCQ	96.71	64.32	3.73	31.22	83.18	60.54	7.20	30.07	96.68	79.82	1.44	4.89

Table 1: The comparison of segmentation accuracy (higher is better for Dice, while lower is better for HD) of state-of-the-art models on the MOTS dataset. ‘‘Average’’ is the aggregative indicator that averages the Dice or HD over 11 categories.

more challenging tumor segmentation of Hepatic Vessel and Colon by considering their relations with the corresponding organs as well as other organs. Particularly, CCQ outperforms existing best-performing methods by +3.77% and -22.56 in terms of Dice and HD on tumor segmentation of Hepatic Vessel as CCQ captures the patch-level relevance between the tumor and organ. Moreover, CCQ improves the tumor segmentation accuracy of Dice on the Colon task by a large margin, *i.e.*, +2.97%.

Long-Range Dependency and Attentive Refinement.

The proposed CCQ can be adapted to organs of various shapes and sizes as it incorporates long-range dependency features. It is challenging to segment Pancreas and other organs using a single-head conditional decoder because the Pancreas, a fish-shaped spongy whose shape is very different from other organs. Compared with Cond-Dec, our CCQ with single-head conditional decoder achieves a significant performance gain on pancreas segmentation by +5.94% on Dice and -10.40 on HD, respectively. Also, CCQ outperforms all the single-head networks for pancreas segmentation. In addition to the long-range dependency, the proposed attentive refinement segmentation can help to improve the boundaries and details of segmentation, and CCQ improves the average HD score by -7.59 .

Ablation Study

To evaluate the effectiveness of Cross-Class Query Learning and Attentive Refinement Segmentation, we have trained nine additional models for comparison. We randomly split training scans into 80% (*i.e.*, 744 scans) for training and the

Method	Ave Dice	Ave HD
w/o Q-Generation	68.89	35.02
w/o Querying	68.91	34.79
w/o Q- F_{enc} -Cat	71.39	29.26
Baseline	67.81	40.60
Baseline+IMG-Self-Attn	68.04	40.38
Baseline+IMG-Self-Attn+Attn-Skip	68.68	36.07
w/o Attn-Skip	70.02	32.60
$N_q = 16$	71.15	32.22
$N_q = 64$	71.44	26.63
CCQ _{full} ($N_q = 32$)	72.38	27.69

Table 2: Ablation study of the proposed CCQ network using Average (Ave) Dice (%) and Average (Ave) HD over 11 categories. Variants of the CCQ are based on different modules and query numbers.

rest 20% for validation (*i.e.*, 186 scans) in the ablation study. Experiment results are presented in Table 2.

Cross-Class Query Learning Module. We compare the CCQ with its three variants, and the results are shown from row 1 to row 3 in Table 2. First, we evaluate the necessity and effectiveness of relation modeling between cross-class semantic concepts Q_{sem} . We remove the Cross-Class Query Generation module (‘‘w/o Q-Generation’’) and the concatenation operation between query vectors Q_{enc} and image features F_{enc} (‘‘w/o Q- F_{enc} -Cat’’), respectively. The model ‘‘w/o Q-Generation’’ does not capture the relations between semantic concepts Q_{sem} , and the model ‘‘w/o Q- F_{enc} -Cat’’ does not guide the cross-attention to cap-

ture the intra-relation among class-relevant query vectors Q_{enc} . Compared to our full model, the performance degradation of -3.49% , -0.99% on average Dice and $+7.33$, $+1.57$ on average HD of the model “w/o Q-Generation” and the model “w/o Q_{enc} -Cat” demonstrates that capturing the relations between categories of organs is crucial for partially-labeled multi-organ segmentation. After removing the Single-Class Focused Querying module in our full CCQ (*i.e.*, “w/o Querying”), the performance loss of -3.47% on Dice and $+7.1$ on HD shows that finding the class-relevant features in the image for segmentation is significant.

Image Encoder & Attentive Refinement Segmentation. For a fair comparison, we choose the previous method (Zhang et al. 2021) as our baseline. We add IMG-Self-Attn (*i.e.*, image transformer self-attention module) to baseline and further replace the conventional skip connection with Attn-Skip (*i.e.*, attentive skip connection) successively. The image transformer self-attention module captures the long-range dependency of the input image. The attentive skip connection encodes the attentive details of the image to the feature map in each layer of the CNN decoder. As shown from row 4 to row 6 in Table 2, the baseline with IMG-Self-Attn improves the performance by $+0.23\%$ and -0.22 respectively on Dice and HD compared with the baseline. Furthermore, adding the Attn-Skip further boosts the performance by $+0.64\%$ and -4.31 on Dice and HD, respectively. In addition, we also replace the conventional skip connection in our full CCQ with Attn-Skip (*i.e.*, “w/o Attn-Skip”), where our full CCQ outperforms the model (“w/o Attn-Skip”) by $+2.36\%$ and -4.91 on Dice and HD, respectively. These results validate the effectiveness of our image transformer self-attention module and attentive skip connection.

Number of Query Vectors. The experimental results of query numbers N_q are shown in row 8 and row 9 of Table 2. The model with $N_q = 32$ outperforms the model with $N_q = 16$ by $+1.23\%$ on average Dice and -4.53 on average HD. The possible reason may be that 16 query vectors are not enough to represent all classes. The model with 64 query vectors has a more accurate boundary (*i.e.*, 1.06 margin on average HD) but less overall segmentation accuracy (*i.e.*, 1.23% margin on average Dice) than our full CCQ. It is probably caused by some redundant vectors in 64 query vectors which may bring the noise for the segmentation.

Visualization

We visualize the segmentation results obtained by our CCQ and the DoDNet (Zhang et al. 2021) on seven tasks. The results are shown in Figure 4, which demonstrate that our model performs better localization and segmentation results on organs and tumors.

Specifically, the proposed attentive refinement segmentation contributes to improving the boundaries and details of segmentation and makes the segmentation masks on Liver Organ & Tumor, Kidney Organ & Tumor, Lung Tumor and Spleen Organ closer to the ground truth. Besides, CCQ improves the tumor segmentation accuracy on the Colon task since our CCQ can segment the Colon tumor with a small

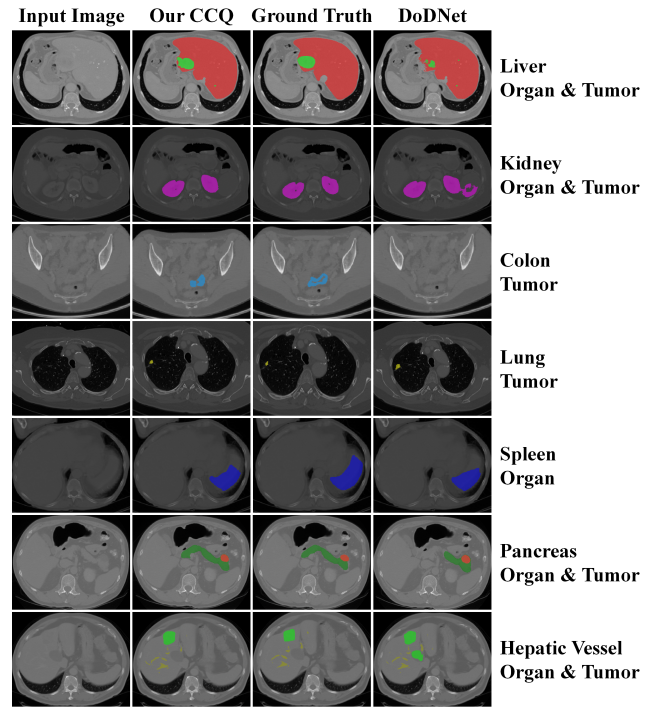


Figure 4: Visualization of segmentation results obtained by CCQ and DoDNet (Zhang et al. 2021).

size which is very challenging, while DoDNet misses the small region. Moreover, CCQ can accurately locate and segment the tumor attached with the Hepatic Vessel, while DoDNet predicts a wrong region. In addition, CCQ performs a better segmentation result on the Pancreas organ, a fish-shaped spongy extended horizontally across the retroperitoneum of the upper abdomen, which has a different shape from other organs.

Conclusion

In this paper, we propose a Cross-Class Query Network (CCQ) to model cross-class semantic concepts for multiple classes in partially labeled organ segmentation. CCQ focuses on generating class-relevant query vectors by learning a set of semantic concepts corresponding to semantic categories and capturing the relations among semantic concepts. Class-relevant query vectors implicitly incorporate and capture the semantic relations and anatomical priors between different classes of organs and tumors. To the best of our knowledge, we are the first to propose cross-class semantic concept modeling in medical image segmentation.

We also propose an attentive refinement segmentation module to incorporate the high-resolution image details into low-resolution. We apply class-relevant semantic queries to generate high-resolution semantic segmentation results without introducing the class-irrelevant noise, which improves segmentation accuracy.

Extensive experimental results demonstrate that CCQ outperforms all the state-of-the-art models on the MOTS dataset which consists of seven organ and tumor segmentation tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.62206174), the Shanghai Pujiang Program (No.21PJ1410900), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (Shang-HAI), and Shanghai Clinical Research and Trial Center.

References

- Bilic, P.; Christ, P. F.; Vorontsov, E.; Chlebus, G.; Chen, H.; Dou, Q.; Fu, C.-W.; Han, X.; Heng, P.-A.; Hesser, J.; et al. 2019. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*.
- Chen, Q.; Xu, J.; and Koltun, V. 2017. Fast image processing with fully-convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2497–2506.
- Chen, S.; Ma, K.; and Zheng, Y. 2019. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*.
- Dmitriev, K.; Kaufman; et al. 2019. Learning multi-class segmentations from single-class datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9501–9511.
- Drozdal, M.; Vorontsov, E.; Chartrand, G.; Kadoury, S.; and Pal, C. 2016. The Importance of Skip Connections in Biomedical Image Segmentation. In *Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings*, volume 10008, 179. Springer.
- Fang, X.; and Yan, P. 2020. Multi-Organ Segmentation Over Partially Labeled Datasets With Multi-Scale Feature Abstraction. *IEEE transactions on medical imaging*, 39(11): 3619–3629.
- Gibson, E.; Giganti, F.; Hu, Y.; Bonmati, E.; Bandula, S.; Gurusamy, K.; Davidson, B.; Pereira, S. P.; Clarkson, M. J.; and Barratt, D. C. 2018. Automatic multi-organ segmentation on abdominal CT with dense V-networks. *IEEE transactions on medical imaging*, 37(8): 1822–1834.
- Heller, N.; Sathianathen, N.; Kalapara, A.; Walczak, E.; Moore, K.; Kaluzniak, H.; Rosenberg, J.; Blake, P.; Rengel, Z.; Oestreich, M.; et al. 2019. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*.
- Kushnure, D. T.; and Talbar, S. N. 2021. MS-UNet: A multi-scale UNet with feature recalibration approach for automatic liver and tumor segmentation in CT images. *Computerized Medical Imaging and Graphics*, 89: 101885.
- Ledig, C.; Heckemann, R. A.; Hammers, A.; Lopez, J. C.; Newcombe, V. F.; Makropoulos, A.; Lötjönen, J.; Menon, D. K.; and Rueckert, D. 2015. Robust whole-brain segmentation: application to traumatic brain injury. *Medical image analysis*, 21(1): 40–58.
- Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J. A.; Van Ginneken, B.; and Sánchez, C. I. 2017. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88.
- Liu, P.; Xiao, L.; and Zhou, S. K. 2021. Incremental Learning for Multi-organ Segmentation with Partially Labeled Datasets. *arXiv preprint arXiv:2103.04526*.
- Navarro, F.; Shit, S.; Ezhov, I.; Paetzold, J.; Gafita, A.; Peeken, J. C.; Combs, S. E.; and Menze, B. H. 2019. Shape-aware complementary-task learning for multi-organ segmentation. In *International Workshop on Machine Learning in Medical Imaging*, 620–627. Springer.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.
- Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; and Rueckert, D. 2019. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53: 197–207.
- Shi, G.; Xiao, L.; Chen, Y.; and Zhou, S. K. 2021. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 70: 101979.
- Simpson, A. L.; Antonelli, M.; Bakas, S.; Bilello, M.; Farahani, K.; Van Ginneken, B.; Kopp-Schneider, A.; Landman, B. A.; Litjens, G.; Menze, B.; et al. 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.
- Wang, W.; Chen, C.; Ding, M.; Yu, H.; Zha, S.; and Li, J. 2021. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 109–119. Springer.
- Wang, Y.; Zhou, Y.; Shen, W.; Park, S.; Fishman, E. K.; and Yuille, A. L. 2019. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *Medical image analysis*, 55: 88–102.
- Wardhana, G.; Naghibi, H.; Sirmacek, B.; and Abayazid, M. 2021. Toward reliable automatic liver and tumor segmentation using convolutional neural network based on 2.5 D models. *International journal of computer assisted radiology and surgery*, 16(1): 41–51.
- Xie, Y.; Zhang, J.; Shen, C.; and Xia, Y. 2021. CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*.
- Zhang, J.; Xie, Y.; Xia, Y.; and Shen, C. 2021. DoDNet: Learning to segment multi-organ and tumors from multiple

partially labeled datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1195–1204.

Zhou, Y.; Li, Z.; Bai, S.; Wang, C.; Chen, X.; Han, M.; Fishman, E.; and Yuille, A. L. 2019. Prior-aware neural network for partially-supervised multi-organ segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10672–10681.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*.